



Virginia Commonwealth University
VCU Scholars Compass

Theses and Dissertations

Graduate School

2013

Accounting for Model Uncertainty in Linear Mixed-Effects Models

Adam Sima
Virginia Commonwealth University

Follow this and additional works at: <https://scholarscompass.vcu.edu/etd>



Part of the [Biostatistics Commons](#)

© The Author

Downloaded from

<https://scholarscompass.vcu.edu/etd/2950>

This Dissertation is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

Accounting for Model Uncertainty in Linear Mixed-Effects Models

by

Adam P. Sima

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Biostatistics
Virginia Commonwealth University
2013

Doctoral Committee:

Jessica M. Ketchum, Department of Biostatistics (Committee Chair)
Robert E. Johnson, Department of Biostatistics (Vanderbilt University)
David J. Edwards, Department of Statistics
Lori A. Michener, Department of Physical Therapy
Nitai D. Mukhopadhyay, Department of Biostatistics
Roy T. Sabo, Department of Biostatistics

© Adam P. Sima 2013

All Rights Reserved

ACKNOWLEDGEMENTS

Arthur C. Clarke famously stated that “Anything that is theoretically possible will be achieved in practice, no matter what the technical difficulties are, if it is desired greatly enough”. When pursuing these achievements, the desire that Clarke refers to can be a fleeting mirage that comes and goes with the hardships and achievements of scientific progress. If it were not for my friends, colleagues and family, this desire would evaporate into the sands of time.

I would like to thank all of my friends for giving me shelter from the unrelenting pressures in my academic journey. While the number of friends that have helped me out is large, I would like to thank Eric V, Nick W, Matt N, Dickie T, Greg N and Adam D for remaining close even though the years and distances have grown long. Throughout my career as a graduate student, Arnold S, Nate S (and Sam), my Glanville family (Nat M, Brian and Heather T, Derek K and Claudia G, JC and Ana, Carrol) and Marlowe PFFBFF have always been able to lift my spirits, especially when it was at its lowest.

My colleagues have also been instrumental in keeping my desire to finish this work alive. It has been an honor to Work with Emily S, Sarah R, Bob and Caroline C, Jiayi H, Amber W, and Cindy S, and I wish them all well in their future careers. I have learned so much from Jessie K and Roy S, and I hope to continue to learn and grow with them into the future.

Last, and certainly not least, my family has made me who I am, and I owe them a debt of gratitude that I will not be able to repay in a thousand lifetimes. My aunts,

uncles, and cousins, of which there are too many to list here, provided a tremendous amount of support. My grandparents on both sides of the family have offered their unwavering kindness. My parents, Paul and Mary Ellen, and siblings, Michael and Katherine, have showed an extraordinary amount of understanding that will never be forgotten.

Thank you.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	ii
LIST OF FIGURES	vii
LIST OF TABLES	ix
ABSTRACT	x
CHAPTER	
I. Introduction	1
1.1 Model Selection from a Conceptual Point of View	1
1.2 A Randomized Clinical Trial for Cervical Radiculopathy	4
1.3 Outline	6
II. Background and Characteristics of Post-Model Selection Es- timators	8
2.1 Model selection in Linear Regression Models	9
2.1.1 Model Selection Strategies	10
2.1.2 Issues with Model Selection	13
2.1.3 Accounting for Model Uncertainty	15
2.2 Linear Mixed-Effects Models	17
2.3 Simulation Study	21
2.4 Cervical Radiculopathy Clinical Trial	30
2.5 Discussion & Conclusions	36
III. Adjusting for Model Uncertainty in Variance Estimation . .	41
3.1 Properties of Post-Model Selection Estimators	42
3.2 Generalized Degrees of Freedom	45
3.3 Estimation of the GDF	49
3.4 Accounting for Model Uncertainty in Variance Estimation . .	52

3.5	Simulation Study	53
3.6	Variance Estimation in the Cervical Radiculopathy Trial . . .	63
3.7	Discussion & Conclusions	65
IV.	Adjusting for Model Uncertainty in Inference	69
4.1	Model Uncertainty and Hypothesis Testing	69
4.2	Properties of Post-Model Selection Fixed-Effect Estimators .	70
4.2.1	Variance Estimation of the Fixed-Effect Parameters accounting for Model Uncertainty	70
4.2.2	General Linear Global Approximation	71
4.3	Estimation of the GGLA	76
4.4	Hypothesis Testing and Construction of Confidence Intervals	78
4.5	Simulation Study	81
4.6	Application to the Cervical Radiculopathy Trial	92
4.7	Discussion and Conclusion	98
V.	Discussion	101
APPENDICES	106
A.1	Fixed Effect Estimation	107
A.1.1	No Parsimony Correction	107
A.1.2	With Parsimony Correction	111
A.2	Residual Variance Estimation	115
A.2.1	No Parsimony Correction	115
A.2.2	With Parsimony Correction	119
A.3	Standard Error Estimation	123
A.3.1	No Parsimony Correction	123
A.3.2	With Parsimony Correction	127
A.4	Type-I Error and Coverage Probability	131
A.4.1	No Parsimony Correction	131
A.4.2	With Parsimony Correction	135
B.1	Proof of Theorem 3.1 and Lemma 3.2	141
B.2	Proof of Theorem 3.3	143
B.3	Proof of Theorem 3.5	145
C.1	Generalized Degrees of Freedom-Proof-of-Concept	148
C.1.1	No Parsimony Correction	148
C.1.2	With Parsimony Correction	152
C.2	Relationship between GDF and Model DF – Proof-of-Concept	157
C.2.1	No parsimony correction	157
C.2.2	With Parsimony Correction	161
C.3	Generalized Degrees of Freedom-Data-Driven Method	165
C.3.1	No Parsimony Correction	165
C.3.2	With Parsimony Correction	169

C.4	Relationship between GDF and Model DF – Proof-of-Concept	174
C.4.1	No parsimony correction	174
C.4.2	With Parsimony Correction	178
C.5	Corrected Variance-Proof-of-Concept	182
C.5.1	No Parsimony Correction	182
C.5.2	With Parsimony Correction	186
C.6	Corrected Variance-Data-Driven Method	190
C.6.1	No Parsimony Correction	190
C.6.2	With Parsimony Correction	194
D.1	Fixed Effect Estimation–Proof-of-Concept	199
D.1.1	No Parsimony Correction	199
D.1.2	With Parsimony Correction	203
D.2	Fixed Effect Estimation-Data-Driven Method	207
D.2.1	No Parsimony Correction	207
D.2.2	With Parsimony Correction	211
D.3	Simulation Standard Deviation of Fixed-Effect Estimates–Proof- of-Concept	216
D.3.1	No Parsimony Correction	216
D.3.2	With Parsimony Correction	220
D.4	Simulation Standard Deviation of Fixed-Effect Estimates-Data- Driven Method	225
D.4.1	No Parsimony Correction	225
D.4.2	With Parsimony Correction	229
D.5	Standard Error Estimation - Proof-of-Concept	233
D.5.1	No Parsimony Correction	233
D.5.2	With Parsimony Correction	237
D.6	Relationship between SEs from GGLA and SEs from GLS – Proof-of-Concept	242
D.6.1	No parsimony correction	242
D.6.2	With Parsimony Correction	246
D.7	Standard Error Estimation - Data-Driven Method	250
D.7.1	No Parsimony Correction	250
D.7.2	With Parsimony Correction	254
D.8	Relationship between SEs from GGLA and SEs from GLS – Data-Driven	259
D.8.1	No parsimony correction	259
D.8.2	With Parsimony Correction	263
D.9	Type-I Error and Coverage Probability - Proof-of-Concept . .	267
D.9.1	No Parsimony Correction	267
D.9.2	With Parsimony Correction	271
D.10	Type-I Error and Coverage Probability - Data-Driven Method	276
D.10.1	No Parsimony Correction	276
D.10.2	With Parsimony Correction	280

BIBLIOGRAPHY	285
-------------------------------	-----

LIST OF FIGURES

Figure

2.1	Simulation results from estimating β_1 without adjusting for model uncertainty ($\rho=0.25$)	25
2.2	Simulation results from estimating β_1 without adjusting for model uncertainty ($\rho=0.75$)	26
2.3	Simulation results from estimating σ^2 without adjusting for model uncertainty ($\rho=0.25$)	27
2.4	Simulation results of estimating the standard error in β_1 without adjusting for model uncertainty ($\rho=0.25$)	29
2.5	Simulation results for Type-I error rate without adjusting for model uncertainty ($\rho=0.25$)	31
3.1	Proof-of-concept simulation results showing the GDF and average model DF ($\rho=0.25$)	55
3.2	Probabilities that the GDF is larger than the model degrees of freedom ($\rho=0.25$)	57
3.3	Proof-of-concept simulation results showing the adjusted and corrected variance estimates ($\rho=0.25$)	59
3.4	Data-driven simulation results showing the GDF and average model DF ($\rho=0.25$)	60
3.5	Data-driven simulation results showing the GDF and average model DF ($\rho=0.75$)	61
3.6	Probabilities that the GDF is larger than the model degrees of freedom ($\rho=0.75$)	62
3.7	Data-driven simulation results showing the adjusted and corrected variance estimates ($\rho=0.25$)	64
4.1	Proof-of-concept simulation results showing the GGLA Estimate of the focus parameter ($\rho=0.25$)	83
4.2	Proof-of-concept standard deviation of the parameter estimates from the simulation study ($\rho=0.25$)	85
4.3	Proof-of-concept simulation results showing the standard errors stemming from the GGLA and corrected methods ($\rho=0.25$)	86
4.4	Probabilities that the standard errors from the GGLA method are larger than the GLS estimators ($\rho=0.25$)	87

4.5	Data-driven simulation results showing the standard errors stemming from the GGLA and corrected methods ($\rho=0.75$)	89
4.6	Proof-of-concept simulation results showing the Type-I error rates ($\rho=0.25$)	91
4.7	Data-driven simulation results showing the Type-I error rates ($\rho=0.25$)	93

LIST OF TABLES

Table

2.1	Summary of Variables in Cervical Radiculopathy Trial	33
2.2	Screening and Multivariable Analysis of Cervical Radiculopathy Data	40
4.1	Multivariable analysis of the cervical radiculopathy data using the corrected method	94
4.2	Multivariable analysis of the cervical radiculopathy data using both the GGLA and combined method	96
4.3	Omnibus test for the effect of cervical traction and confidence inter- vals for the estimated NDI for each treatment group at 2 and 4 weeks using all methods	97

ABSTRACT

Accounting for Model Uncertainty in Linear Mixed-Effects Models

by

Adam P. Sima

Chair: Jessica M. Ketchum

Standard statistical decision-making tools, such as inference, confidence intervals and forecasting, are contingent on the assumption that the statistical model used in the analysis is the true model. In linear mixed-effect models, ignoring model uncertainty results in an underestimation of the residual variance, contributing to hypothesis tests that demonstrate larger than nominal Type-I errors and confidence intervals with smaller than nominal coverage probabilities. A novel utilization of the generalized degrees of freedom developed by *Zhang et al.* (2012) is used to adjust the estimate of the residual variance for model uncertainty. Additionally, the general global linear approximation is extended to linear mixed-effect models to adjust the standard errors of the parameter estimates for model uncertainty. Both of these methods use a perturbation method for estimation, where random noise is added to the response variable and, conditional on the observed responses, the corresponding estimate is calculated. A simulation study demonstrates that when the proposed methodologies are utilized, both the variance and standard errors are inflated for model uncertainty. However, when a data-driven strategy is employed, the proposed methodologies show limited usefulness. These methods are evaluated with a trial assessing the performance of

cervical traction in the treatment of cervical radiculopathy.

CHAPTER I

Introduction

1.1 Model Selection from a Conceptual Point of View

Statistical models are used as a way to map real world phenomena to mathematical constructs. Conditional on a particular statistical model, statisticians and other methodologists have developed extensive methods to estimate, make inference and predict future responses based on a set of predictor variables. However, there is no universal accepted strategy for proposing such models, particularly if hypotheses are postulated *a priori*. To foster a discussion of these strategies, consider a study where the goal is to determine the relationship between a predictor and response variable. An easy way to determine the relationship between these two variables is to include just these variables (and perhaps an adjustment for the overall mean) in a model and use standard statistical techniques to perform estimation and inference. This strategy, which is common in practice, is attractive because it is an easy model to interpret and has close to optimal statistical power properties.

However, it is possible that the response variable is also related to one or more other predictor variables. Since there usually is a relationship among the predictor variables themselves, as well as the between each of the predictor and response variables, the conclusion using the entire set of predictor variables in a statistical model may not coincide or agree with the marginal analysis described above. To ensure

completely unbiased parameter estimates, a model containing all predictor variables should be fit. This strategy can also adjust for any moderating, confounding or mediating effects the variables have amongst themselves. However, this model may suffer difficulty of interpretation or power issues. More importantly, this model is more likely to be ‘overfit’, meaning that parameter estimates are dependent on the sample generated and are not typically generalizable to the larger population or a different sample.

The balancing of bias in estimation and simplicity in interpretation of a model cuts to the heart of model selection. Model selection procedures attempt to balance the bias introduced from omitting important predictors and the efficiency, or overfitting, due to including too many predictors. Generally speaking, for models to be useful they must include all important predictors, as omitting any important predictors can result in biased parameter estimates. Conversely, including too many predictors in a model will yield inflated variance estimates which can potentially limit the utility of inference procedures and confidence intervals in estimation and prediction. Thus, proposing statistical models is an often unacknowledged skill that truly blends statistics and field science.

Optimally, models should be proposed using *a priori* information derived from the field matter. To ensure that the chosen model contains unbiased and efficient parameter estimates, only variables that are either related to the response or are relevant for hypothesis testing should be included. There is often no clear agreement as to the set of important predictors in any particular clinical situation. Thus, statisticians have developed a broad range of data-driven model selection techniques to assist in the model building process. These techniques are designed to balance the bias and variance of the final model. There are many different model selection strategies such as best subsets and progressive selection, and these will be more broadly discussed in the subsequent sections.

Model selection techniques are most appropriately used when there is enough data to split into a training and test set. To help eliminate problems of overfitting a model to a particular set of predictors, a model selection procedure is used on the training set. When a final model is found, the test set is then used for estimation, inference, and prediction (*Hastie et al. (2001)*). In practice, splitting the data into two groups is not common, as power considerations can make splitting the data counterproductive. Clinicians have also resisted splitting the data because of concerns about generalizability, however, proper randomization can allay this issue. Rather, researchers typically use a model selection strategy on the whole set of data and then, using the same data, answer their research questions using the appropriate estimation, inference and prediction methods.

Performing inference on the same data that was used for model selection can have detrimental consequences on the variance estimates and hypothesis testing characteristics. This is due to the fact that common model selection procedures are designed to select models that slightly favor overfit models as the property of unbiasedness in the fixed-effect estimates is believed to be more favorable compared to the efficiency property. The literature contains many instances when the Type-I error of hypothesis tests and coverage probabilities of confidence intervals that depart from the nominal value when separate training and test sets are not used. Furthermore, the variance estimates themselves are biased downward when model selection and inference are performed on the same dataset.

The quantities discussed above are further biased due to the fact that estimation and inference were performed on a model that is assumed to be true, rather than one that is known to be true. If model selection is performed, it is inherently assumed that the true model is not known. After selecting a model, this model is used as if it were true; ignoring the uncertainty in the model that motivated the use of the model selection procedure in the first place. Ignoring the model uncertainty results

in an underestimation of the variance estimates, which in turn, could result in higher Type-I errors and smaller than nominal coverage probabilities.

1.2 A Randomized Clinical Trial for Cervical Radiculopathy

Cervical radiculopathy occurs when the nerves in the cervical spine become compressed by the cervical vertebrae. It is more commonly referred to as a *pinched nerve* and can result in pain in the neck that can extend down into the shoulders and upper extremities. Loss of sensation in the upper extremities is another common symptom. Treatment for cervical radiculopathy typically involves non-surgical treatment regimens such as posture education, manual therapy, exercise and cervical traction.

An overview of some common manual therapy and exercise regimens can be found online in the eAppendix published by *Young et al.* (2010). One particular treatment used for cervical radiculopathy is cervical traction, which involves pulling the head up from the neck allowing the spaces in the vertebrae of the neck to expand, thus relieving the pressure on the nerves. In a clinical trial, *Joghataei et al.* (2004) found that patients treated with cervical traction had improved grip strength after 5 physical therapy visits compared to a group of controls. However, no pain or quality of life outcomes were recorded in this study.

A multi-center randomized clinical trial was conducted by *Young et al.* (2010) to determine if patients treated with cervical traction along with a multimodal treatment program (education, exercise and manual therapy) demonstrate better quality of life outcomes compared to a control group treated with only the multimodal treatment. Only a summary of the trial will be discussed here; the details of the trial can be found in the aforementioned reference. A total of 121 patients were recruited from 7 different clinics located in Virginia, Georgia, Alabama and West Virginia. Of those patients, 81 were randomized to either receive cervical traction or a sham traction regimen along with the multimodal treatment program and were not missing their 4-week follow-up.

The primary outcomes of the study were the Numeric Pain Rating Scale (NPRS), the Neck Disability Index (NDI) and the Patient-Specific Functional Scale (PSFS) while a number of secondary outcomes were also recorded. These outcomes were measured at baseline and again at 2 and 4 weeks post-treatment. Besides the time of measurement and treatment group, variables including the patient's age (years), sex, body mass index (BMI) (kg/m^2), whether the patient had any previous symptoms, whether the injury was work-related, the duration of symptoms ($>$ or ≤ 3 weeks), the number of physical therapy visits and whether the patient had bothersome pain or numbness were recorded. Calculated scores made up the remainder of these variables, including baseline NDI, the Fear Avoidance Belief Questionnaire (FABQ) score at both 2 and 4 weeks, and baseline pain, calculated using the Numeric Pain Rating Scale. These variables were recorded for this trial but not used in the primary analysis in the report of the study. A more detailed explanation of these variables is given in the following chapter. For simplicity's sake, the only response variable considered in this work will be the NDI, which is a variable bounded in the range $[0, 50]$ where higher scores indicate higher levels of disability. Further details of how the NDI will be modeled, and the measures used to model it, are discussed in subsequent chapters.

A linear mixed-effect model was used to analyze the data from this trial because of its characteristic as a multivariable model and its ability to adjust for potential non-independence within a subject, which is inherent due to the longitudinal nature of the trial design. All previous work in accounting for model uncertainty has focused on linear regression models where any two realizations of the response variable can be considered independent. Currently, the problems caused by performing model selection on the same data used for testing and ignoring model uncertainty has not been studied in linear mixed-effects models. These models are used when the response variables cannot be considered independent. This work addresses these problems in covariance pattern models, where the error terms cannot be considered independent

and uses the cervical radiculopathy trial is used as a motivating example.

Although the length of text in this work may argue otherwise, the clinical consequences of ignoring model uncertainty serves as the impetus of this research, as opposed to the statistical problem-solving. The progress of science will always be constrained by that which we don't know. This work attempts to quantify some of this uncertainty in the model used for an analysis and incorporate it into statistical models so that the results from studies are more generalizable to real-world situations.

1.3 Outline

An overview of model selection procedures in both linear regression and linear mixed-effect models are presented in Chapter II. This section details some issues that have been discussed in the context of using the same data to perform both model selection and hypothesis testing for linear regression models and ignoring the model uncertainty in inference. This chapter will conclude with a simulation study to show that these problems extend to the linear mixed-effects model.

Chapter III discusses the behavior of the estimate of the residual variance when failing to account for model uncertainty. An estimate of the residual variance that adjusts for model uncertainty using the concept of generalized degrees of freedom will be proposed. This work extends the work of *Ye* (1998) and *Zhang et al.* (2012) to linear mixed-effect models and uses a Monte-Carlo algorithm based on perturbations of the response variable to estimate the generalized degrees of freedom. The simulation study from Chapter II will be extended to assess the results of the proposed methodology.

Extending the work presented in Chapter III, Chapter IV shows that the variance of the fixed-effect parameters may be biased when model uncertainty is ignored. A perturbation method that is similar in concept, but different in execution, to the estimation method used for the generalized degrees of freedom is used to estimate

both the fixed-effect parameters as well their covariance matrix. This work extends the work of *Shen et al.* (2004a). The aforementioned simulation study will be further extended to discuss these results.

In all of the aforementioned chapters, the impact of accounting for model selection on the randomized clinical trial for treatment of patients with cervical radiculopathy are discussed. A summary of the findings, limitations, and future work are discussed in Chapter V.

CHAPTER II

Background and Characteristics of Post-Model Selection Estimators

In view of the the outline discussed in Chapter 1, Section 2.1 defines the statistical model for model selection for the linear regression model, used when the response variable can be considered independent. Several model selection strategies are reviewed, as are the problems that occur when model selection and inference are performed on the same dataset. Section 2.2 extends the linear regression model previously defined for model selection, by *Danilov and Magnus* (2004) and *Shen et al.* (2004a), to the settings in which the linear mixed-effects model. In Section 2.3, a simulation study will shows that the same problems that plague the independent-response case are present when the error terms can be considered non-independent, such as in a covariance pattern model. Section 2.4 presents an analysis of the cervical radiculopathy trial using the mixed effects-model. Finally, Section 2.5 gives an outline of future work to be considered and recommendations for practitioners undergoing model selection.

2.1 Model selection in Linear Regression Models

Classically, the linear regression equation for a sampling unit is written as:

$$Y_i = \mathbf{x}_i' \cdot \boldsymbol{\beta} + \epsilon_i \quad (2.1)$$

where Y_i is the response for the i^{th} sampling unit, \mathbf{x}_i is a vector of fixed covariates corresponding to the i^{th} sampling unit with corresponding parameter values $\boldsymbol{\beta}$, and ϵ_i is random error, assumed to have null mean and variance σ^2 . Combining over n units, the model can be written as:

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon} \quad (2.2)$$

where \mathbf{Y} is a $n \times 1$ vector of responses, \mathbf{X} is a fixed $n \times p$ matrix of covariates and $\boldsymbol{\beta}$ are the fixed-effect parameter values for the corresponding covariates in \mathbf{X} . The random error term, $\boldsymbol{\epsilon}$, is assumed to have a multivariate normal distribution with a null mean and variance $\sigma^2 \cdot \mathbf{I}_{n \times n}$. This variance assumption implies that all observations are independent from any other observation and have the same error variance. If either all or none of the variables in \mathbf{X} are subject to model selection, then (2.2) is sufficient for incorporating a model selection procedure. However, often times researchers want to keep certain variables in the model so that inference can be made on them. For instance, researchers would always want the intervention variable of a clinical trial to remain in the model so they can perform a hypothesis test for the corresponding parameter. A model intercept is also typically also kept in the model to account for an overall mean shift from zero and not subject to model selection. Thus, the variables in (2.2) can be partitioned as:

$$\mathbf{Y} = \mathbf{X}_1 \cdot \boldsymbol{\beta}_1 + \mathbf{X}_2 \cdot \boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \quad (2.3)$$

where \mathbf{X}_1 is a $n \times p_1$ design matrix not subject to model selection and whose corresponding parameters, β_1 , will always be estimated. Conversely, \mathbf{X}_2 is a $n \times p_2$ design matrix of covariates subject to model selection and the regression parameters β_2 could either be estimated or set to zero depending on the result of the model selection procedure. The set of covariates defining \mathbf{X}_1 will be referred to as the *inference set* and \mathbf{X}_2 will be referred to as the *selection set*. The error term, ϵ , remains the same as (2.2).

2.1.1 Model Selection Strategies

There are numerous methods to select the parameters in β_2 to be estimated in the linear regression model. These methods can broadly be combined into four general strategies. The most extensive strategy that has been researched is the best-subset strategy. This strategy involves fitting all 2^{p_2} possible models and, for each model, calculating a summary statistic to be used as a metric. Depending on the statistic, the model corresponding to the highest or lowest value of this metric is the model the is considered the optimal model.

One group of metrics that are used for models of almost any form, and continue to garner much research interest, are known as information criteria. Briefly, information criteria are based on the relative Kullback-Leibler distance which measures how much information is lost when a model is used to estimate the truth. Since the truth is not known and cannot be quantified, relative distances are used, such that small values of the relative distance indicate a better approximation to the truth. The expectation of the relative Kullback-Leibler distance under the parameter-space is of the form $-2 \cdot \log L + c \cdot K$, where L is the likelihood function and K is the number of the parameters in the model, and c is a constant with respect to the data (*Burnham and Anderson (2002)*).

Many different formulations of c have been proposed. Some of these criteria are

only used with certain models or for particular types of response variables. The more common information criteria can be used in a variety of circumstances. Two of the most common types of the general information criteria are the AIC ($c=2$) and the BIC ($c=\log(n)$), with n being the total number of observations in the sample. To account for small samples, *Hurvich and Tsai* (1990) created an information criterion known as AIC_c , and uses $c = AIC + \frac{2 \cdot K \cdot (K+1)}{n-K-1}$. Rather than use the lowest value of the information criterion, it is often recommended to use a model with less parameters if this model is within 2 units of the recommended model. This is meant to adjust for any overfitting caused by searching for the optimal model, and is generally referred to as a *parsimony correction*.

The best-subset strategy does not always use the information criteria discussed previously. Other common metrics developed specifically for linear regression models are the coefficient of multiple determination (R^2), the adjusted coefficient of multiple determination (R^2_a), Mallows's C_p criterion, and the prediction sums of squares (*PRESS*). Models with high values of R^2 and R^2_a are favored over models with small values of these statistics, while the opposite is true for the *PRESS* value. For Mallows's C_p , favorable models are found when the model statistic is close to, but not greater than, the number of non-zero regression parameter in the model (*Kutner et al.* (2004)).

Another model selection strategy is known as progressive selection. Rather than fit all 2^{p_2} models, progressive selection either sequentially adds or subtracts parameters to be estimated in the model. Four types of progressive selection, forward, backward, forward stepwise and backward stepwise selection, are popular since they are easily available in software packages. Forward selection starts with an intercept-only model and performs separate analyses for each of the p_2 variables in \mathbf{X}_2 , obtaining p -values for each parameter. The variable corresponding to the smallest p -value that is smaller than a pre-specified level is added to the hypothesis set and the process repeats itself

with the $p_2 - 1$ remaining variables. The process stops when all the p -values of the non-included variables are larger than a pre-specified level. Backward selection proceeds similarly but, rather than add variables to the model, it starts with all parameters being estimated and sets the parameters equal to zero if the corresponding p -values are larger than the specified level. The forward stepwise algorithm is similar to the forward selection algorithm but, after a variable is added, the variable is reanalyzed in addition to the intercept to ensure that the p -value corresponding to the added variable is lower than the pre-specified level, and the process continues. The backward stepwise selection algorithm is similar, as it ensures it is not excluding any parameter estimates that have p -values less than the pre-specified level.

Often times researchers follow model selection strategies that are not strictly a best-subsets or progressive selection strategy. These methods are known as *ad hoc* strategies and include any sort of mixture of the best-subsets and progressive selection strategies. Also included in this category is pre-screening, where the individual bivariate association of each member of selection set is assessed with the response variable and, if it meets a particular threshold of statistical or clinical importance, it is included to be estimated in the final model. The parameter is set to be zero otherwise. Resampling methods such as bootstrap and jackknife model selection are considered *ad hoc* selection methods.

Lastly, complex and computationally intensive algorithms have been developed that achieve model selection through less traditional means. These non-traditional methods include, but are not limited to model averaging, penalized regression (ie. LASSO), adaptive selection, and many of the other algorithms that are currently being developed. These works typically further the traditional model selection procedures by allowing model selection for cases where the traditional algorithms are not suited, such as the case where $n < p$. An overview of many these methods are discussed in *Hastie et al. (2001)*.

Each of the model selection strategies has certain asymptotic properties. These describe how the model selection strategy performs compared to the true underlying model. A model selection procedure is considered *consistent* if it chooses a correct model with probability one as the sample size tends toward infinity *Shibata* (1984). If the model chosen by the selection procedure is defined by \hat{M} , the set of true models as M_o , and its complement M_o^C , then this consistency property can be expressed as $P\left(\hat{M} \subset M_o\right) \rightarrow 1$ as $n \rightarrow \infty$. The definition of the consistent property allows for a set of true models, as opposed to a single true model. The distinction between a set of true models, as opposed to a single true model, is further explored in Section 3.1 in the context of describing the behavior of the parameter estimates when model uncertainty is taken into account.

Besides being consistent, model selection procedures can be *overconsistent*, meaning the probability that an incorrect model is chosen tends toward 0, but the probability of the correct model being identified converges on an upper limit strictly less than 1 (*Leeb* (2006)), or *efficient*, where the loss of the selected model approaches the minimum loss of all considered models *Hurvich and Tsai* (1990). Of the common information criteria, the AIC is considered efficient (although in certain cases it can be consistent) while the BIC is consistent. The asymptotic convergence of the efficient and consistent model selection properties were explored by *Shao* (1997) and considered to be in probability for both consistent and efficient model selection; almost sure convergence is achieved for consistent model selection. Thus, the fixed-effect estimates will have this effect as well. Overconsistent model selection has neither of these asymptotic convergence properties.

2.1.2 Issues with Model Selection

As mentioned in Chapter 1, the strategy of performing model selection and inference on the same data has been found to have serious detrimental consequences on

characteristics of the post-model selection inference, or the process of making inference on a variable after performing both model selection and inference on the same set of data. *Shen et al.* (2004a) and *Mundry and Nunn* (2009) use simulation studies to show the Type-I error rate of the post-model selection hypothesis tests can reach as high as 50% when performing a model selection procedure and inference on the same dataset. *Demetrescu et al.* (2011) shows similar results when using a time series model. When the parameter of interest is non-zero, coverage probabilities of confidence intervals are used to assess the performance of the post-model selection estimators. *Hurvich and Tsai* (1990), *Zhang* (1992), *Kabaila* (1995), *Kabaila* (1998), *Kabaila and Leeb* (2006) and *Giri and Kabaila* (2008) concluded that the coverage probabilities of confidence intervals resulting from the linear regression model are smaller than the nominal value, typically 0.95. Relatedly, *Arditi* (1989), *Ye* (1998) and *Danilov and Magnus* (2004) each discuss how the estimated variance, $\hat{\sigma}^2$, of the residuals often underestimates the true variance σ^2 . This downward bias results in inflated statistics that describe the quality of fit of the model, such as the coefficient of determination R^2 .

Additionally, it has been noted that the distribution of the post-model selection fixed-effect estimates likely do not have a normal distribution as would be expected if the true model was known (*Sen* (1979), *Potscher* (1991), *Potscher* (1995), *Potscher and Novak* (1998), *Leeb* (2005), *Leeb and Poetscher* (2008), *Leeb and Potscher* (2005), *Leeb and Potscher* (2003), *Leeb* (2009), *Poetscher and Leeb* (2009) and *Berk et al.* (2010)). These distributions are often multi-modal and resembles mixture distributions so the hypothesis testing procedures that are typically used are not even asymptotically valid.

Many of the above references use a specific model selection strategy, such as finding the model with the lowest AIC or a backward selection method. However, *Leeb and Potscher* (2005) argue that, regardless of the model selection strategy, these

consequences stem from the process of model selection, and not the specific model selection procedure itself. This implies that, at least in small samples, the results of post-model selection inference published in the literature could potentially report biased and misleading results that may not be able to be repeated (*Ioannidis (2008)*) do to the overestimating of the significance of the effects and underestimating the length of the confidence intervals.

2.1.3 Accounting for Model Uncertainty

Even if asymptotic normality of the fixed-effect parameters is justifiable, estimates of the parameters in the statistical model are typically calculated assuming the model is known. If the true model were known, then only estimation and prediction procedures would be needed as the non-zero parameters would also be known. The commonly used estimation and prediction procedures, either numerical or exact, are contingent upon the true model being known. In practice, this situation rarely happens. Rather, the underlying model is unknown and hypothesis testing must be introduced to help not only to gain information about the relationship between the response and predictor variables, but to understand the structure of the true model. The inferences that are made are model dependent, as the estimates, hypothesis tests and predicted values can change if a different models are used.

This suggests that the issues regarding the overfitting properties of model selection procedures are linked to model uncertainty. If there were no model uncertainty, there would be no need for model selection procedures and the issues discussed previously would be non-existent. However, except in very controlled situations, such as experiments in the bench sciences sciences such as chemistry or physics, this is not the case in practice, especially in the biomedical fields. Therefore, addressing the issue of model uncertainty should control the issues of overfitting. In fact, the proposed methodology in Chapters 3 and 4 will treat the issues of model uncertainty

and overfitting as one problem.

As discussed in the Chapter 1, performing model selection procedures attempts to balance bias and efficiency of the model. This also assumes that the true model is unknown and there is some uncertainty in the true structure of the model. Under certain conditions this model uncertainty can be ignored as the effects of the uncertainty of the true model will have little, if any, impact on the choice of model, estimation, inference and predicted values. These cases are limited to situations where the elements of the design matrix are orthogonal, the sample size, n , is large and there are only a few important predictors that have a strong relationship with the response so that the after accounting for these variables, near perfect predictions can be made.

Model uncertainty is rarely taken into account in practice, and ignoring it can have serious consequences on the statistical conclusions. Because of the consistent, over-consistent or efficient property of the model selection criteria used for the model selection procedure, the fixed-effect parameters are asymptotically unbiased. However, the residual variance estimate accounting for model uncertainty has been found to be greater than or equal to the residual variance estimate assuming the true model is known (*Shibata* (1984)). Furthermore, *Zhang* (1992) proved that confidence intervals computed after model selection procedures will have smaller than nominal coverage probabilities. Thus, the issues with model selection that were discussed above were expected and caused by both overfitting from performing model selection and inference on the same set of data as well as ignoring the uncertainty due to model selection.

Little methodology has been proposed in the literature to adjust for model uncertainty. The methodology that does propose solutions to this problem generally fall into one of two categories. The methodology in the first category attempts to obtain empirical distributions for the estimated quantities and, based on these distributions, estimates, confidence intervals and hypothesis tests can be performed. *Salt*

et al. (2007), *Austin* (2008), *Wang and Lagakos* (2009) and *Finos et al.* (2010) all use various techniques to achieve this goal. On the other hand, *Ye* (1998) and *Shen et al.* (2004a) calculate alternate estimates of model parameters that directly take into account model uncertainty.

2.2 Linear Mixed-Effects Models

Model selection for the linear mixed-effects model is more complicated than in the linear regression case. This is because the linear mixed-effects model (LMM) allows for relaxing of either the homogeneous variance or independence assumptions that are present in the linear regression model. This can be seen directly from linear effects model, which, for one sampling block, is written as:

$$\mathbf{Y}_i = \mathbf{X}_i \cdot \boldsymbol{\beta} + \mathbf{Z}_i \cdot \mathbf{b}_i + \boldsymbol{\epsilon}_i \quad (2.4)$$

where \mathbf{Y}_i is $n_i \times 1$ vector of responses, \mathbf{X}_i is a fixed $n_i \times p$ design matrix for the fixed-effects with corresponding parameters $\boldsymbol{\beta}$, \mathbf{Z}_i is a $n_i \times g$ matrix of known covariates with subject-specific random effects \mathbf{b}_i , assumed to have a null mean and variance \mathbf{G} . The vector $\boldsymbol{\epsilon}_i$ is an error term, assumed to have a multivariate normal distribution with null mean and variance $\sigma^2 \cdot \mathbf{R}$ so that the matrices \mathbf{G} and \mathbf{R} are the $g \times g$ and $n_i \times n_i$ covariance matrices of \mathbf{b}_i and $\boldsymbol{\epsilon}$, respectively. Furthermore, it is also assumed that \mathbf{b}_i and $\boldsymbol{\epsilon}_i$ are independent. Also, it will be convenient to note that the variance of \mathbf{Y}_i is $\mathbf{Z}_i \cdot \mathbf{G} \cdot \mathbf{Z}_i' + \sigma^2 \cdot \mathbf{R}_i$, which can be denoted as $\sigma^2 \cdot \mathbf{V}_i$

Often times researchers do not consider all of the elements of \mathbf{V}_i to be unknown. Rather, certain structures are chosen to obtain a parsimonious error structure. These error structures only have a relatively few parameters and are relevant to a particular design. This not only aids in estimation, as there are only a few covariance terms to be estimated, but allows researchers to ensure that the covariance structure matches

the experimental design of the study. One common covariance structure is compound symmetry, which assumes all observations in a block have the same pairwise variance. This type of error structure is commonly used in cluster-randomized trials so that each sampling unit in a cluster will be related to each other, but independent of a sampling unit from another cluster. An autoregressive structure is commonly used in longitudinal trials where the relationship between observation is expected to decrease as observations become further removed from each other. *Verbeke and Molenberghs* (2000) discuss these structures in further detail.

Combining across all blocks, the linear mixed-effects model is written as:

$$\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{Z} \cdot \mathbf{b} + \boldsymbol{\epsilon} \quad (2.5)$$

where $\mathbf{X} = (\mathbf{X}'_1, \mathbf{X}'_2, \dots, \mathbf{X}'_n)'$, $\mathbf{Y} = (\mathbf{Y}'_1, \mathbf{Y}'_2, \dots, \mathbf{Y}'_n)'$, $\mathbf{Z} = (\mathbf{Z}'_1, \mathbf{Z}'_2, \dots, \mathbf{Z}'_n)'$ and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}'_1, \boldsymbol{\epsilon}'_2, \dots, \boldsymbol{\epsilon}'_n)'$. The quantities \mathbf{b} and $\boldsymbol{\epsilon}$ are random effects that are assumed to be marginally independent and have variance $\mathbf{Z} \cdot \mathbf{G} \cdot \mathbf{Z}'$ and $\sigma^2 \cdot \mathbf{R}$, respectively. Therefore, the overall variance, \mathbf{V} , is the sum of the variances of each of the random effects, or $\mathbf{Z} \cdot \mathbf{G} \cdot \mathbf{Z}' + \sigma^2 \cdot \mathbf{R}$. A further discussion of the mixed-effects model can be found in *Verbeke and Molenberghs* (2000).

Just as in the linear regression model, the LMM model can be augmented to reflect model selection. This model can be represented as:

$$\mathbf{Y} = \mathbf{X}_1 \cdot \boldsymbol{\beta}_1 + \mathbf{Z}_1 \cdot \mathbf{b}_1 + \mathbf{X}_2 \cdot \boldsymbol{\beta}_2 + \mathbf{Z}_2 \cdot \mathbf{b}_2 + \Lambda \cdot \boldsymbol{\gamma} \quad (2.6)$$

where, just as in the linear regression model, \mathbf{X}_1 is a $n \times p_1$ design matrix of fixed-effects not subject to model selection and the parameters, $\boldsymbol{\beta}_1$, will always be estimated. The matrix of fixed-effects \mathbf{X}_2 is a $n \times p_2$ is a set of covariates subject to model selection, and the regression parameters $\boldsymbol{\beta}_2$ may be estimated or set to zero depending on the result of the model selection procedure. The vectors $\mathbf{Z}_1 \cdot \mathbf{b}_1$ and $\mathbf{Z}_2 \cdot \mathbf{b}_2$

are partitions of $\mathbf{Z} \cdot \mathbf{b}$ in (2.5) such that the former is not subject to selection and the latter can have elements of \mathbf{b}_2 set to a constant zero, thus resulting in the corresponding variance parameters in \mathbf{G} being set to zero. The matrix $\mathbf{\Lambda}=(\boldsymbol{\epsilon}^{(1)},\boldsymbol{\epsilon}^{(2)},\dots,\boldsymbol{\epsilon}^{(R)})$ is a set of error terms each having a multivariate normal distribution with null mean and unique covariance structure. The variance of $\boldsymbol{\epsilon}^{(r)}$ is $\sigma^2 \cdot \mathbf{R}^{(r)}$ ($r=1,2,\dots,R$). The vector $\boldsymbol{\gamma}$ is a column of the identity matrix that is used to select an appropriate $\boldsymbol{\epsilon}^{(r)}$, so that if a data-driven model selection procedure favors the $\boldsymbol{\epsilon}^{(r)}$, $\hat{\boldsymbol{\gamma}}$ will be the r^{th} column of the identity matrix.

Model selection with LMMs can proceed in one of three fashions: fixed-effects selection, random-effects selection, or covariance-structure selection. In fixed-effects selection, it is assumed that $\mathbf{Z}_2=\mathbf{0}$ and $r=1$, so that only the parameters that correspond to the hypothesis set are subjected to model selection. Conversely, \mathbf{X}_2 can be set to $\mathbf{0}$ and r be set to 1 so that the elements of \mathbf{b}_2 , and hence the covariance parameters in \mathbf{G}_2 can be determined by a model selection procedure by selecting the appropriate \mathbf{Z}_2 matrices, resulting in random-effects selection. Both \mathbf{Z}_2 and \mathbf{X}_2 can each be set to $\mathbf{0}$ so that covariance structure selection is performed. These selection types often occur simultaneously so that, in practice, fixed-effects, random-effects, and covariance structure selection occur simultaneously.

Model selection strategies for LMMs have been proposed that take the complexities of the LMM into account. However, these strategies are less common in LMMs than in the linear regression model. A best-subsets strategy remains a tenable model selection strategy, particularly for random-effect and covariance structure selection as the number of covariance structures considered are typically small. Information criteria have received much interest as a metric because of their optimal statistical properties. These metrics can be adjusted for any of the three selection types possible in LMMs and for any type of estimation procedure. For example, when using residual maximum likelihood (REML), the REML likelihood can be used in the information

criteria to focus on random effect selection, while the full-likelihood can be used for fixed-effect or the combined fixed- and random-effect selection. Since information criteria are so highly regarded, the various different information criteria remains too numerous to sufficiently discuss in this work. *Dimova et al. (2011)* and *Vallejo et al. (2010)* assess the performance of some of the common information criterion developed for LMMs.

The different types of model selection procedures in LMMs and the co-dependence of the fixed- and random-effects in estimation makes model selection tricky at best; thus, most model selection procedures in LMMs are *ad hoc*. *Cheng et al. (2010)* provide an overview of model selection strategies for LMMs. *Littell et al. (2006)* discuss various likelihood and information criteria that can be used for model selection in LMMs. For both fixed- and random-effect selection, *Verbeke and Molenberghs (2000)* recommend searching for a covariance structure using the all the variables in the selection and hypothesis sets. Then, using a metric such as the AIC, they recommend performing a backwards selection procedure on the fixed-effects holding the covariance structure fixed. Other complex and computationally intensive algorithms have been developed for LMMs, such as extending the LASSO model (*Foster et al. (2009)*) and model averaging (*Ibrahim et al. (2011)*).

A discussion of how these strategies affect post-model selection estimation and inference is even less developed than the discussion of model selection strategies in LMMs. One difficulty in discussing this topic is that it is known that misspecifying the covariance structure can result in higher than nominal Type-I errors and smaller than nominal coverage probabilities (*Gurka et al. (2011)*). Another complexity in assessing these quantities is that inference can be made conditionally or marginally on the random effects $\mathbf{Z} \cdot \mathbf{b}$. These two different inferential strategies have different interpretations and differences in model selection criteria (*Vaida and Blanchard (2005)*) which can have impacts on the selected variance structure, and ultimately,

inference.

Prior to discussing how model selection impacts the estimation of the variance and its effect on the Type-I error in LMMs, an important restriction must be made. For any work considered here, it is assumed that both the structure and values of \mathbf{V} are known but the parameter σ^2 is unknown. This restriction is made to reduce the dimensionality of the variance parameters to one parameter that can be thought of as a proportionality constant. Although the penultimate goal of this research is to propose methodology to account for model uncertainty in LMMs for any unknown variance structure, the lack of previous research would make this goal beyond the scope of this dissertation. However, this work will study how misspecifying the values of \mathbf{V} will affect the post-model selection estimates and hypothesis test characteristics with an eye on this future work. This restriction is the same as restricting the scope of the work to covariance pattern models, as both \mathbf{Z}_1 and \mathbf{Z}_2 are set to $\mathbf{0}$.

This restriction will allow for closed form solutions of the variance estimates determined by the generalised least squares estimators. Specifically, the generalised least squares estimate for the fixed-effect parameter vector after model selection is $\hat{\boldsymbol{\beta}}_{\hat{M}} = \left(\mathbf{X}'_{\hat{M}} \cdot \mathbf{V}^{-1} \cdot \mathbf{X}_{\hat{M}} \right)^{-1} \mathbf{X}'_{\hat{M}} \cdot \mathbf{V}^{-1} \cdot \mathbf{Y}$, where $\mathbf{X}_{\hat{M}}$ is the final design matrix after model selection. The vector $\hat{\boldsymbol{\beta}}$ has a variance $\sigma^2 \left(\mathbf{X}'_{\hat{M}} \cdot \mathbf{V}^{-1} \cdot \mathbf{X}_{\hat{M}} \right)^{-1}$. When the value of σ^2 is unknown, the generalised least squares estimate of the residual variance, $\hat{\sigma}^2$, can be used in place of σ^2 . This estimate can be written as $\hat{\sigma}^2 = \frac{(\mathbf{Y} - \hat{\mathbf{Y}}_{\hat{M}})' \mathbf{V} (\mathbf{Y} - \hat{\mathbf{Y}}_{\hat{M}})}{N - p_{\hat{M}}}$, where $\hat{\mathbf{Y}}_{\hat{M}} = \mathbf{X}_{\hat{M}} \cdot \hat{\boldsymbol{\beta}}_{\hat{M}}$ and $p_{\hat{M}}$ is the number of columns in $\mathbf{X}_{\hat{M}}$. Both the residual variance and the variance of the estimated parameter vector will be discussed further in subsequent chapters.

2.3 Simulation Study

A simulation study was conducted to determine how inference is affected when both model selection and inference are performed on the same dataset. Data was sim-

ulated in the setting of (2.6) with $\mathbf{Z}_2=\mathbf{0}$ and $M=1$ so that only fixed-effects selection was performed. Ten time-independent covariates, or covariates that remain the same for all observations within a block, were simulated from a normal distribution with a mean vector $\mathbf{0}$ and covariance matrix Σ , where the (i,j) element of Σ is $0.5^{|i-j|}$. Finally, to create a complete design matrix, the covariates were repeated and put into blocks of size 4 and a column of ones were added as an intercept term. These covariates were divided up so that the hypothesis set contained the intercept and the first simulated covariate, while the selection set contained the other nine covariates. For ease of discussion, this variable will be referred to as the *focus variable* since variables analyzed in this manner are typically the focus of a study. The parameters corresponding to the hypothesis set were simulated as $\beta_1 = \mathbf{0}$ while the selection set β_2 was varied between the null vector $\mathbf{0}$ and $(1,1,0,0,0,0,0,0,1)'$, which will be referred to as Case 1 and Case 2, respectively. The error term, $\epsilon^{(1)}$, was simulated from a multivariate normal distribution with null mean and a compound symmetric covariance matrix with $\sigma^2=1$ and $\rho=0.25$. The number of blocks varied between 20 and 40.

Rather than estimate the intraclass correlation value, ρ , this value was assumed to be known and allowed to vary by increments of 0.02 so that the effect of misspecifying the intraclass correlation coefficient can be assessed. The generalised least squares estimates could be used to determine the fixed-effects estimate $\hat{\beta}$ as well as the residual variance $\hat{\sigma}^2$. For each simulated dataset, all $2^9 = 512$ possible models were fit and, for each model, the best model was determined by selecting the model corresponding to the smallest AIC value. The AIC was computed using the true likelihood, as opposed to alternatives such as the residual maximum likelihood and is attractive due to its efficient nature as a model selection procedure. The parameter corresponding to the focus variable tested for no significant effect ($H_0:\beta_1=0$) versus a general, two-sided alternative at an $\alpha = 0.05$ significance level. The critical value from the z -distribution

was used rather than the t -distribution so that all inference was made on the same power function. This will enable the tests to be comparable even if the model degrees of freedom differ. This process was repeated 1000 times for each combination of simulation parameters. For each value of the intraclass correlation, the effect of model selection bias and model uncertainty in estimating the focus variable parameter (β_1), σ^2 , the standard error of the estimate of β_1 ($\text{cov}(\hat{\beta}_1)$) and the Type-I error rate was assessed over all simulations. All simulations were performed in the IML environment in SAS (V9.3).

In addition to the simulation paradigm presented above, simulations were extended that varied some of the conditions that were held fixed. This includes factoring in the parsimony correction discussed previously, where a smaller model within 2 units of the information criterion metric is favored. The Bayesian Information Criterion (BIC) was also used to show that using the AIC was not the root cause of any issues with post-model selection inference. Two additional fixed-effect structures were investigated with these vectors having values $\beta=(0,1,0,0,0,0,0,0,0,0)$ and $\beta=(0,1,1,1,1,0,0,0,0,1)$ so that the only difference between these two cases (Cases 3 and 4, respectively) is that the focus variable parameter is changed from zero to one. Rather than assessing the Type-I error, the coverage probability of an asymptotic confidence interval is assessed. The true intraclass correlation was also changed to 0.75 to assess the impact of this variable on any conclusions. The full results of all the simulations can be found in Appendices A.1-A.4. Generally, these results are invariant to the simulation parameters and, in the cases where there is some variation, the results are discussed in this section. Otherwise, these results will only be found in the Appendices.

Figure 2.1 shows the results for the assessment of the bias in estimating the parameter of interest over the different sample sizes and simulation settings. Regardless of the value of the intraclass correlation near the true value, the parameter estimate

does not show significant bias across the simulation settings. This is an expected result that stems from the consistent property of the AIC metric. However, when the true parameter vector contains non-zero terms and the value used for the intraclass correlation diverges from the true value, bias in the fixed-effect parameter estimate is introduced for small samples (Figure 2.1b). This bias reduces in magnitude as the sample size increases (Figures 2.1c,d).

Interestingly, the characteristics discussed above change as a function of the true value of the intraclass correlation, as can be seen in Figure 2.2. When the true intraclass correlation is set to 0.75, the mean of the parameter estimate over all simulations seems to be consistently smaller than the true value. The observed bias is present regardless of the intraclass correlation used in estimation and diverges from the true parameter value when the model intraclass correlation tends to the upper limit (Figures 2.2a,b). This is an unexpected result that is not present when the true fixed-effect parameter is non-zero and disappears when the sample size increases (Figures 2.2c,d).

The bias in estimating the residual variance σ^2 can be seen in Figure 2.3. The generalised least square estimates of the residual variance underestimate the true value, regardless of simulation template used. Even though this bias is small in size, there is a consistency in the results evidenced by the very narrow 95% confidence intervals that surround the mean values. These confidence bounds are so narrow the melt together with the mean value. This bias decreases somewhat for larger sample sizes (Figures 2.3c,d), but the true value still remains outside of the confidence bounds. This result was expected in light of the previously cited literature. Regardless of the sample size or simulation template, the estimate of σ^2 becomes wildly inaccurate when the model intraclass correlation tends to the upper limit.

Prior to assessing how ignoring model selection uncertainty affects the Type-I error, the bias in the standard errors of the parameter estimates were compared to

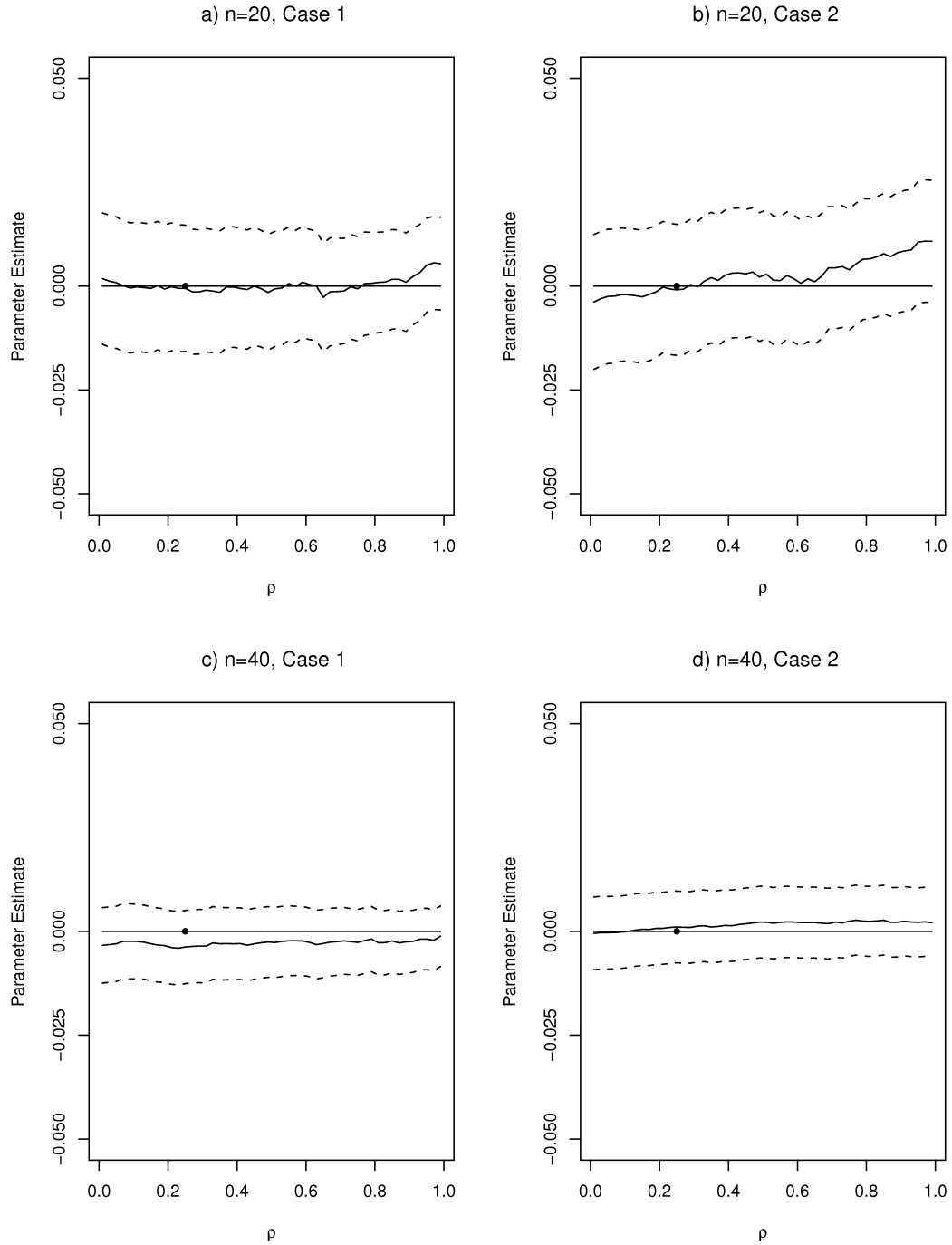


Figure 2.1: Simulation results from estimating the fixed-effect parameter of interest with $\rho=0.25$. The jagged solid line depicts the mean fixed-effect estimate over all simulations while the dashed-line represents 95% confidence bounds. The dot represents the true simulated value. The true value of the fixed-effects parameter falls within the confidence bounds for reasonable values of the intracluster correlation value.

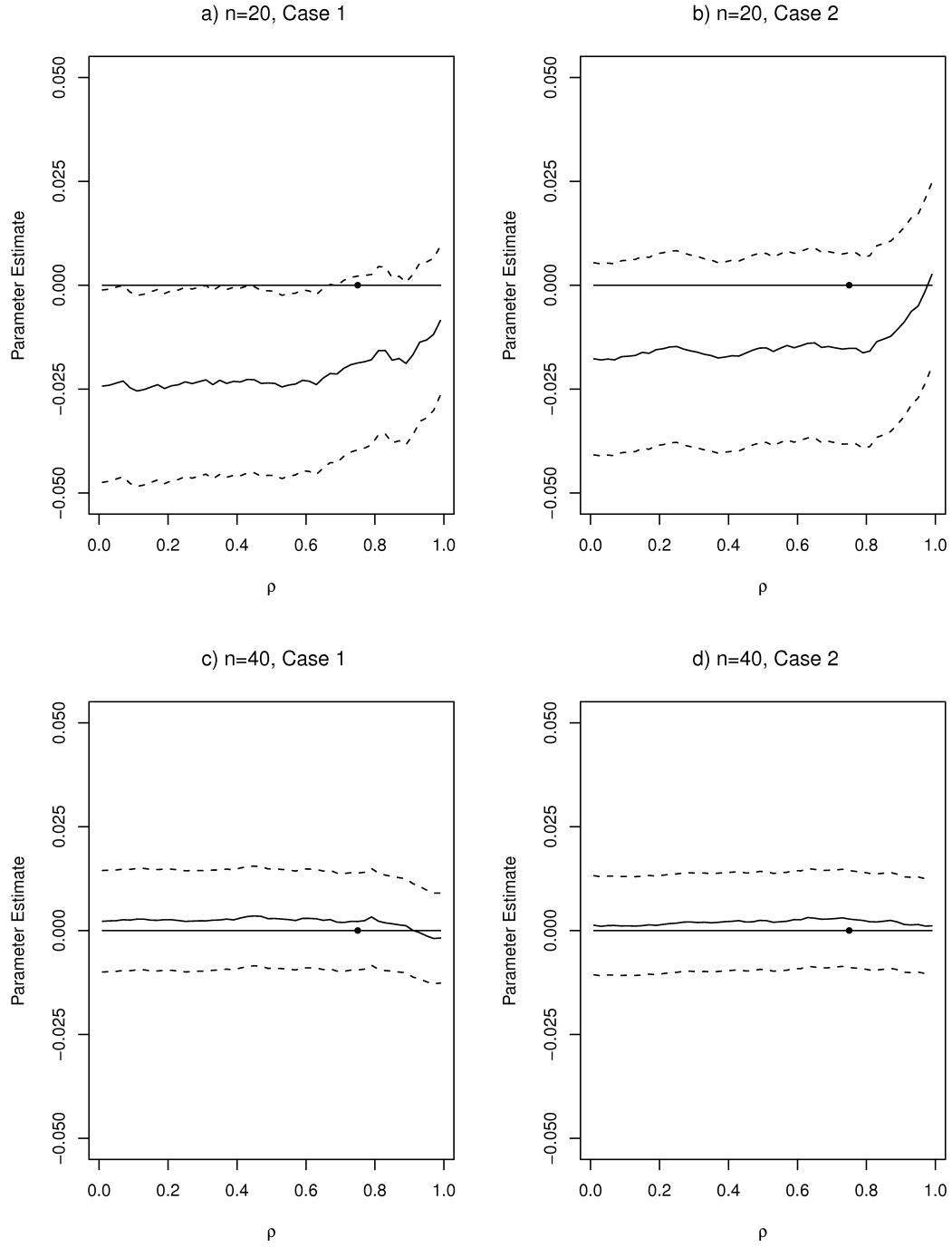


Figure 2.2: Simulation results from estimating the fixed-effect parameter of interest with $\rho=0.75$. The jagged solid line depicts the mean fixed-effect estimate over all simulations while the dashed-line represents 95% confidence bounds. The dot represents the true simulated value. The true value of the fixed-effects parameter falls within the confidence bounds for reasonable values of the intracluster correlation value, although not as justifiably when $\rho=0.25$.

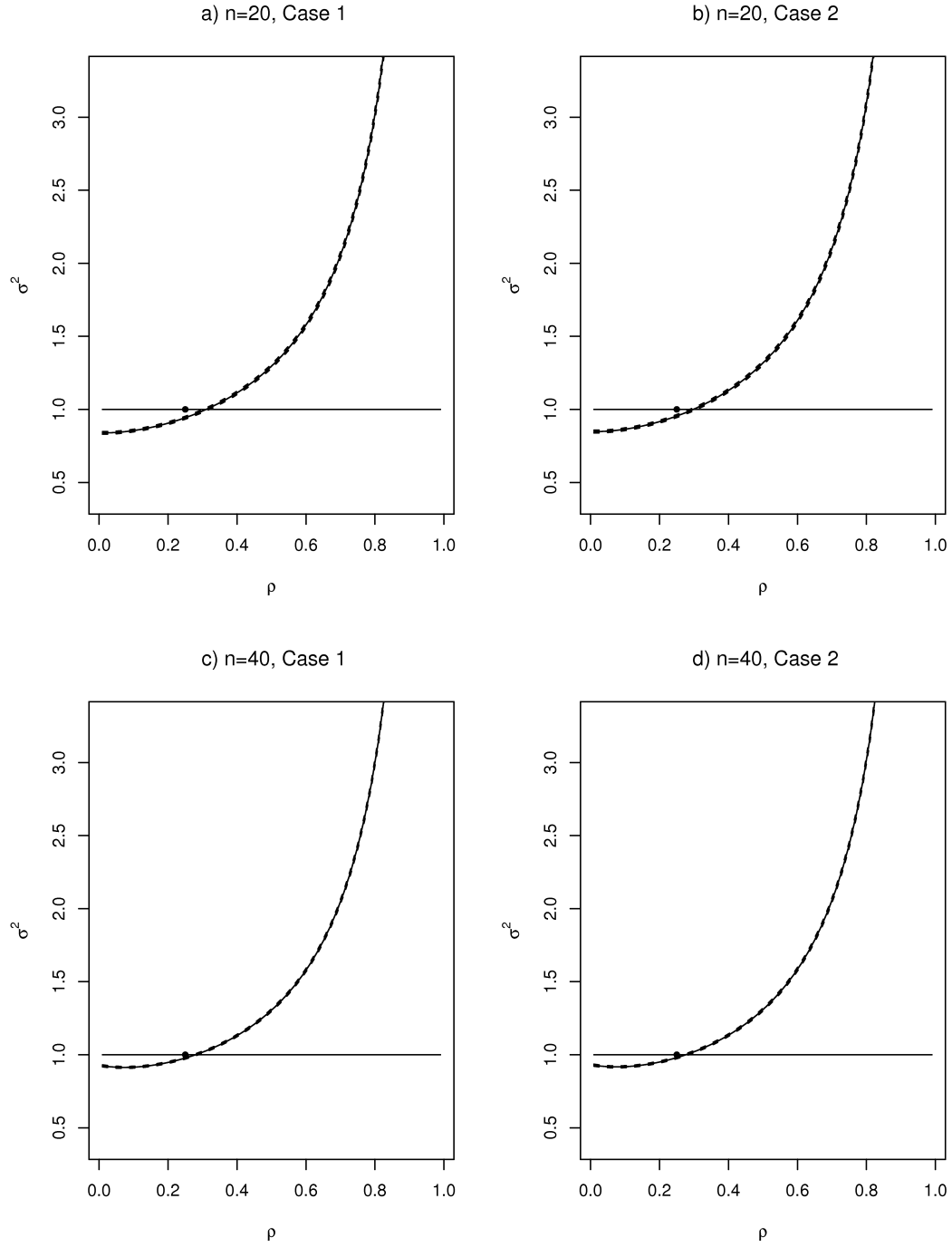


Figure 2.3: Simulation results from estimating σ^2 without adjusting for model uncertainty ($\rho=0.25$). The nonlinear solid line depicts the mean estimate of σ^2 over all simulations while the dashed-line represents 95% confidence bounds. The 95% confidence bounds are so small in some cases that they may be indistinguishable from the mean value. The dot represents the true simulated value, and falls outside of the confidence bounds for all situations.

the simulated values. Because the simulation study was not conditioned on a fixed set of covariates, the standard error of the fixed-effect parameter estimates varied between the simulations. The value of the standard error, estimated from the variables recommended from the model selection procedure, was subtracted from the similar quantity using the variables that had a just the truly non-zero coefficients. This difference is shown in Figure 2.4. There is a general upward bias in the standard errors compared to the true value (Figures 2.4a,b). This is somewhat expected because the AIC tends to favor models with a larger amount of fixed-effects over smaller models. The additional covariates in the final model will have the effect of increasing the overall variance of the parameter estimates (Figures 2.4b,d). This effect is somewhat tempered by increasing the sample size since the estimated model is typically closer to the true model as the asymptotic limit is reached (Figures 2.4c,d). One surprising result from this study is that these values are not robust against the value of the intraclass correlation used for the analysis, so that if a poor estimate of the intraclass correlation is used, the variance of the parameter estimates will be very different from the true value.

The Type-I error rates resulting from the simulations can be seen in Figure 2.5. As expected, the Type-I error rate is higher than the nominal level of 0.05 when intraclass correlation used for estimation is smaller than that of the simulated value, regardless of the sample size. Furthermore, hypothesis tests become too conservative when the intraclass correlation used is much greater than the true value. However, for small samples and intraclass correlation values close to the true value (Figures 2.5a,b), the estimated Type-I error is nearly twice the nominal value, indicating that the hypothesis test after model selection is quite liberal with respect to the nominal significance level. An increased sample size tempered this relationship somewhat (Figures 2.5c,d), but inflated Type-I error rates persisted for reasonable intraclass correlation values. The extent of this bias is somewhat surprising judging by the

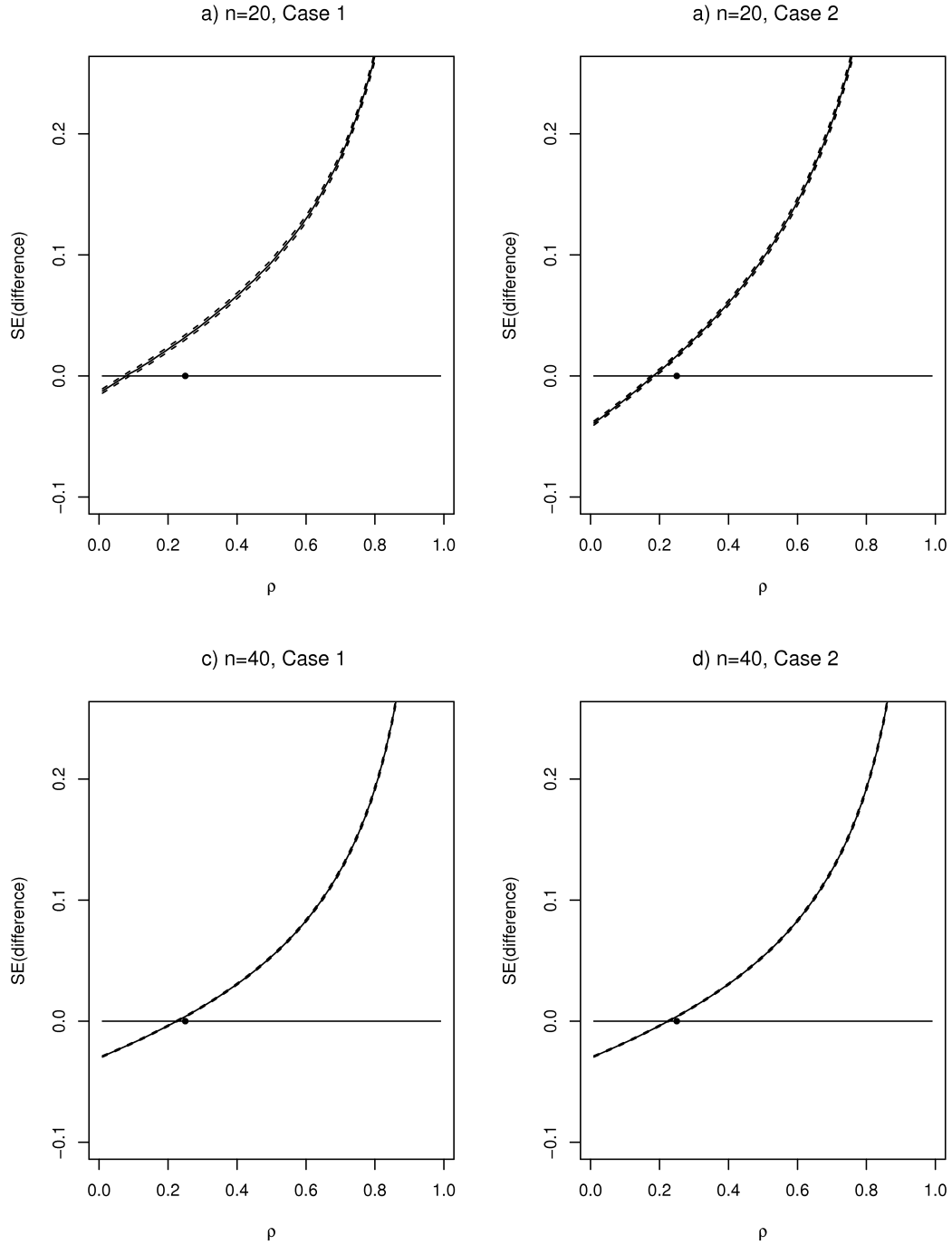


Figure 2.4: Simulation results from estimating the standard error in β_1 without adjusting for model uncertainty ($\rho=0.25$). The quantity shown is the difference between the estimate of the standard error as if the true model were known and that which was estimated through a model selection procedure. The nonlinear solid line depicts the mean estimate of the standard error over all simulations while the dashed-line represents 95% confidence bounds. The dot represents the true simulated value. The standard error after model selection is typically an overestimate due to overfitting.

results shown in Figures 2.1 and 2.4. If the fixed-effect estimate for the focus variable is somewhat unbiased and the standard error shows an upward bias, it would be expected that the ratio of these values would result in the opposite conclusion; that the resulting test would show a conservative bias. However, this is not shown as the results agree with previously cited literature for independent response variables.

2.4 Cervical Radiculopathy Clinical Trial

As introduced in Chapter 1, a clinical trial was conducted to assess the use of cervical traction for patients diagnosed with cervical radiculopathy. A detailed description of the conduct of the study can be found in *Young et al.* (2010). Each treatment arm received a standard of care regimen of manual therapy and exercise. The patients in this trial were randomized either to a group that was given the full traction treatment, which consists of applying an upward force on the neck in order to stretch out the spine, or a control group. The control group was given a sham traction procedure, where minimal force was applied. Outcomes were measured 2 and 4 weeks after physical rehabilitation began, as well as at baseline.

The original study presented “unadjusted” and “adjusted” mean differences of all outcomes between the treatment groups at 2 and 4 weeks. The adjustment that the authors mention pertains to any intra-subject variation included as part of the LMM that was used for the primary analysis. For prospective studies, this assumes that all other factors that may have influence on the outcomes were controlled for through the randomization process. However, there is the potential for this assumption to be false, potentially resulting in potentially biased estimates due to the omission of factors such as baseline levels, demographic information or other treatment characteristics. After adjusting for these variables, the effect of the cervical traction procedure may be different than the effect estimated without adjusting for these variables. This may be due to poor randomization or a relationship between these variables and the treatment

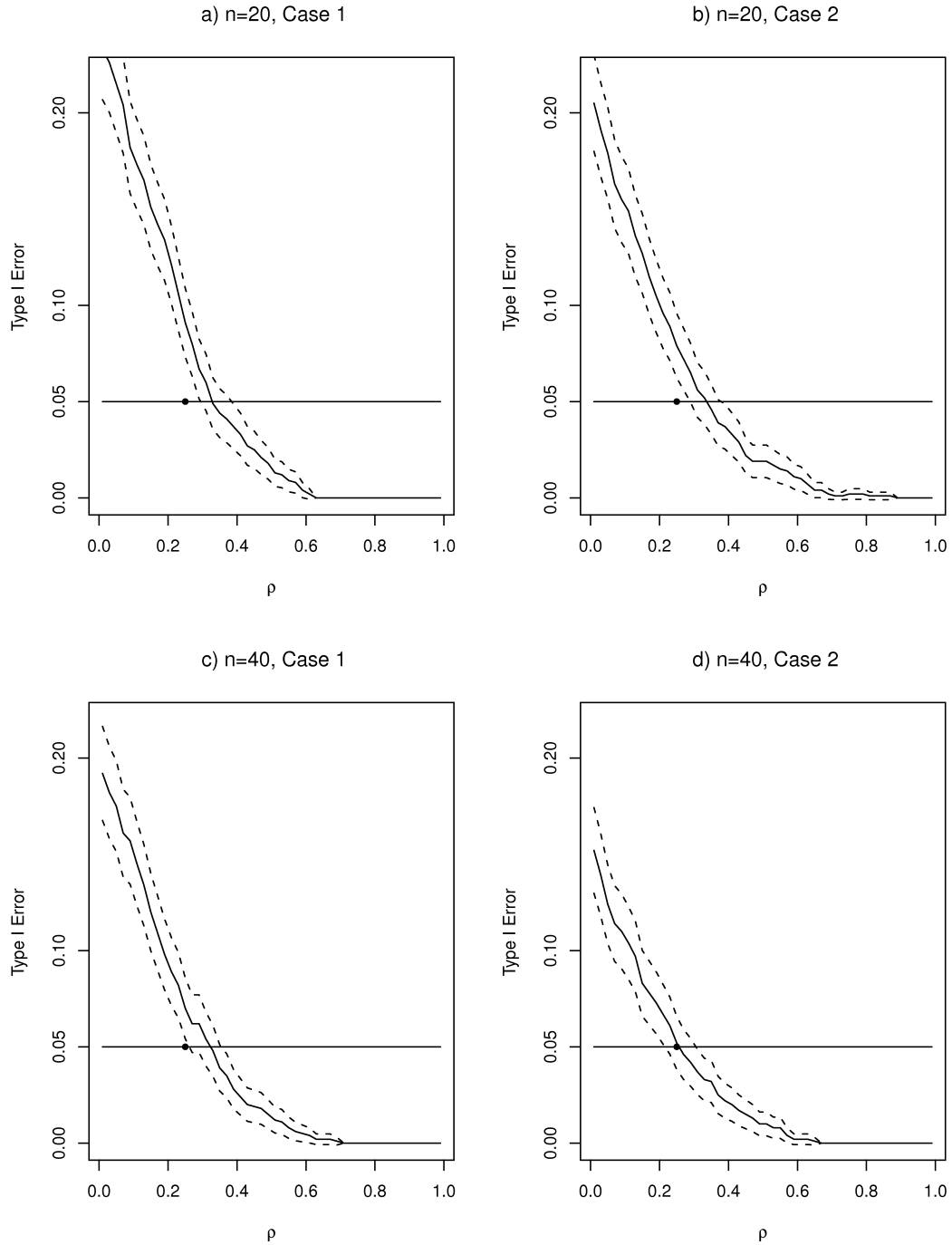


Figure 2.5: Simulation results from estimating Type-I error rate ($\rho=0.25$). The jagged solid line depicts the mean estimate of the standard error over all simulations while the dashed-line represents 95% confidence bounds. The dot represents the true simulated value. The Type-I errors are typically greater than the nominal value, indicating liberal hypothesis tests.

variable.

An analysis that adjusts for the intra-subject variation as well as other potential explanatory predictors was completed on the 2 and 4 week measurements. The LMM is able to accommodate a model that considers baseline, 2 and 4 week outcomes as a response variable, however, this was not done in this analysis to ensure the parameter estimates each have a clear meaning. The Neck Disability Index (NDI) is used for the response variable presented in this work, although any of the outcomes could be used if the assumption of the normally-distributed error terms is appropriate. This measure is a score between 0 and 50, with higher values reflective of more severe injuries.

Potential explanatory variables included the demographic variables (age, sex, BMI), baseline injury characteristics (baseline NDI, baseline Numeric Pain Rating Scale, previous occurrence of symptoms, duration of symptoms, whether the injury was work related, Fear Avoidance Belief Questionnaire score (FABQ) (2 and 4 week), bothersome symptoms (pain, numbness), treatment characteristics (treatment group, number of treatments) and design variables (time of measurement). The Numeric Pain Rating Scale is a measure of pain ranging in $[0,10]$, with higher scores indicative of more pain. As its name suggests, the FABQ score represents fear and avoidance and has a range in $[0,96]$. Only patients who have completed the full 4-week treatment regimen and have no missing values for any of the covariates were considered for this study. Of the original 81 patients who were randomized, 68 had no missing values and were included in the analysis. Forty patients (59%) were in the group treated with the neck traction and 28 (41%) received the sham traction treatment. A summary of all the variables, including NDI, can be seen in Table 2.1. Marginally, there are no significant differences between any of the treatment groups, although, nominally, the treatment group tended to have higher NDI and FABQ scores and more previous symptoms, indicating that this group may contain patients with more severe injuries.

Variable	Traction ($n=40$)	Sham	Overall ($n=68$)
		Traction ($n=28$)	
Age (y)	47.6 (9.0)	47.4 (9.8)	47.5 (9.3)
BMI (kg/m^2)	29.9 (6.3)	28.8 (6.0)	29.4 (6.2)
Sex (n,%)			
Male	12 (30)	8 (29)	20 (29.4)
Female	28 (70)	20 (71)	48 (70.6)
Previous Symptoms (n,%)			
Yes	11 (28)	12 (43)	23 (33.8)
No	29 (72)	16 (57)	45 (66.2)
Duration of symptoms (n,%)			
< 3 weeks	9 (20)	6 (21)	15 (22.1)
≥ 3 weeks	31 (80)	22 (79)	53 (77.9)
Work Related Injury (n,%)			
Yes	8 (20)	4 (14)	12 (17.7)
No	32 (80)	24 (86)	56 (82.3)
FABQ ¹ (2 week)	40.5 (22.9)	35.1 (19.8)	38.3 (21.7)
FABQ ¹ (4 week)	34.3 (23.0)	29.7 (19.3)	32.4 (21.5)
Baseline Pain ²	22.3 (10.3)	20.0 (9.0)	21.3 (9.8)
Bothersome Pain (n,%)			
Yes	32 (80)	24 (86)	56 (82.4)
No	8 (20)	4 (14)	12 (17.7)
Bothersome Numbness (n,%)			
Yes	11 (28)	7 (25)	18 (26.5)
No	29 (72)	21 (75)	50 (73.5)
Number of Treatments	7.0 (2.2)	6.8 (1.7)	
NDI ³ (Baseline)	19.9 (8.2)	18.0 (7.6)	19.2 (8.0)
NDI ³ (2 weeks)	15.9 (8.3)	13.7 (7.2)	15.0 (7.9)
NDI ³ (4 weeks)	13.5 (9.3)	11.9 (7.6)	12.8 (8.6)

Table 2.1: Summary of Variables in Cervical Radiculopathy Trial: ¹ Range of scores 0-96, higher levels of FABQ represent higher levels of fear avoidance: ² Range of scores 0-10, 0=No pain, 10=Worst Pain Imaginable: ³ Range of scores 0-50, higher levels of NDI represent more severe injuries

A reference cell parameterization was used for the categorical variables. Females, patients without pain as their most bothersome symptom, patients without numbness as their most bothersome symptom, patients without previous symptoms, patients without work related injuries, patients with symptom duration less than 3 weeks and the discharge time point were chosen as reference cells. Using the previously defined language, the model intercept, treatment group, time of measurement and the interaction of treatment group and time of measurement made up the hypothesis set so that inference is always be made on these variables. The selection set consisted of all of the other variables. It is assumed that the structure of the variance within an individual, \mathbf{V}_i , is known up to a constant. The matrix \mathbf{V}_i was taken to be the correlation matrix determined using maximum-likelihood estimation of the final model.

Prior to the model selection procedure being applied, each of the variables were screened using a LMM adjusting for the intra-subject variation and the time of the NDI measurement. For each of these models, the variance was fixed at 0.7286, which is the maximum-likelihood estimate of the model with only an intercept and time in the model. The results of the screening can be seen in Table 2.2. The FABQ scores, baseline pain index, having bothersome pain, and baseline NDI were all significantly related to the the 2 and 4-week post-treatment NDI score ($P < 0.05$). The treatment group was not significant at either of the time points.

A best-subsets strategy using the AIC was used to find a final model. After applying this procedure, a patient's age, sex, the duration of previous symptoms, and whether pain was a bothersome symptom were not included in the final model. Just as in the screening analysis, the elements of \mathbf{V} were assumed to be known using the maximum likelihood estimate of the final model (0.5238). The parameter estimates, standard errors (SEs), p -values from a hypothesis test that the parameter is null-valued and 95% CIs from this model can be seen in Table 2.2. Of the variables included in the final model, only baseline NDI, FABQ and the number of treatments were

significantly related to NDI. The patients having a more severe injury at baseline and higher fear avoidance beliefs were more likely to have a higher NDI at the 2 and 4 week follow-up time points while patients that had more visits to the physical rehabilitation service had better injury outcomes. None of the variables in the hypothesis test were significantly related to the NDI score. An omnibus test for no effect of the cervical traction, which was a compound test of both of the parameters corresponding to the treatment group and the treatment group and time interaction, had a p -value of 0.873, indicating a lack of evidence that cervical traction has an impact on the post-treatment NDI score. The results of this analysis agree with the study conclusions from *Young et al.* (2010) with respect to the effect of traction procedure.

Estimates of the effect size for each of the model effects are given in the last column of Table 2.2. These effect sizes are calculated using the z -statistic from the hypothesis test and provides an estimate of Cohen’s d effect size, calculated as $\frac{2 \cdot |z|}{\sqrt{df}}$ where z is the test statistic and df is the model degrees of freedom, which, for the cervical radiculopathy analysis presented, is $136 - 12 = 124$. The effect sizes are often difficult to interpret, but a standard metric for these type of effects sizes states that an value of $d < 0.3$ signifies a small effect, $0.3 \leq d < 0.8$ represents a medium effect and $d \geq 0.8$ represents a large effect, although these guidelines can vary by discipline.

Lastly, estimates and 95% CIs of the NDI scores at the 2 and 4 week follow up visits were computed separately for each treatment group. These quantities were calculated using the overall mean values shown in Table 2.1. The mean NDI score for the cervical traction group at 2 weeks post-treatment was found to be 14.8 (95% CI: 12.3, 17.2) while the control group had a mean score of 14.2 (95% CI: 11.7, 16.6), which has a difference in NDI of 0.6 (95% CI: -1.8, 3.0). At 4 weeks post-treatment, the mean NDI for the cervical traction group decreased to 13.3 (95% CI: 10.9, 15.8) while the mean NDI score for the control group was 13.2 (95% CI: 10.8, 15.6), for a mean difference of 0.1 (95% CI: -2.3, 2.5). The interpretation of these quantities are

similar to the results published by *Young et al.* (2010), although the magnitude for all of the mean values are nominally larger than those reported *Young et al.* (2010). The predicted differences are much smaller in the adjusted analysis presented here.

This study, along with the results from *Young et al.* (2010), give evidence that the addition of cervical traction may not be effective in treating cervical radiculopathy. Furthermore, the results from this were able to adjust for the longitudinal design of the study as well as the covariates included in the final analysis. Researchers could use these results to plan studies to evaluate other novel interventions for cervical radiculopathy. Consider a novel intervention that, when applied with the standard of care therapy, aims to reduce the NDI by 25% after 4 weeks compared to a similar group treated with the just the standard of care. Using the study discussed previously as a guide, the researchers would be trying to show that the novel therapy reduces NDI by 1.58 units and would assume a standard deviation of 4.92. Assuming a Type-I error of 0.05 and using a 2-sided t -test, this new study would require 154 patients per group to achieve 80% power. However, the simulation results indicate that the residual variance is typically underestimated, which would result in this study being under-powered.

2.5 Discussion & Conclusions

This chapter has summarized some of the issues that result from performing post-model selection inference and ignoring model uncertainty. Specifically, the literature cites many instances where variance estimates show downward bias and Type-I error rates are higher than the nominal level in cases where the response variable can be considered independent. Similarly, coverage probabilities are lower than the nominal value when the regression parameters are non-zero. A statistical representation of the model used in model selection was extended to the correlated data case.

A simulation study showed that the problems that are present when the response

variable can be considered independent persist when this assumption is no longer valid. Further simulations show that these results are robust against other information criteria (BIC), the use of the parsimony correction and other intraclass correlation values ($\rho=0.75$). When the variable of interest was non-zero, the coverage probabilities of 95% confidence intervals for the parameters in the hypothesis set were lower than the nominal value.

Overall, these results show an inflated effect size of each of the parameters, resulting in the increased expectation that a test of the parameter corresponding to the focus variable has a relationship with the response when, in truth, it may not be. This agrees with the argument put forth by *Ioannidis* (2008). Additionally, these results agree with the previously published literature using the linear regression model where the response variable can be considered independent. This is not a surprising result, as the linear regression model is a special case of the LMM.

The results from the simulation study are assumed to extend to other situations not discussed here. These include time-dependent covariates, or when the observations within a block vary, the inclusion of random effects and different covariance patterns. It is expected that the problems of post-model selection inference extend to other complex models such as longitudinal models and hierarchical models with more complex variance structures. It is also assumed that the problems of post-model selection inference extend to situations where model selection is performed on the covariance structure, either with or without model selection being performed on the regression parameters. These situations must be explored individually and confirming this extension remains an open topic of study so the generalizability of these results to other situations is only based on an educated conjecture.

The simulation study was limited by the fact that the form, or structure, of the covariance matrix was considered known. This choice was made so because it is beyond the scope of this work to assess how the covariance parameters behave

in the presence of model uncertainty. However, the effect of misspecification of the intraclass correlation value was observed. In practice, this value is estimated and not known. It is unknown how the data driven model selection procedures impact the estimation of the parameters belonging to the covariance structures. However, even for reasonable estimates of these parameters, this study has shown that there are most likely serious consequences of using post-model selection inference.

The results from a trial on patients with cervical radiculopathy were extended from *Young et al.* (2010) to adjust for potential differences in several predictor variables. These predictor variables were subjected to a model selection procedure to potentially reduce the effects of biased parameter estimates from a misspecified model. The model suggested from the model selection procedure will be used in an evaluation of future methodological work. This model was also used to perform a power analysis for a future study. Since this power analysis was based upon the post-model selection standard deviation which, according to the simulation study, it likely underestimates the true standard deviation and does not take into account model uncertainty.

As mentioned previously, little work has been done in this area to propose methods to account for model uncertainty in linear mixed-effect models. For LMMs, one potential solution is to extend *Ye* (1998) by using the generalised degrees of freedom for linear mixed-effect models calculated by *Zhang et al.* (2012) to achieve a better estimate of the variance will be presented in the following chapter. *Shen et al.* (2004a) is extended to inflate the covariance matrix of the fixed-effect parameter estimates to achieve tests that are closer to the nominal level. These results will allow statisticians and researchers to be able to incorporate some degree of model uncertainty into their results.

Not only are the results from a study using model selection strategies impacted by the previously discussed biases, but future studies can also be affected. In practice, the variation of the response variable is the most difficult thing to quantify, both for

clinicians and statisticians. Using previously published results is an efficient way of estimating this variation. However, if the estimates of the variation are underestimated in the published literature, and these results are used in power analysis, the variability seen in the study will be higher than expected. This could result in a study that may not have sufficient power to achieve its goal, leading to the shelving of a promising treatment. As the simulations show, the true variance is generally higher than the post-model selection estimate. This means that the hypothetical study presented will most likely require more than the 308 total patients to achieve the power the 80% power that is expected.

Variable	Marginal Model			Post-Model Selection		
	Estimate	95% CI	Estimate	SE	p-value	95% CI
Intercept			-0.20	3.54	0.955	(-7.13, 6.74)
Time			-0.97	0.92	0.292	(-2.77, 0.83)
Treatment Group	-1.94	(-1.75, 5.64)	0.59	1.24	0.632	(-1.84, 3.03)
Treatment Group×Time			-0.49	1.18	0.681	(-2.81, 1.83)
Age	-0.16	(-0.35, 0.04)				
BMI	0.21	(-0.08, 0.51)	0.07	0.09	0.440	(-0.11, 0.24)
Sex	-1.32	(-5.32, 2.68)				
Previous Symptoms	-0.10	(-3.96, 3.76)	1.21	1.17	0.299	(-1.08, 3.51)
Duration of Symptoms	1.15	(-3.25, 5.55)				
FABQ	0.22	(0.16, 0.27)	0.15	0.03	< 0.001	(0.10, 0.21)
Work Related Injury	4.04	(-0.67, 8.75)	-0.55	1.50	0.715	(-3.50, 2.40)
Baseline Pain	0.32	(0.14, 0.49)	0.06	0.06	0.278	(-0.05, 0.18)
Bothersome Pain	5.60	(0.97, 10.2)				
Bothersome Numbness	-2.26	(-6.37, 1.85)	-2.01	1.25	0.108	(-4.45, 0.43)
Number of Treatments	-0.29	(-1.20, 0.63)	-0.60	0.27	0.025	(-1.13, -0.08)
Baseline NDI	0.73	(0.56, 0.89)	0.52	0.07	< 0.001	(0.37, 0.66)
Residual Variance			24.24			(19.36, 31.24)

Table 2.2: Screening and multivariable analysis of cervical radiculopathy data. This table includes a marginal model that is adjusted for the intra-patient variation as well as the time point that NDI is measured. Effects with a p -value smaller than 0.05 are indicated by an asterisk (*)

CHAPTER III

Adjusting for Model Uncertainty in Variance Estimation

In order to adjust the parameter estimates for model uncertainty, a theoretical view of how these quantities behave must be explored. This chapter will propose a model selection paradigm to allow the behavior of both the fixed-effect and residual variance estimates to be studied. Following this exploration, the concept of generalized degrees of freedom (GDF) and its application to estimating the variance will be introduced. The concept of GDFs was originally proposed by *Ye* (1998) for adjusting linear regression models for model uncertainty. *Zhang et al.* (2012) extended this concept to LMMs, but used a different derivation for its use in adaptive model selection. This method is adapted for calculating the post-model selection estimate of the residual variance and will be evaluated with extended results from the simulation initially shown in Section 2.3. The application of this technique to the cervical radiculopathy trial will be presented and, finally, the conclusions and limitations of using the GDF in variance estimation will be discussed.

Some of the notation in this section was originally defined and used by the aforementioned authors. For clarity purposes, an effort was made to be as consistent as possible with these works.

3.1 Properties of Post-Model Selection Estimators

As the simulation study in the previous chapter demonstrated, there are issues with using the same data to select and make valid statistical inference. This strategy often results to overfitting, leading to biased estimates of the residual variance and undesirable properties of the hypothesis tests. More importantly, this strategy ignores the uncertainty in the underlying model that initiated the use of the model selection strategy. In order to adjust for model uncertainty, it is first necessary to understand the behavior of the estimates when the underlying model is uncertain.

The selection set in (2.6) can be further partitioned as $\mathbf{X}_2 = (\mathbf{X}_{2,M_o}, \mathbf{X}_{2,M_o^C})$, so that \mathbf{X}_{2,M_o} represents the variables corresponding to parameters with non-zero values and \mathbf{X}_{2,M_o^C} contains the variables corresponding to null-valued parameters. The parameter vectors for the partitioned parameter matrix are β_{2,M_o} and β_{2,M_o^C} , respectively. It is assumed that the model selection procedure will either restrict the models considered to only those containing $(\mathbf{X}_1, \mathbf{X}_{2,M_o})$, or will include $(\mathbf{X}_1, \mathbf{X}_{2,M_o})$ nearing probability one. This paradigm restricts the scope of model selection to only cases where the variables corresponding to non-null parameters are included in the final estimated model. Consistent and efficient model selection, present in the BIC and AIC, respectively, can be thought of examples of such a strategy as the probability of achieving an optimal model converges to unity as the sample size increases. This paradigm allows an investigation into the properties of the post-model selection estimators.

To illustrate this concept, consider a model where $\mathbf{Y} = \beta_o + \mathbf{x}_1 \cdot \beta_1 + \epsilon$ is the true, but unknown, model. In addition to the vector \mathbf{x}_1 , the vector \mathbf{x}_2 is considered a potential explanatory variable that has no relationship with the response, thus having a regression parameter of 0. In this example, the vector of ones corresponding to the intercept is the hypothesis set, \mathbf{X}_{2,M_o^C} consists of \mathbf{x}_2 and \mathbf{X}_{2,M_o} consists of \mathbf{x}_1 . The aforementioned restriction will prohibit the model $\mathbf{Y} = \beta_o + \mathbf{x}_2 \cdot \beta + \epsilon$ or an

intercept-only model from being considered since it does not contain \mathbf{x}_1 . The model $\mathbf{Y} = \beta_o + \mathbf{x}_1 \cdot \beta_1 + \mathbf{x}_2 \cdot \beta_2 + \epsilon$ is considered only if the estimation method produces an unbiased estimator of the parameters. It goes without saying that the true model is considered in this paradigm. With this paradigm, M_o will be referred to as the true model, understanding that this is a set of models. Its complement is referred to as M_o^C .

Theorem 3.1 discusses how the use of a model selection procedure affects the fixed-effect estimates:

Theorem 3.1. *In the setting of (2.5), (2.6) and the previous partitioning of both the selection set and the model space, let $\hat{\beta}_{\hat{M}}$ be the generalised least squares estimator on the recommended final model from the model selection procedure. Then, $E(\lim_{n \rightarrow \infty} \hat{\beta}_{\hat{M}}) = \beta$. Furthermore, $E(\hat{\beta}_{\hat{M}}) = \beta$ if $\mathbf{X}_{2,M_o^C} = \mathbf{0}$.*

This result is expected due to the consistent or efficient nature of the model selection procedure. However, it is important because if Theorem 3.1 is not true and the resulting expectation of the fixed-effects parameter estimates were biased, then the generalised least squares estimate of the residual variance is inaccurate. The generalised least squares estimate of the residual variance is the mean square error (MSE) of the predicted values and it is well known that the MSE degenerates to a variance and squared bias term. Thus, if the parameter estimates, and hence predicted values, have a bias that was non-zero, a biased estimate of the residual variance would be obtained. However, since the fixed-effect parameter estimates are unbiased, the residual variance has clear meaning.

The order of the expectation operator and limit in the result of Theorem 3.1 may be transposed depending on how the model uncertainty is quantified. Transposing these quantities makes the limit superfluous, so the fixed-effect estimates can be unbiased regardless of the sample size, as opposed to asymptotically unbiased. However, this is due to the particular paradigm of accounting for model selection that

was discussed. Theorem 3.1 is true for any consistent or efficient model selection procedure. Further information regarding this derivation can be found in Appendix B.1. Lemma 3.2 adds to the behavior of the post-model selection estimates of the fixed-effects.

Lemma 3.2. *Under the conditions of Theorem 3.1, the generalized least squares estimate of the fixed-effects after model selection, $\hat{\boldsymbol{\beta}}_{\hat{M}}$, has an asymptotic multivariate normal distribution with mean $\boldsymbol{\beta}$ and variance $\boldsymbol{\Sigma}$.*

Lemma 3.2 follows directly from Theorem 3.1. The proofs for Theorem 3.1 and Lemma 3.2 are straightforward extensions of *Shen et al. (2004a)* and hinge on the assumption that the variance \mathbf{V} is known. If the response vector, design matrices and error term are pre-multiplied by $\mathbf{V}^{-\frac{1}{2}}$, where $\mathbf{V}^{-1} = \mathbf{V}^{-\frac{1}{2}} \cdot \mathbf{V}^{-\frac{1}{2}}$, then the conditions found in *Shen et al. (2004a)* are met. More details of this work can be found in Appendix B.1.

Lemma 3.2 has two distinct implications. First and foremost, the results of this lemma allow for inference and construction of confidence intervals using the generalised least square estimates and standard multivariate distribution theory. Second, it conflicts with work from others that claim the post-model selection distribution of the parameter estimates is not multivariate normal and cannot be known. However, these works focus on model selection procedures that allow for consideration of all models. Only a set of the true models is considered in the framework introduced previously, so that the asymptotic normality that results from Lemma 3.2 holds.

Even though the fixed effect estimators are asymptotically unbiased, not all parameter estimates in the LMM have this property. As Theorem 3.3 states, the estimator of the residual variance is not unbiased.

Theorem 3.3. *After a model selection procedure performed under the conditions found in Theorem 3.1, the expected value of the estimate of the residual variance*

estimated by the generalised least squares estimator will be no less than the expectation of the residual variance estimate if the true model is known. That is, $E(\hat{\sigma}_M^2) \geq E(\hat{\sigma}_{M_o}^2) = \sigma^2$. If the true model is known, or if there are no covariates that are unrelated to the response, then the generalised least squares estimate will be unbiased. This is stated as: $\mathbf{X}_{2,MC} = \mathbf{0}$, then $E(\hat{\sigma}_M^2) = E(\hat{\sigma}_{M_o}^2) = \sigma^2$.

Theorem 3.3 states that even though the expectation of the fixed-effects estimates are unbiased, the expectation of the variance does not necessarily retain this quality. From a philosophical point of view, Theorem 3.3 says that model uncertainty has a cost in that the results derived from a model that is not known to be true are less certain than a model that is known to be correct. In both linear regression models and LMMs, this uncertainty is captured in the parameter σ^2 . As mentioned previously, researchers rarely know the true model but often make inference and construct confidence intervals as though the true model is known. In view of Theorem 3.3, these inferences and confidence intervals are biased so that they reflect smaller variability. Furthermore, the simulation results in the previous chapter show that $\hat{\sigma}^2$ underestimates σ^2 due to overfitting. This bias could be eliminated if the results were adjusted for model uncertainty. The corrected variance estimate could potentially be an unbiased estimator for the variance if the underfitting was adjusted appropriately. The proof of Theorem 3.3 can be found in Appendix B.2.

3.2 Generalized Degrees of Freedom

The concept of generalized degrees of freedom (GDF) was first developed by Ye (1998), although similar concepts were discussed by Efron (1986) and Girard (1989). With a similar motive as this work, the goal of Ye (1998) was to account for model selection in the estimate of the residual variance σ^2 in a linear regression model. This work used a fortuitous relationship between the derivative of the predicted values with

respect to the responses and the projection, or “hat” matrix. Using this definition, this work was able to define the GDF as “...the sum of the *average* sensitivities of the fitted value $\hat{\mu}_i(\mathbf{Y})$ to a small change in y_i . Thus it measures the flexibility of the modeling procedure M .” This definition was used in the linear regression setting to produce an estimate of σ^2 adjusting for model uncertainty and was accomplished by replacing the model degrees of freedom p with the value of the GDF. Since the GDF is typically larger than the number of parameters, this resulted in an inflated estimate of the residual variance that is indicative of the result from Theorem 3.3.

As stated in Ye (1998), the definition of GDF is advantageous for two reasons. First, flexible models, or models whose predicted values would not be sensitive to small changes in the response variable, would have relatively small GDF and models that were very sensitive to small changes in the response would have large values of the GDF. Thus, the GDF is a measure of the flexibility of the modeling procedure and of model fit. Inflexible, poor fitting models would result in high values of the GDF and the confidence intervals resulting from these models would be much wider than the similar intervals for flexible models. The other major advantage is that, through distribution theory, a simple relationship between certain model quantities and the GDF could allow for conceptually simple estimation methods. The estimation method developed by Ye (1998) is known as the *perturbation method*. An explanation and extension of this estimation method will be discussed in the following section.

The GDF defined by Zhang *et al.* (2012) was meant to minimize the Kullback-Leibler distance through optimizing the penalty term. This differs from the formulation proposed by Ye (1998) that relates the GDF to the derivative of the predicted values with respect to the observed responses, which is stated as:

$$\mathcal{D}^{\mathcal{M}} = \sum_{i,j} h_{ij}^{\mathcal{M}} = \sum_{i,j} \frac{\partial E(\hat{\mu}_{ij})}{\partial y_{ij}} \quad (3.1)$$

where $\mathcal{D}^{\mathcal{M}}$ is the GDF value and the quantity $h_{ij}^{\mathcal{M}}$ will be referred to as the *generalized*

leverage for the j^{th} observation from block i . For the generalised least squares estimators, the predicted values $\hat{\mu}_{ij}$ and the residual variance estimate $\hat{\sigma}^2$ are both functions of the response vector \mathbf{Y} . Except in a some special case scenarios, the predicted response is a function of the unknown elements in \mathbf{V} , which in turn, are functions of \mathbf{Y} . The relationship between a single response and potentially many elements of \mathbf{V} is a monumental task that will be dependent on the estimation method for the elements of \mathbf{V}_i as well as the structure of \mathbf{V}_i . Rather than assess the predicted values with respect to the observed responses, consider a similar definition of the GDF that assesses the predicted values with respect to the unknown true value μ_{ij} . Then, Definition 3.4 can be used as an extension of the GDF proposed by Ye (1998) for LMMs.

Definition 3.4. For LMMs, the generalized degrees of freedom (GDF) of the model can be represented as:

$$\mathcal{D}^{\mathcal{M}} = \sum_{i,j} h_{ij}^{\mathcal{M}} = \sum_{i,j} \frac{\partial \text{E}(\hat{\mu}_{ij})}{\partial \mu_{ij}} \quad (3.2)$$

where $h_{ij}^{\mathcal{M}}$ is the diagonal element, or leverage, of the hat matrix for the j^{th} observed value from block i and y_{ij} is the value of the corresponding observed response variable. The quantity $\hat{\mu}_{ij}$ is the predicted value using the generalised least squares estimator of the fixed-effect parameters.

Rather than keep the GDF in terms of the notation in Definition 3.4, the formulation of the LMM can be used for the predicted values to achieve a quantity enables a more efficient estimation. In the LMM discussed in Chapter 2, Definition 3.4 results in Theorem 3.5:

Theorem 3.5. The GDF for the LMM model is defined by $\mathcal{D}^{\mathcal{M}} = \sum_{i,j} h_{ij}^{\mathcal{M}}$, where:

$$h_{ij}^{\mathcal{M}} = \frac{1}{\sigma^2} \text{E} \left(\hat{\mu}_{ij} \cdot (\mathbf{Y}_i - \boldsymbol{\mu}_i)' \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j \right) = \frac{1}{\sigma^2} \text{cov} \left(\hat{\mu}_{ij}, (\mathbf{Y}_i - \boldsymbol{\mu}_i)' \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j \right)$$

with $\boldsymbol{\mu}_i = \mathbb{E}(\mathbf{Y}_i)$ and \mathbf{e}_j as the j^{th} column of the identity matrix with dimension n_i .

The proof of Theorem 3.5 could be extended to covariance pattern models by pre-multiplying the LMM equation by $\mathbf{V}^{-\frac{1}{2}}$ and proceeding with already established proofs. However, neither Ye (1998) or Shen *et al.* (2004a) offer much detail on these derivations. A more detailed explanation can be found in Appendix B.3. Theorem 3.5 can be shown to degenerate to the GDF defined by Ye (1998) when the response variable can be considered independent.

The concept of GDF was further developed by Efron (2004) and Weisberg (2005), primarily for use in adaptive model selection. Adaptive model selection is related to the information criteria discussed in the previous section, however, it assumes that the optimal penalty c is unknown and to be estimated from the data. This leads to the *adaptive model selection criterion*, $-\log L + \hat{c} \cdot K$, where, as previously defined, $\log L$ is the logarithm of the likelihood function and K is the number of parameters in the model. Zhang *et al.* (2012) extended the definition of the GDF to estimate \hat{c} in the LMM and finds that when used as a model selection strategy, the resulting models have smaller Kullback-Leibler distance than AIC or BIC, meaning that the predictions are closer to the response variable. Further, the adaptive selection procedure generally is as good as or better than the commonly used methods both in terms of the number of correctly and incorrectly entered variables. The estimate of the GDF calculated by Zhang *et al.* (2012) degenerate to the GDF computed by Ye (1998) when independence is assumed across all observations.

In a similar fashion to Theorem 3.3, the generalized leverage used by Zhang *et al.* (2012) is:

$$\sum_k h_{ijk}^{\mathcal{M}}(Z) = \sum_k \text{cov}(\hat{\pi}_{jk} \mu_{ij}, Y_{ik}) - \frac{1}{2} \text{cov}(\hat{\pi}_{jk}, Y_{ij} Y_{ik}) \quad (3.3)$$

where π_{ijk} is the $(j, k)^{th}$ element of $\frac{1}{\sigma^2} \mathbf{V}^{-1}$. Notice that the first term in (3.3) bears a remarkable resemblance to the latter term in Theorem 3.5. In fact, if both σ^2 is

replaced by an estimate $\hat{\sigma}^2$ in Theorem 3.5 and μ_{ij} is replaced by an estimate in (3.3), then $\hat{\sigma}^{-2} \text{cov}(\hat{\mu}_{ij}, (\mathbf{Y}_i - \boldsymbol{\mu}_i)' \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j) = \sum_k \text{cov}(\hat{\pi}_{jk} \hat{\mu}_{ij}, Y_{ik})$. The first term in (3.3) acts as a measure of how close the predicted values are to the observed responses. Under the same reasoning, the second term acts as a measure of how close the predicted variance between two observations within a block are to an empirical estimate. Thus, the GDF proposed by *Zhang et al.* (2012) incorporates both the fixed- and random-effects in the estimation of the GDF. Because of this relationship between the predicted values and the additional information added on from the comparing the covariance terms, the GDF used for the LMM will be the one derived by *Zhang et al.* (2012). This will allow for an unknown σ^2 to be used in the calculation of the GDF while restricting \mathbf{V} to be known.

3.3 Estimation of the GDF

As defined in (3.3), the GDF cannot be calculated since the quantity is a function of the unknown parameter vector $\boldsymbol{\mu}$. The GDF developed for the linear regression models by *Ye* (1998) uses a data perturbation technique to estimate the generalized degrees of freedom. *Shen et al.* (2004a) uses a similar method to adjust the covariance matrix of the parameter estimates in linear regression for model uncertainty. *Shen and Huang* (2006) and *Efron* (2004) use data perturbation to find optimal penalty terms to minimize loss in model selection. *Shen et al.* (2004b) extended the concept of generalized degrees of freedoms to the exponential family in general, but concentrate on binomial and Poisson models. *Zhang et al.* (2012) use data perturbation to estimate the GDF of LMMs and include the case where the structure of \mathbf{V} is unknown and estimated. This section will give an overview of this procedure while the details can be found in the aforementioned manuscripts.

The perturbation method requires creating what *Shen et al.* (2004a) refer to as a

“pseudo”-response variable, defined as:

$$\mathbf{Y}_i^* = \mathbf{Y}_i + \tau \cdot \sigma_* \cdot \tilde{\mathbf{Y}}_i \quad (3.4)$$

where \mathbf{Y}_i is the response vector defined in (2.4), $\tilde{\mathbf{Y}}_i$ has a multivariate normal distribution with null mean and variance \mathbf{V} , and σ_* and τ are constants with τ known as the *perturbation size*. The perturbation size is bounded in $(0,1]$, however, recommended values of this parameter are in the range $[0.5,0.8]$. It is customary to set $\sigma_* = \sigma$, but this is not a necessary requirement. The pseudo-response variable defined in (3.4) differs from *Shen et al.* (2004a) and *Zhang et al.* (2012), however, it is straightforward to show that these response variables are mathematically equivalent.

The estimation of the GDF is based on the pseudo-response variable and, rather than use this variable as calculated, this variable is conditioned on the observed responses. Conditioned on these data, the pseudo-response variable \mathbf{Y}_i^* has a multivariate normal distribution with mean \mathbf{Y}_i and variance $(\tau \cdot \sigma_*)^2 \cdot \mathbf{V}_i$. *Zhang et al.* (2012) show that the estimate of the generalized leverage can be written as:

$$\sum_k (\tau)^{-2} \text{cov}^*(\hat{\mu}_{ij} \hat{\pi}_{ijk}, Y_{ik}^*) - \frac{1}{2} \mathbf{V}_i(j, k) \text{var}^*(Y_{ij}^* Y_{ik}^*) \text{cov}^*(\hat{\pi}_{ijk}, Y_{ij} Y_{ik}).$$

Any operator followed by an asterisk (*) will mean that the operator is conditional on the observed responses, so that $\text{cov}^*(\cdot, \cdot) = \text{cov}(\cdot, \cdot | \mathbf{Y})$. The estimate of the total generalized degrees of freedom is the summation of the estimated generalized leverages over all blocks.

A Monte Carlo simulation is used to calculate $\hat{h}_{ij}^{\mathcal{M}}$ in order to be robust against the choice of \mathbf{Y}^* . this necessitates simulating a large number, D , of pseudo-response variables Using (3.4), a vector $\mathbf{Y}^{*d} = (\mathbf{Y}_1^{*d}, \mathbf{Y}_2^{*d}, \dots, \mathbf{Y}_N^{*d})'$ for $d=(1,2, \dots, D)$ is simulated. These pseudo-response variables are used to create predicted values $\hat{\boldsymbol{\mu}}^d(\mathbf{Y}^{*d}) = (\hat{\boldsymbol{\mu}}_1^d(\mathbf{Y}^{*d}), \hat{\boldsymbol{\mu}}_2^d(\mathbf{Y}^{*d}), \dots, \hat{\boldsymbol{\mu}}_N^d(\mathbf{Y}^{*d}))'$ using the generalised least squares estimators. These

predicted quantities are written so that it is clear they are functions of the psuedo-response variable, and similar notation will be used for the quantities $\hat{\pi}_{ijk}(\mathbf{Y}^{*d})$. After all Monte Carlo simulations are complete, the estimate of the GDF, $\hat{\mathcal{D}}^{\mathcal{M}}$, can be computed by summing the average estimate of variance and covariance terms over all blocks. This is expressed as:

$$\begin{aligned} \sum_i \sum_{j,k} (D \cdot \tau^2 \cdot \sigma_*^2)^{-1} & \left[\hat{\pi}_{ijk}(\mathbf{Y}^{*d}) \hat{\mu}_{ij}(\mathbf{Y}^{*d}) - \frac{1}{D} \sum_d \hat{\pi}_{ijk}(\mathbf{Y}^{*d}) \hat{\mu}_{ij}(\mathbf{Y}^{*d}) \right] \left[Y_{ik}^{*d} - \frac{1}{D} \sum_d Y_{ik}^{*d} \right] \\ & - \frac{\mathbf{V}_i(j, k)}{D^2} \left[\hat{\pi}_{ijk}(\mathbf{Y}^{*d}) - \frac{1}{D} \sum_d \hat{\pi}_{ijk}(\mathbf{Y}^{*d}) \right] \left[Y_{ij}^{*d} Y_{ik}^{*d} - \frac{1}{D} \sum_d Y_{ij}^{*d} Y_{ik}^{*d} \right]^2 \quad (3.5) \end{aligned}$$

The estimate in (3.5) is the same estimate used by *Zhang et al.* (2012), although the notation has been changed slightly and some simplification has been performed. Further simplification can be made if the \mathbf{V} can be considered known, but to keep the estimator of the GDF as general as possible, the estimator is presented as shown. The Law of Large Numbers stipulates that as the number of simulated pseudo-response vectors goes to infinity, the estimated generalized degrees of freedom will converge to the true generalized degrees of freedom (*e.g.* $D \rightarrow \infty, \hat{\mathcal{D}}^{\mathcal{M}} \rightarrow \mathcal{D}^{\mathcal{M}}$). To ensure convergence, it is recommended that D be larger than the total number of observations $N = \sum n_i$.

Many recommendations have been made for an initial value of σ^2 . *Ye* (1998) recommend an “iterative procedure”, but offer no details. Because of the potential complexity of this iterative strategy, it is not recommended. *Zhang et al.* (2012) uses the most complex model for the simulated response vectors, however this is not recommended when accounting for model uncertainty because it potentially uses two different models: one for estimation of the variance parameters and one for estimation of the fixed-effect parameters. These models may have important differences which

may not make sense, particularly in small samples. Furthermore, it makes little sense to use information from one model when the goal is to use a completely separate model for inference and estimation. The naïve variance estimate can always be used in the algorithm discussed above, however, in light of the results in the previous section, this will typically underestimate the true variance.

This work assumes that \mathbf{V} is known, however, this is not true in general situations. This causes the second term in (3.5) to be very small relative to the first term and contribute little to the overall GDF. Relaxing this assumption will result in larger GDF values.

3.4 Accounting for Model Uncertainty in Variance Estimation

Since it is assumed that both the structure and values of \mathbf{V} are known, the generalised least squares estimators can be used to estimate both the fixed-effect parameters, $\boldsymbol{\beta}$, and residual variance, σ^2 . The generalised least squares estimator of σ^2 after model selection can be represented by:

$$\hat{\sigma}_{adj, \hat{M}}^2 = \frac{(\mathbf{Y} - \hat{\mathbf{Y}}_{\hat{M}})' \mathbf{V}^{-1} (\mathbf{Y} - \hat{\mathbf{Y}}_{\hat{M}})}{N - p_{\hat{M}}} \quad (3.6)$$

where the subscript of the variance parameter estimator specifies that the estimator is adjusted for covariates and will henceforth be referenced as the *adjusted variance*. The vector $\hat{\mathbf{Y}}_{\hat{M}}$ are the predicted values from the final model with the parameter values estimated via the generalised least squares fixed-effect estimator for the hypothesis set and those variables from the selection set that were chosen to be included in the final model by the data-driven model selection procedure. The total number of estimated variables in the final model is $p_{\hat{M}}$, which is also the model degrees of freedom for the

final model, and the total number of observations N .

Equation (3.6) is valid only when the final model is known to be the true model. By virtue of using a model selection procedure, a true model is not known. To account for the model uncertainty, the following estimator for σ^2 is proposed:

$$\hat{\sigma}_{cor}^2 = \frac{(\mathbf{Y} - \hat{\mathbf{Y}}_{\hat{M}})' \mathbf{V}^{-1} (\mathbf{Y} - \hat{\mathbf{Y}}_{\hat{M}})}{N - \hat{\mathcal{D}}^{\mathcal{M}}} \quad (3.7)$$

so that the estimate of the GDF defined in (3.3) is used in place of the total number of model degrees of freedom. The subscript “ \hat{M} ” is superfluous in the notation as only the model recommended by the model selection procedure would be adjusted for model uncertainty. This estimator, which will be referred to as the *corrected variance*, will have a positive relationship to the GDF so that robust models will have smaller corrected variance and sensitive models will have large values of the corrected variance. Unlike the adjusted variance, the corrected variance is not constrained by linear models theory and, as such, may not have the properties that are expected or desired. For example, it is well known that $\hat{\sigma}_{adj}^2 \geq 0$ with equality only if the covariates form a perfect predictor. The corrected variance may be negative for very sensitive models if $\hat{\mathcal{D}}^{\mathcal{M}} > N$. Rather than to naïvely set the value of the corrected variance to some value, it is recommended that no action is performed as this is an indicator that this may be a poor fitting model.

3.5 Simulation Study

The simulation study discussed in the previous chapter was extended to confirm that the the estimated variance adjusting for model uncertainty using the GDF was larger than the naïve estimate when no adjustment was made. Two different values of σ_* were used. As a proof-of-concept, σ_* was set to a value of 1, which was the simulated value of σ . To mimic a real-life situation, the naïve standard deviation

estimated from the model recommended from either the model selection procedure was used in the algorithm. This will be referred to as the data driven method. For both the proof-of-concept and data driven methods, the perturbation size was set at 0.5 and the total number of perturbations simulated for each dataset was $D=200$. Just as in the previous section, many different simulation templates were used. All tended to tell a similar story, so that only a few are discussed in the text while the remainder can be found in Appendices C.1-C.4.

Prior to discussing the behavior of the corrected variance, the GDF is assessed in conjunction with the average dimension of the model at each value of the intraclass correlation, along with the dimension of the true model. The dimension of the true model is the number of non-zero parameters in the selection set along with the two parameters in the hypothesis set. The results of the extension of the simulation study for Cases 1 and 2 for the proof-of-concept simulation, along with sample sizes of 20 and 40, can be seen in Figure 3.1. As expected, the model selection procedures have a tendency to recommend overfit models, or models that include more variables than are truly needed for intraclass correlations around the true value for any combination of simulation parameters. Interestingly, as the intraclass correlation approaches the upper limit, it appears that this property disappears and the *correct sized* model is typically estimated. This does not mean the *correct* model is estimated, as this assessment is beyond the scope of this work. The average dimension of the final model converges to the expected value as the sample size increases, a result that is expected from the efficient or consistent nature of the model selection procedure (Figures 3.1a vs 3.1c, Figures 3.1b vs 3.1d).

The GDF calculated in the proof-of-concept simulation tended to be larger than the mean dimension of the model for all reasonable values of the intraclass correlation, regardless of the simulation parameters. This should result in the corrected variance being larger than adjusted variance for reasonable values of the intraclass

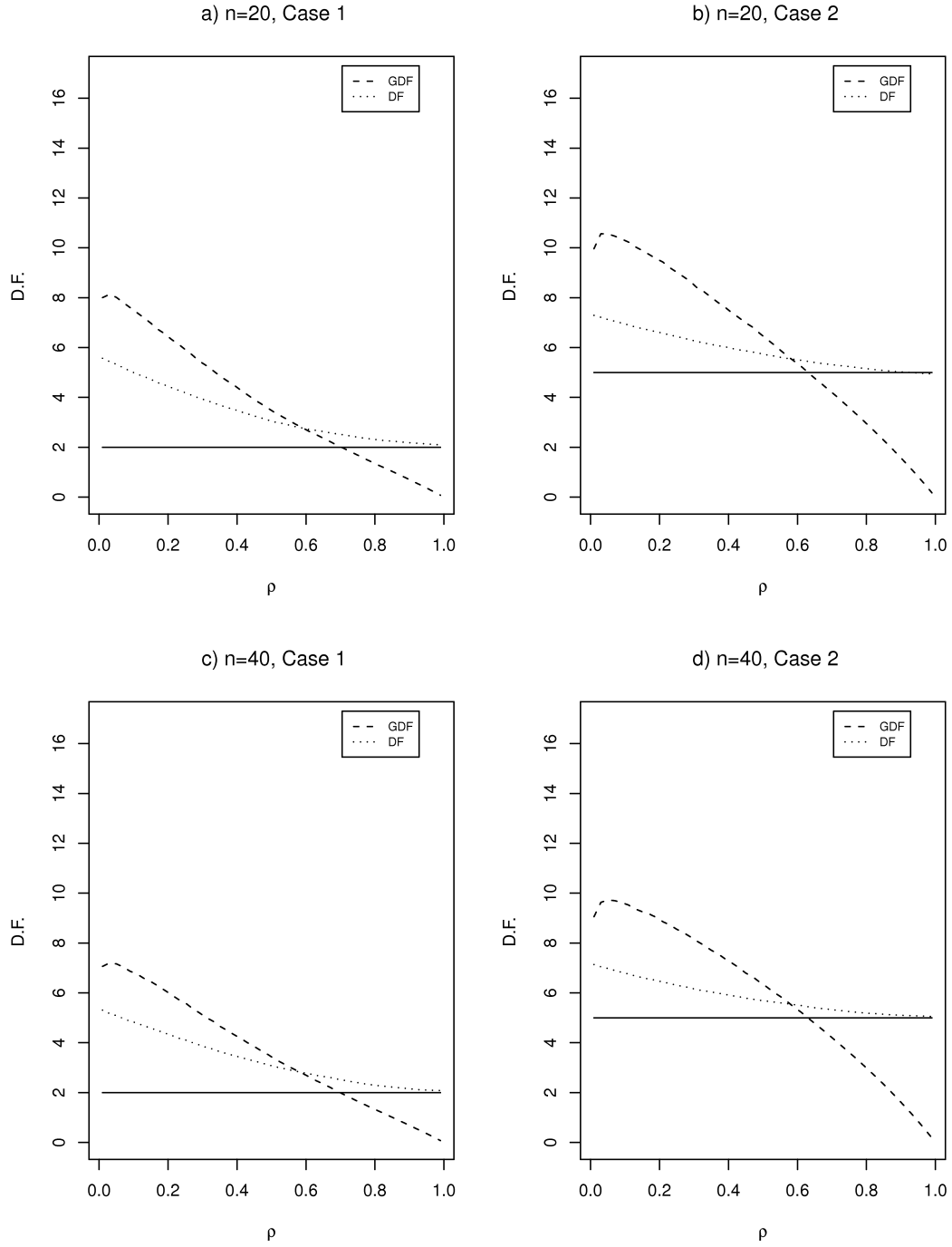


Figure 3.1: Proof-of-concept simulation results showing the GDF and average model DF ($\rho=0.25$). The solid line represents the expected dimension of the final model while the dotted and dashed lines represent the model and generalized degrees of freedom, respectively. As expected from overspecification, the model degrees of freedom are larger than the expected dimension, while the GDF inflates the model degrees of freedom.

correlation, leading to an expected result per Theorem 3.3. For intraclass correlation values much higher than the simulated value, the GDF can be smaller than both the mean and the expected dimension of the model. By (3.7), this should result in corrected variances smaller than the adjusted variances. Since these values of the intraclass correlation are unlikely to be used as an estimate, this phenomena can be ignored. As sample size increases, the mean GDF values do decrease (Figures 3.1a vs 3.1c, Figures 3.1b vs 3.1d), although this decrease is difficult to detect from the plots. There is also a slight concavity to the GDF profile over the intraclass correlation values with a maximum being achieved close to zero. These characteristics are robust to any of the simulation parameters explored.

In addition to comparing the mean values of the GDF and the model degrees of freedom, the tendency of the GDF of being larger than the model degrees of freedom was assessed. In a similar manner to the magnitudes of these quantities, the proportion of times the GDF was larger than the model degrees of freedom was plotted over the values of the intraclass correlation used for the analysis. The plots for the four simulation parameters for the proof-of-concept simulations can be seen in Figure 3.2. Overall, the GDF is higher than the model degrees of freedom for reasonable values of the intraclass correlation, although this proportion drops off precipitously as the intraclass correlation value approaches the upper limit. The proportion does not change as a function of the covariates (Figure 3.2a vs 3.2b, Figure 3.2c vs. 3.2d), but becomes closer to 1 over a broader range of intraclass correlation values as the sample size increases (Figure 3.2c, 3.2d).

Figure 3.3 shows the adjusted and corrected variances for the proof-of-concept simulations. For reasonable values of the intraclass correlation, the corrected variance estimate is larger than the adjusted estimate since the GDF are typically larger than the model degrees of freedom. This result agrees with Theorem 3.3. In all of the proof-of-concept simulations, $N \gg \hat{\mathcal{D}}^{\mathcal{M}} > p_{\hat{M}}$, so the impact of the GDF does

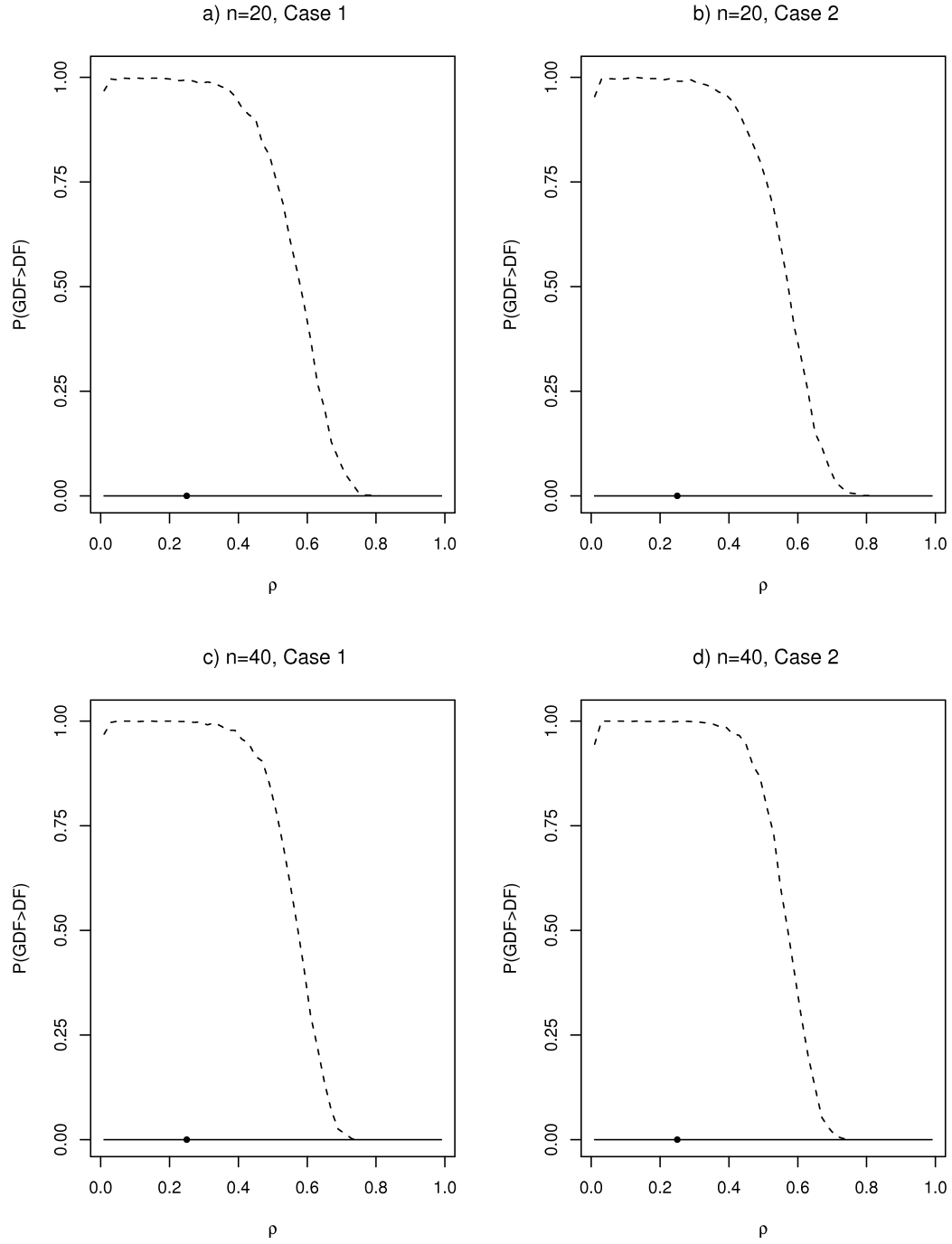


Figure 3.2: Probabilities that the GDF is larger than the model degrees of freedom ($\rho=0.25$). The dashed line is the probability and shows that the GDF almost certainly inflates the model degrees of freedom.

not have a large impact on the corrected variance, leading to the marginal increase of the corrected variance compared to the adjusted variance.

The average dimension and GDF calculated from the data-driven method is shown in Figure 3.4. Unlike the proof-of-concept method, the data-driven method is convex and approaches an upper limit at an intraclass correlation value of 0. When the true value of this parameter is 0.25, the GDF tends to be larger than the average dimension of the model recommended by the model selection procedure. Just as in the proof-of-concept simulations, the GDF decreases as the sample size increases (Figures 3.4a vs 3.4c, Figures 3.4b vs 3.4d).

When the value of the intraclass correlation increases to 0.75, the value of the GDF is not typically greater than the model degrees of freedom. Rather, the GDF is typically less than or equal to the model degrees of freedom, which will result in either no inflation of the residual variance or a corrected variance that is smaller than the adjusted variance. This characteristic limits usefulness of the GDF for general purposes. These results can be seen in Figure 3.5. The implications of this characteristic are discussed in the Discussion and Conclusions section of this chapter and the in the corresponding section in Chapter 4.

As expected, the average size of the GDF is relative to the model degrees of freedom is related to the difference in size shown in Figure 3.5. The proportion of times the GDF is larger than the model degrees of freedom can be seen in Figure 3.6 when the true value of the intraclass correlation is 0.75. As expected, for reasonable values of the intraclass correlation, the GDF is not always greater than the model degrees of freedom. In particular, the model degrees of freedom is typically larger than the GDF when the sample size is large (Figures 3.6c, 3.6d). This will limit the usefulness in correcting the estimate of the residual variance for model uncertainty.

The corrected variance calculated from the data-driven method behaves similar to that of the proof-of-concept for almost all of the intraclass correlation values except

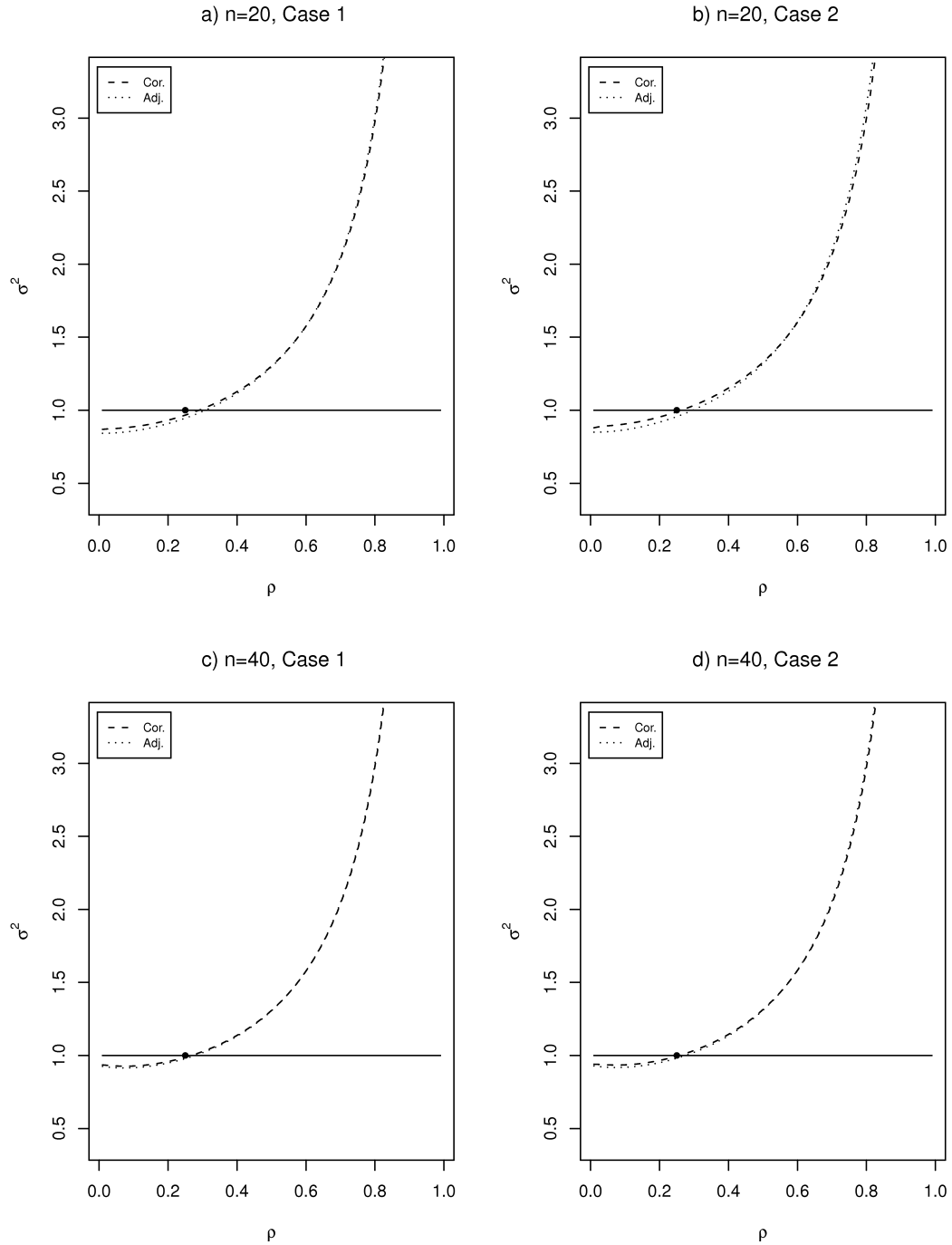


Figure 3.3: Data-driven simulation results showing the adjusted and corrected residual variance estimates ($\rho=0.25$). The solid line represents the true value of the residual variance while the dashed and dotted lines represent corrected and adjusted residual variance estimates, respectively. These values are close together and underestimate the true variance.

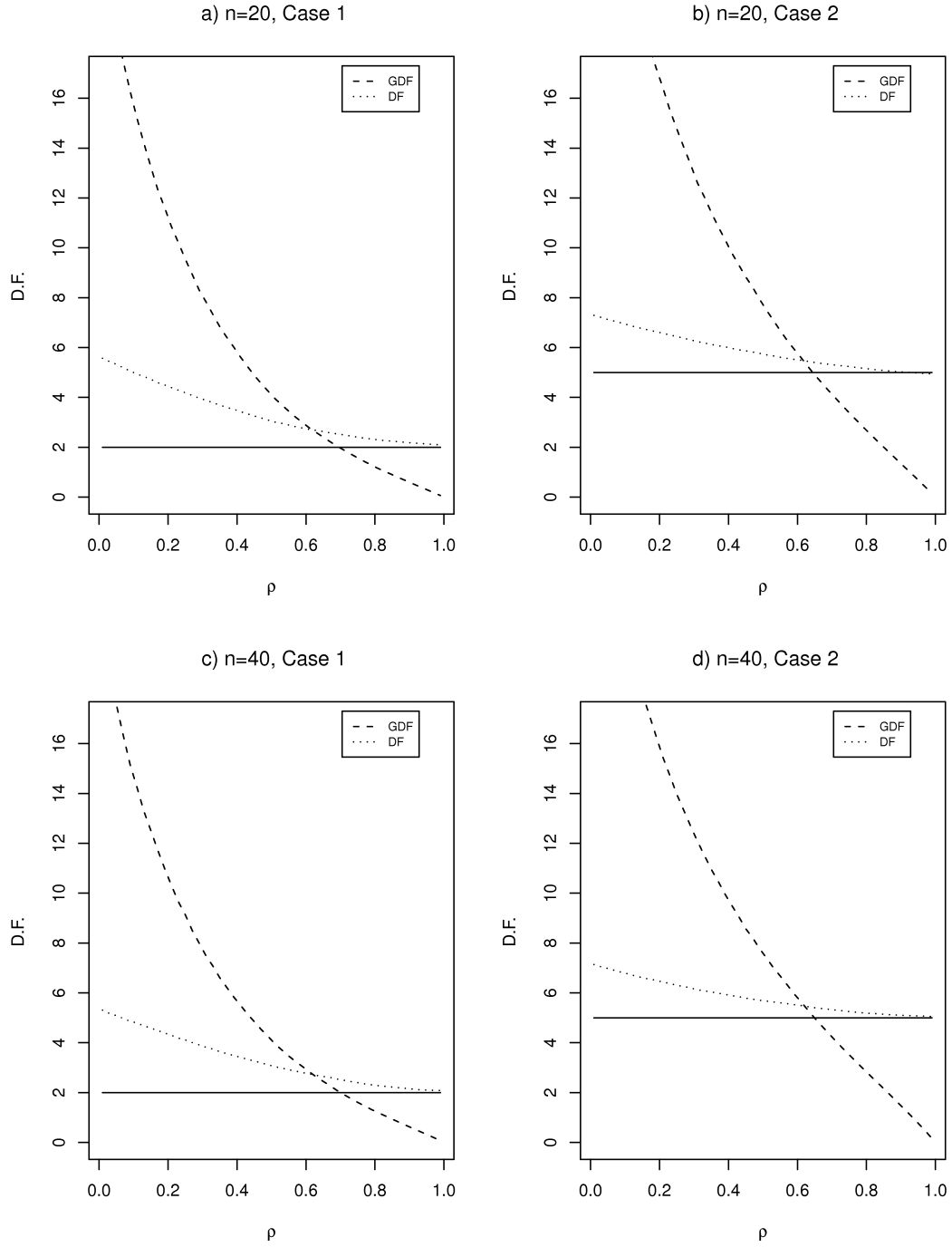


Figure 3.4: Data-driven simulation results showing the GDF and average model DF ($\rho=0.25$). The solid line represents the expected dimension of the final model while the dotted and dashed lines represent the model and generalized degrees of freedom, respectively.

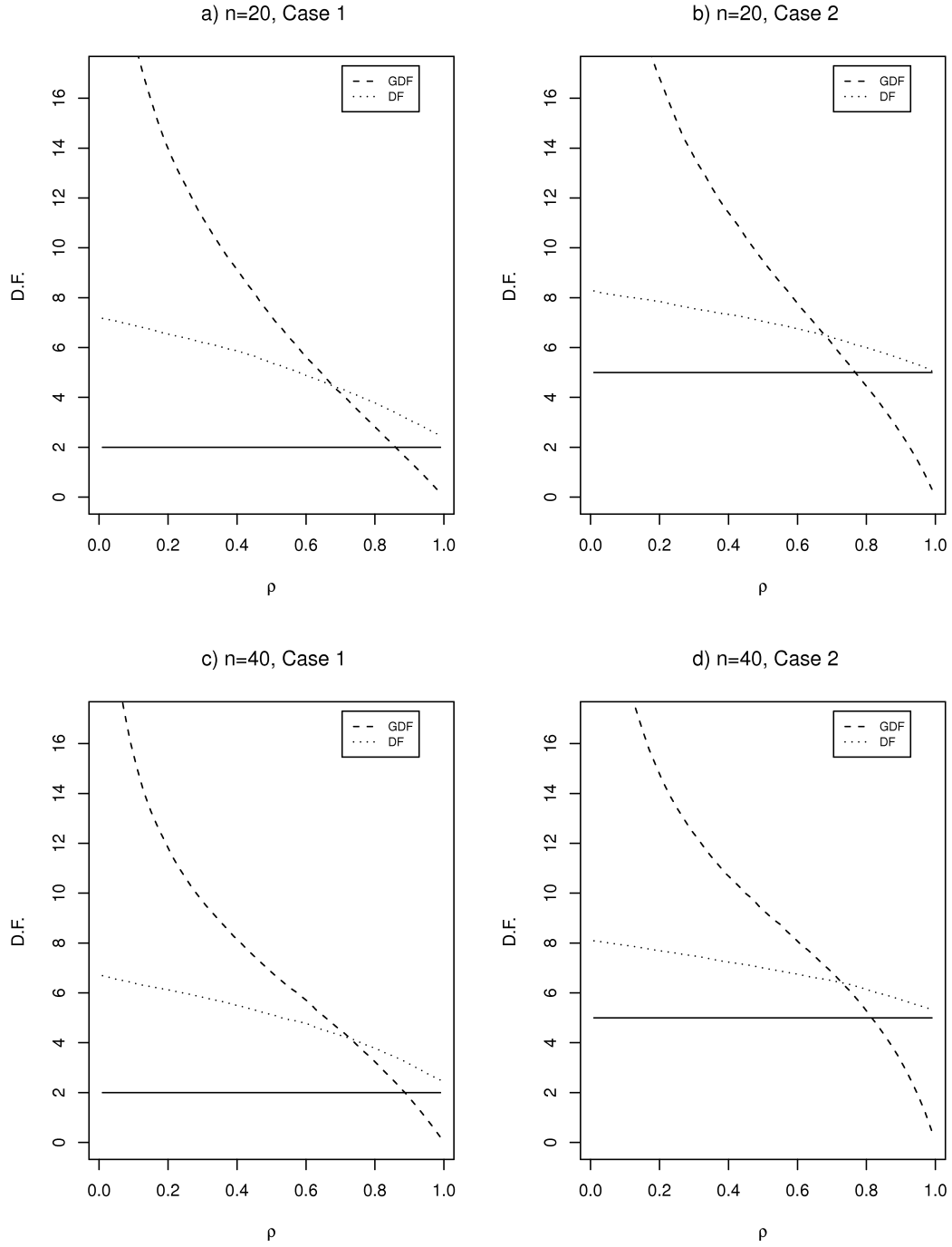


Figure 3.5: Data-driven simulation results showing the GDF and average model DF ($\rho=0.75$). The solid line represents the expected dimension of the final model while the dotted and dashed lines represent the model and generalized degrees of freedom, respectively.

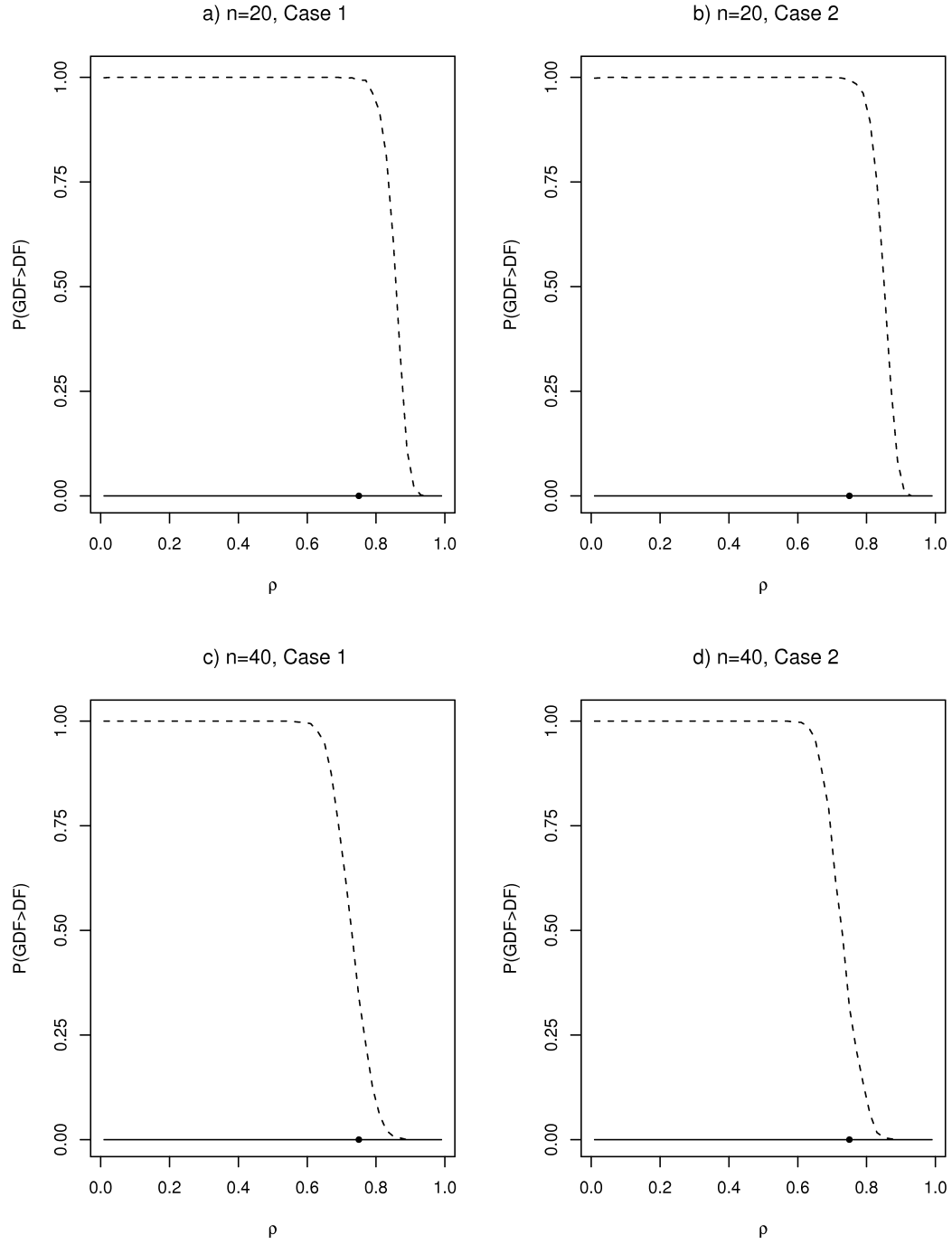


Figure 3.6: Probabilities that the GDF is larger than the model degrees of freedom ($\rho=0.75$). The dashed line is the probability and shows that the GDF is almost may not always to be larger than the model degrees of freedom.

when this value is close to zero. As the intraclass correlation used for the model approaches zero, the corrected variance increases so there is a distinct minimum. This characteristic occurs regardless of the value of the simulated intraclass correlation, but tends to mirror the adjusted variance closer as the sample size increases. This characteristic can be seen in Figure 3.7 and is further discussed in the Discussion and Conclusion section of this chapter.

The extended simulations for the data driven method, found in Appendix C.4, show that the corrected variances all resemble the profile in Figure 3.6, even if the GDF is smaller than the average degrees of freedom in the model. This is because the GDF has a lower bound of 0, so that as the total number of observations increase, the ratio of the corrected variance to the adjusted variance approaches unity $\left(\lim_{N \rightarrow \infty} \frac{N - \hat{\mathcal{D}}^{\mathcal{M}}}{N - p_{\hat{\mathcal{M}}}} = 1\right)$. Therefore, even though the the GDF performs poorly when the intraclass correlation is large, the corrected variance converges to the adjusted variance.

3.6 Variance Estimation in the Cervical Radiculopathy Trial

The GDF for the cervical radiculopathy trial was calculated to account for model uncertainty using the data-driven algorithm described previously. This means that σ_* was set to the observed standard deviation from the model discussed in Chapter 2, which is 4.92. The final model, was made up of twelve different columns corresponding to a model degrees of freedom of 12, has a GDF value of 34.39. Using this value, the corrected variance was found to be 29.46, which is larger than the adjusted variance of 24.24.

This trial was able to shed some light on the effect of a simulation parameter that was held constant in this work: σ^2 . The ratio between the GDF and dimension of the model from the cervical radiculopathy trial is 2.86, while the similar ratio computed from the simulation results in Figure 3.3 is 1.40. Thus, models with higher residual

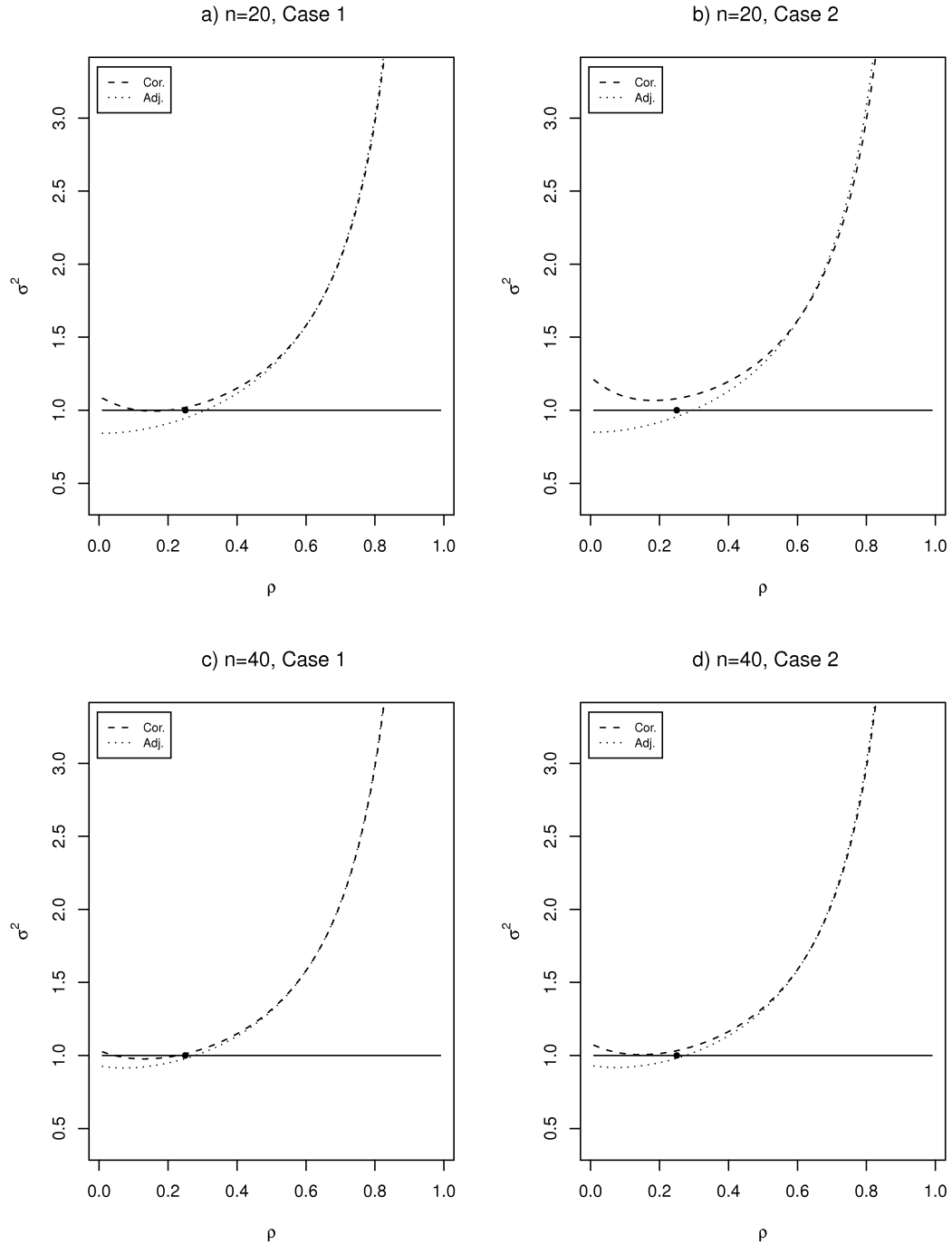


Figure 3.7: Data-driven simulation results showing the adjusted and corrected residual variance estimates ($\rho=0.25$). The solid line represents the true value of the residual variance while the dashed and dotted lines represent corrected and adjusted residual variance estimates, respectively. These lines are very close together and reflect the fact that the corrected variance does not inflate the variance by a larger amount.

variances will have larger GDF values. This is intuitive because the residual variance can be thought of as a measure of model fit since perfect fitting models will have a residual variance of 0. As this value increases, the predicted values become farther away from the observed values, indicating a poorer fit.

This information could be used to design a prospective study using a novel intervention outlined in the previous chapter. Without adjusting for model uncertainty, the study required 308 total patients, or 154 per group. A similar power analysis that incorporates model uncertainty through the corrected variance yields a recommended sample size of 372, or 186 per group. Adjusting for model uncertainty would require a considerable larger amount of patients that need to be enrolled into the study than one that ignores uncertainty in the model. Appropriately designed studies should incorporate this characteristic into power analysis.

3.7 Discussion & Conclusions

This chapter introduced the paradigm in which model selection occurred and which model uncertainty can be taken into account. This resulted in conclusions describing the asymptotic unbiasedness of the fixed-effect estimates as well as the asymptotic normality of this vector. Additionally, the expected value of the adjusted variance after model selection was found to be no less than the adjusted variance if the true model were known. Equality of this statement could be achieved if the true model is known, and in the absence of this condition, equality is not guaranteed.

A statistic referred to as the generalized degrees of freedom (GDF) was introduced and can be thought of as the average sensitivity of the predicted responses to a change in the observed response. An extension of *Ye (1998)* was used to show a derivation of the GDF for the LMM, but the work by *Zhang et al. (2012)* allowed for the use of a GDF that not only measured the flexibility of modelling the predicted values but also measured the flexibility in estimating the covariance parameters and was used

for adjusting the residual variance for model selection. The restriction on \mathbf{V} was kept not because of an inability to solve the GDF, but due to an uncertainty of how the elements of \mathbf{V} behave in the presence of model uncertainty.

The GDF was used with the generalised least squares estimate of the residual variance to produce an estimate of the this statistic that is adjusted for model uncertainty. The proof-of-concept simulations demonstrated that for reasonable values of the intraclass correlation, the GDF is larger than the model degrees of freedom, resulting in a corrected variance estimate that is larger than the estimate using the model degrees of freedom. However, since the total number of observations were much larger than either the model degrees of freedom or GDF, this resulted in little change in the residual variance estimate.

The data-driven method demonstrated a weakness of the GDF when applied to real data. When the adjusted variance was used in the perturbation method, the GDF was not larger than the average dimension of the model for high values of the intraclass correlation. This resulted in a corrected variance that smaller than the adjusted variance. However, since the total sample size was much larger than either degrees of freedom, this resulted in a corrected variance almost equal to the adjusted degrees of freedom.

In view of Theorems 3.1 and 3.2, it may be interesting to know the behavior of $E \left(\lim_{N \rightarrow \infty} \hat{\sigma}_{\hat{M}}^2 \right)$. The simulations showed that as the sample size increased, the residual variance approached $\hat{\sigma}_{\hat{M}_o}^2$. This result would be intuitive because, as the sample size increases, a true model is more likely being chosen, yielding an unbiased estimate of the residual variance. The aforementioned relationship may be able to be shown using the paradigm discussed in Section 3.1 and in Appendix B.2. However, this identity hinges on theoretical aspects of the consistent or efficient nature of the model selection strategy that needs to be carefully studied prior to making any claim.

Even though the GDF for the LMM to adjust for model uncertainty and the

GDF derived for adaptive model selection by *Zhang et al.* (2012) are similar, they may indeed be measuring different quantities. Although not shown in this text, the second term in (3.5) contributes very little to the overall GDF. Thus, it is not expected that using the results from *Zhang et al.* (2012) had a severe influence on the results shown in the simulation study. However, the formulation of the GDF in (3.5) itself may indicate a potential problem with using this quantity for the GDF to account for model uncertainty. A quantity that can be considered a count variable, or a variable that is a summation of indicator variables that are not ordinal, is considered unitless. This holds whether the count variable is the number of columns in a design matrix (p), the sample size (n_i or N), or the realization of a Poisson-distributed random variable. The estimator of the GDF proposed by *Zhang et al.* (2012), which is used for estimator of the dimension of the model in this work and a unitless penalty term in the original work, has units that are not combinable. The first term in (3.5) is unitless, but the second term has units that are proportional to the units of the data raised to the fourth power. These quantities cannot be added, and as such, may be an indication that the GDF as stated in (3.5) can be improved upon.

The restriction imposed in the simulation study that the covariates are ‘time-independent’ paired with the compound symmetric structure of the error terms can limit some generalizability of the results. Under this scenario, the predicted values are independent of the intraclass correlation. If this situation were not true, so that either the covariates were ‘time-dependent’ or that the structure of the error terms was a different one-parameter structure, such as auto-regressive, the profile of the GDF over the possible values of the internal elements of \mathbf{V}_i would change. It is expected that the values of the GDF would increase as the absolute value of the internal elements of \mathbf{V}_i increased. Confirmation of this will be left to future work.

Lastly, applying the proposed methodology to the cervical radiculopathy trial showed how model uncertainty affects the results in a clinical setting. Using the GDF

resulted in almost tripling degrees of freedom in the model, yielding in a corrected variance that was almost 1.25 times larger than the original estimate. A sample size calculation for a potential prospective study showed that using the corrected degrees of freedom resulted in a substantial increase in the number of subjects needed for a sufficiently powered study as compared to one using the adjusted variance estimate.

CHAPTER IV

Adjusting for Model Uncertainty in Inference

4.1 Model Uncertainty and Hypothesis Testing

The previous chapter gave an overview of estimating the residual variance while accounting for model uncertainty. The residual variance is a key component when performing hypothesis tests or constructing confidence intervals, which were shown in Chapter 2 to have less than optimal coverage probabilities. This section uses the corrected variance developed in the previous chapter to perform hypothesis tests and construct confidence intervals while adjusting for model uncertainty. Rather than just use the corrected variance in the hypothesis tests and confidence intervals suggested by the generalised least squares theory, a statistic based on the general global linear approximation proposed by *Shen et al.* (2004a) is used. The methodology presented in this chapter is examined by further extending the simulation results and with the cervical radiculopathy trial discussed in Chapters 2 and 3.

4.2 Properties of Post-Model Selection Fixed-Effect Estimators

4.2.1 Variance Estimation of the Fixed-Effect Parameters accounting for Model Uncertainty

Coverage probabilities and hypothesis tests of a linear combination of parameter estimates are functions of both the fixed-effect parameter estimates and the variance of these estimates. As discussed in Chapter 3, the parameter estimates are asymptotically unbiased when estimated after an efficient or consistent model selection procedure has been applied. However, the residual variance is not unbiased in the presence of model uncertainty. Since the residual variance plays an important role in the variance of the fixed-effect estimates, this latter quantity should also not be unbiased.

The notation in the previous section can be used to assess how ignoring model uncertainty affects the variance estimation of the fixed-effect estimates. As Theorem 4.1 demonstrates, ignoring model uncertainty can result in potentially underestimating the covariance matrix of the fixed-effect estimates. This can result in higher than expected incidences of Type-I errors and coverage probability that are smaller than expected.

Theorem 4.1. *When using the generalised least squares estimate for β , $\hat{\beta}_{\hat{M}}$, then*

$$E \left[\text{cov} \left(\hat{\beta}_{\hat{M}} \right) - \text{cov} \left(\hat{\beta}_{M_o} \right) \right] \geq \mathbf{0} \quad (4.1)$$

or equivalently, the difference of the above matrices is non-negative definite. If $\mathbf{X}_{2,M_o^c} = \mathbf{0}$, then $E \left[\text{cov} \left(\hat{\beta}_{\hat{M}} \right) - \text{cov} \left(\hat{\beta}_{M_o} \right) \right] = \mathbf{0}$.

Similar to Theorem 3.3, Theorem 4.1 states that model uncertainty will result in higher variances than if the model uncertainty is ignored. However, Theorem 4.1 may

be difficult to interpret because it deals with the ‘size’ of matrices. If the off-diagonal elements of both $\text{cov}(\hat{\beta}_{\hat{M}})$ and $\text{cov}(\hat{\beta}_{M_o})$ are ignored and the focus is concentrated on the the elements along the diagonal of the covariance matrix, Theorem 4.1 simply states that the variance of the fixed-effect estimates from a model recommended by a model selection procedure is greater than if the model is assumed to be known, or that ignoring model uncertainty yields smaller standard errors. The proof of Theorem 4.1 is a generalization of a similar proof in *Shen et al.* (2004a) for error terms assumed to be independent. If both the covariates and response were pre-multiplied by $\mathbf{V}^{-\frac{1}{2}}$, the conditions found in *Shen et al.* (2004a) are met, thus proving Theorem 4.1.

The variance of the fixed-effects from the generalised least squares solution is $\sigma^2 \cdot (\mathbf{X}' \cdot \mathbf{V}^{-1} \cdot \mathbf{X})^{-1}$, and when σ^2 is unknown and a model selection procedure has been performed, it is replaced by $\hat{\sigma}_{adj, \hat{M}}^2$. The corrected variance can be substituted for the adjusted variance to adjust for model uncertainty. However, as the simulation results from the previous chapter demonstrate, the size of the inflation of the corrected variance compared to the adjusted variance will be small for even moderate sample size. It is expected that this will have little impact on the Type-I error rates and coverage probabilities. Thus, a different type of estimator must be proposed that can adjust the variance of the fixed-effect estimates.

4.2.2 General Linear Global Approximation

It is known that the size of the elements of the variance matrix of the fixed-effect estimates is related to the number of columns of the variance matrix, which is equal to the estimated parameters in the final model. If the number of estimated parameters in the final model increases, the variance of the fixed-effect estimators will increase. In the previous section, an adjustment for model uncertainty was incorporated through one quantity, the GDF. This was the result of an exploitation of the relationship between the degrees of freedom and expected value of the predicted values. This

is a fortuitous relationship that cannot be generalized to cases outside of the scope of the original problem. Therefore, incorporating model uncertainty in the variance matrix of the parameter estimates will require a statistic other than the generalised least squares estimator. Rather than just estimating the variance of the parameter estimates, an separate estimate of the parameters themselves is proposed so that its variance can be calculated.

If only unbiased estimators of β are considered, then, by definition, the uniform minimum variance unbiased estimator (UMVUE) quality of the generalised least squares estimator can be used to identify estimators with a higher variance. Since the generalised least squares estimator of β has this quality (*Myers and Milton (1998)*), then any other unbiased estimator will have a variance no less than the generalised least squares estimator. One such estimator is discussed in the following section.

Any new statistic for estimating the parameters must have certain qualities to permit consideration. Most importantly, the estimator must be an unbiased estimator of the statistics. Because of this property and the UMVUE property of the generalised least squares estimator, it is known that the variance of a new statistic will not be smaller than that of the generalised least squares estimator. Second, just as in the estimation of the GDF, it is desirable for the size of the variance of the statistic to be related to the quality of fit of the final model so that poor fitting models will yield large variance and vice versa.

One well-known approximation that meets the aforementioned criteria is an estimation method based on the Taylor expansion, which, for a vector of statistics using the response variable \mathbf{Y} , can be written as:

$$\mathbf{g}(\mathbf{Y}) = \mathbf{g}(\boldsymbol{\mu}) + \sum_i \sum_j \frac{\partial \mathbf{g}(\mathbf{Y})}{\partial y_{ij}} \bigg|_{\mathbf{Y}=\boldsymbol{\mu}} (Y_{ij} - \mu_{ij}) \quad (4.2)$$

where $\mathbf{g}(\mathbf{Y})$ is the $(k \times 1)$ vector of statistics that is to be estimated. Under certain

conditions, such as $\mathbf{g}(\cdot)$ being continuously differentiable with respect to the true values $\boldsymbol{\mu}$ as well as \mathbf{Y} being close to $\boldsymbol{\mu}$, the following properties are true:

$$\begin{aligned} \mathbb{E}[\mathbf{g}(\mathbf{Y})] &\approx \mathbf{g}(\boldsymbol{\mu}) \\ \text{cov}[\mathbf{g}(\mathbf{Y})] &\approx \sigma^2 \cdot \mathbf{D}\mathbf{V}\mathbf{D}' \end{aligned} \tag{4.3}$$

In (4.3), $\mathbf{D}_i = \left(\frac{\partial \mathbf{g}(\mathbf{Y})}{\partial \mu_{i1}}, \dots, \frac{\partial \mathbf{g}(\mathbf{Y})}{\partial \mu_{in_i}} \right)'$ and $\mathbf{D} = (\mathbf{D}_1, \dots, \mathbf{D}_N)$ so that \mathbf{D} is a $k \times N$ matrix.

However, if $\mathbf{g}(\mathbf{Y})$ is the estimate of the regression parameters after model selection ($\hat{\boldsymbol{\beta}}_{\hat{M}}$), then it is known that $\mathbf{g}(\cdot)$ is not continuously differentiable and methods such as the Taylor expansion may not be able to provide useful estimates. Rather, an estimate of the form:

$$\mathbf{d}(\mathbf{Y}) = \boldsymbol{\alpha}_0 + \sum_{i,j} (y_{ij} - \mu_{ij}) \cdot \boldsymbol{\alpha}_{ij} \tag{4.4}$$

is proposed, where $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_{ij}$ are $(k \times 1)$ vectors ($i = (1, 2, \dots, N), j = (1, 2, \dots, n_i)$). Keeping with the language introduced by *Shen et al.* (2004a), $\mathbf{d}(\mathbf{Y})$ will be referred to as the *general global linear approximation* (GGLA) of $\mathbf{g}(\mathbf{Y})$. The expected value and variance of $\mathbf{d}(\mathbf{Y})$ are:

$$\begin{aligned} \mathbb{E}[\mathbf{d}(\mathbf{Y})] &= \boldsymbol{\alpha}_0 \\ \text{var}[\mathbf{d}(\mathbf{Y})] &= \sigma^2 \boldsymbol{\Sigma} = \sigma^2 \cdot \sum_i \sum_{j,k} V_i(j, k) \cdot \boldsymbol{\alpha}_{ij} \cdot \boldsymbol{\alpha}_{ik} \end{aligned} \tag{4.5}$$

In a similar strategy to the estimation of the GDF, estimators of the $\boldsymbol{\alpha}_{ij}$ are proposed so they are related to the robustness of the model procedure. The definition of the generalized leverage in Definition 3.4 can be used to propose an estimator of the $\boldsymbol{\alpha}_{ij}$ that incorporates model uncertainty. The estimators, which can be found in Definition 4.2, are related to the concept of the GDF described in Chapter 3.

Definition 4.2. The optimal vectors for the $\boldsymbol{\alpha}_{ij}$ that incorporate model uncertainty

are:

$$\begin{aligned}
\alpha_{ij}^o &= \frac{\partial}{\partial \mu_{ij}} E[\mathbf{g}(\mathbf{Y})] \\
&= \frac{1}{\sigma^2} E[\mathbf{e}_j' \cdot \mathbf{V}_i^{-1} \cdot (\mathbf{Y}_i - \boldsymbol{\mu}_i) \cdot \mathbf{g}(\mathbf{Y})] \\
&= \frac{1}{\sigma^2} \text{cov}(\mathbf{e}_j' \cdot \mathbf{V}_i^{-1} \cdot (\mathbf{Y}_i - \boldsymbol{\mu}_i), \mathbf{g}(\mathbf{Y}))
\end{aligned} \tag{4.6}$$

with \mathbf{e}_j being the j^{th} column of a $n_i \times n_i$ identity matrix. All other notation has been defined.

To ensure that the estimate of the GGLA is unbiased, the optimal estimate of α_0^o is $E[\mathbf{g}(\mathbf{Y})]$. The derivation of Theorem 4.2 is a straightforward substitution in the proof for Theorem 3.5.

The parameters α_{ij} primarily serve to inflate the variance as an adjustment due to poor fitting models, as the size of the elements of α_{ij} will be related to the flexibility of the model fitting procedure. When summed over all observations, these parameters can also impact the estimate of $\mathbf{g}(\mathbf{Y})$, particularly in small samples. It is also possible that certain elements of α_{ij} may be large while others may be near zero, indicating that only certain elements of $\mathbf{g}(\mathbf{Y})$ are susceptible to model uncertainty while others are fairly robust. The values of the elements α_{ij} near zero will correspond to the robust elements of $\mathbf{g}(\mathbf{Y})$, and the variance of these terms will be less affected by model uncertainty. Conversely, elements in $\mathbf{g}(\mathbf{Y})$ that are very susceptible to large changes when the response variable is perturbed by small amounts will have elements of α_{ij} large in magnitude.

Since the GGLA is related to the theory that was used to estimate the GDF, it would be logical to use the same type of estimation method to estimate parameters in the GGLA. The perturbation method can be used for this estimation, and a more detailed description of this method for the GGLA is found in the subsequent section. The GGLA does not propose a solution to σ^2 and \mathbf{V}_i so that other methods must be used provide these estimates. As in the rest of this work, the elements of \mathbf{V}_i

are considered known. Either the adjusted or corrected variance can be used for the parameter σ^2 . In keeping with the theme of this entire work, it is recommended that the corrected variance be used.

As (4.4) is written, the GGLA is not a statistic since the estimator contains the unknown parameters $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_{ij}$ and μ_{ij} . The estimation of $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_{ij}$ are discussed in the following section. The predicted value $\hat{\mu}_{ij}$, calculated from the generalised least squares estimates, can be substituted in for μ_{ij} creating an estimate of $\mathbf{g}(\mathbf{Y})$ that is data-dependent. Since the predicted values are functions of the response, failing to account for this variability will likely bias the results to some extent. When the estimates of $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_{ij}$ and μ_{ij} are used, the resulting vector is defined as:

$$\hat{\mathbf{d}}(\mathbf{Y}) = \hat{\boldsymbol{\alpha}}_0 + \sum_{i,j} (y_{ij} - \hat{\mu}_{ij}) \cdot \hat{\boldsymbol{\alpha}}_{ij} \quad (4.7)$$

The variance of the GGLA is interpreted similar to that of the variance of the post-model selection fixed-effect estimator using generalised least squares. For instance, the diagonal elements of $\text{var}[\mathbf{d}(\mathbf{Y})]$ are standard errors of the GGLA while the off-diagonal elements are covariances between fixed-effect estimates. This permits testing of linear hypotheses and constructing multivariate confidence intervals, which will be discussed in Section 4.4.

The variance of the GGLA has some important qualities. The variance matrix is symmetric and positive definite, which are crucial elements when constructing hypothesis tests. This topic will be further discussed in Section 4.4. The variance matrix itself is unstructured as in that it has $\frac{q(q-1)}{2}$ different values, which is the maximum for a matrix of its size. This restricts the scope of statistics that can be used for $\mathbf{d}(\mathbf{Y})$. Strictly speaking, the GGLA could be used to estimate $\mathbf{d}(\mathbf{Y}) = \mathbf{Y}_i$ so that the GGLA is itself a model for $\boldsymbol{\mu}$. However, if it was that other blocks shared the covariance structure, the GGLA can not be used. Furthermore, if the response vector

is assumed to have any sort of parametric error structure other than unstructured, the GGLA can not be used.

4.3 Estimation of the GGLA

As previously noted, the GGLA in Theorem 4.2 is based on the results from the estimation of the GDF calculated in Theorem 3.5. Because of the similarities, it would make sense that a similar estimation method could be used to obtain estimates of the vectors α_0 and α_{ij} . If a psuedo-response variable, as in (3.4), is considered the response and this response conditions on the observed response variable, then an estimate of the optimal vectors α_0 and α_{ij} are expressed in Corollary 4.3:

Corollary 4.3. *If the response variable is \mathbf{Y}_i^* , where $\mathbf{Y}_i^* = \mathbf{Y}_i + \tau \cdot \sigma_* \tilde{\mathbf{Y}}$ as in (3.4), then, under the conditions of Definition 4.2, the estimates of α_0 and α_{ij} can be calculated by:*

$$\begin{aligned}\hat{\alpha}_0^o &= E[\mathbf{g}(\mathbf{Y}^*)] \\ \hat{\alpha}_{ij}^o &= \frac{\partial}{\partial Y_{ij}} E^*[\mathbf{g}(\mathbf{Y}^*)] \\ &= \frac{1}{(\tau \cdot \sigma_*)^2} E^* [\mathbf{e}'_j \cdot \mathbf{V}_i^{-1} \cdot (\mathbf{Y}_i^* - \mathbf{Y}_i) \cdot \mathbf{g}(\mathbf{Y}^*)] \\ &= \frac{1}{(\tau \cdot \sigma_*)^2} \text{cov}^* (\mathbf{e}'_j \cdot \mathbf{V}_i^{-1} \cdot (\mathbf{Y}_i^* - \mathbf{Y}_i), \mathbf{g}(\mathbf{Y}^*))\end{aligned}\tag{4.8}$$

The data perturbation method discussed previously can be used to estimate the parameters of the GDF. Under a Monte-Carlo simulation method similar to the one described in the previous chapter, the vector \mathbf{Y}_i^{*d} ($d=1,2,\dots,D$) is simulated from a normal distribution with mean \mathbf{Y}_i and variance \mathbf{V}_i and combined to form a pseudo-response vector $\mathbf{Y}^{*d}=(\mathbf{Y}_1^{*d}, \mathbf{Y}_2^{*d}, \dots, \mathbf{Y}_N^{*d})'$. For each pseudo-response vector, $\mathbf{g}(\mathbf{Y}^{*d})$ is calculated. From the results of Corollary 4.3, the estimate of $\hat{\alpha}_0^o$ can be approximated by the arithmetic mean of statistic $\mathbf{g}(\mathbf{Y}^*)$ over all D Monte-Carlo simula-

tions. Similarly, the vector $\hat{\alpha}_{ij}^o$ can be estimated using conditional covariance of the quantity $\mathbf{e}'_j \cdot \mathbf{V}_i^{-1} \cdot (\mathbf{Y}_i^* - \mathbf{Y}_i)$ with statistic $\mathbf{g}(\mathbf{Y}^*)$, respectively. If b_{ij}^d is defined as $\mathbf{e}'_j \cdot \mathbf{V}_i^{-1} \cdot (\mathbf{Y}_i^* - \mathbf{Y}_i)$, $\bar{b}_{ij} = D^{-1} \sum_d b_{ij}^d$, and $\bar{\mathbf{g}} = D^{-1} \sum_d \mathbf{g}(\mathbf{Y}^{*d})$ then the estimates of $\hat{\alpha}_0^o$ and $\hat{\alpha}_{ij}^o$ are :

$$\begin{aligned}\hat{\alpha}_0^o &= \frac{1}{D} \sum_d \mathbf{g}(\mathbf{Y}^{*d}) \\ \hat{\alpha}_{ij}^o &= \frac{1}{(D-1) \cdot (\tau \cdot \sigma_*)^2} \sum_d (b_{ij}^d - \bar{b}_{ij}) \cdot (\mathbf{g}(\mathbf{Y}^{*d}) - \bar{\mathbf{g}})\end{aligned}\tag{4.9}$$

The quantity $(D-1)$ in the divisor of $\hat{\alpha}_{ij}^o$ is used rather than (D) because one degree of freedom is spent estimating $\bar{\mathbf{g}}$. This divisor is typically used in estimating the variance of a statistic using a bootstrap or any other Monte Carlo estimation method. The original derivation of the GGLA by *Shen et al.* (2004a) for error terms assumed to be independent and homogenous uses the divisor of $(D-1)$, and this divisor was kept to be notationally consistent. Regardless of the choice of divisor, as becomes D moderate to large, the estimates will be similar and the effect of the divisor will be negligible.

The estimates described in (4.9) have some important qualities. Both $\hat{\alpha}_0^o$ and $\hat{\alpha}_{ij}^o$ can be thought of as means and have a asymptotic distribution that is governed by the Central Limit Theorem. As such, it can be concluded that $\hat{\mathbf{d}}(\mathbf{Y})$ will have an asymptotic multivariate normal distribution with mean $E[\mathbf{g}(\mathbf{Y})]$ and variance $\sigma^2 \cdot \Sigma$. When $\hat{\alpha}_0$, $\hat{\alpha}_{ij}$, σ^2 , and \mathbf{V}_i are unknown, estimates can be used, although this situation would result in approximate asymptotic distributions.

The statistic $\hat{\mathbf{d}}(\mathbf{Y})$ is based on two quantities that are estimated using the perturbation method that require multiple pseudo-response vectors to be computed from a Monte-Carlo simulation. The notation of these simulations is the same for the estimation of the GDF and the GGLA; in practice, the pseudo-response vectors should be simulated separately. Just as in the estimation of the GDF, D should be chosen

to be at least larger than the total sample size N . These restrictions will ensure that the estimates of $\hat{\alpha}_0^o$ and $\hat{\alpha}_{ij}^o$ are robust against the specific pseudo-response vectors as well as robust against the choice of divisor discussed previously. Lastly, when all observations can be assumed to be independent, the estimates in (4.9) reduce to the estimates of the *Shen et al. (2004a)*.

4.4 Hypothesis Testing and Construction of Confidence Intervals

If model uncertainty were ignored and hypothesis testing or construction of confidence intervals on a linear combination of the parameters was to be done, classical linear models theory provides the basis for this action. In a null hypothesis for a linear combination of the parameters, written as $H_o : \mathbf{L} \cdot \boldsymbol{\beta} = \mathbf{h}_o$, the statistic using the generalised least squares estimator $\mathbf{L} \cdot \hat{\boldsymbol{\beta}}_{\hat{M}} - \mathbf{h}_o$ will have an asymptotic multivariate normal distribution with mean $\mathbf{0}$ and variance $\sigma^2 \left(\mathbf{L} \left(\mathbf{X}'_{\hat{M}} \cdot \mathbf{V}^{-1} \cdot \mathbf{X}_{\hat{M}} \right)^{-1} \mathbf{L}' \right)^{-1}$. The matrix \mathbf{L} , which will be referred to as the *contrast matrix*, is a full-rank matrix that has dimension $q \times p$. Ignoring model uncertainty, a Wald statistic $Q_{\hat{M}}$ is calculated as:

$$Q_{\hat{M}} = \hat{\sigma}_{adj, \hat{M}}^2 \left(\mathbf{L} \cdot \hat{\boldsymbol{\beta}}_{\hat{M}} - \mathbf{h}_o \right)' \cdot \left(\mathbf{L} \cdot \left(\mathbf{X}_{\hat{M}} \cdot \mathbf{V}^{-1} \cdot \mathbf{X}_{\hat{M}} \right)^{-1} \mathbf{L}' \right)^{-1} \cdot \left(\mathbf{L} \cdot \hat{\boldsymbol{\beta}}_{\hat{M}} - \mathbf{h}_o \right) \quad (4.10)$$

Under the null hypothesis, $Q_{\hat{M}}$ has an asymptotic χ^2 distribution with q degrees of freedom.

The asymptotic normality will also allow the construction of confidence intervals. A single contrast $\ell_m \cdot \hat{\boldsymbol{\beta}}_{\hat{M}}$ has an asymptotic normal distribution with mean $\ell_m \cdot \boldsymbol{\beta}$ and variance $\sigma^2 \left(\ell_m \left(\mathbf{X}'_{\hat{M}} \cdot \mathbf{V}^{-1} \cdot \mathbf{X}_{\hat{M}} \right)^{-1} \ell_m' \right)^{-1}$, where ℓ_m is the m^{th} row of \mathbf{L} . A

marginal $100(1 - \alpha)\%$ confidence interval can be written as:

$$\left(\ell_m \cdot \hat{\beta}_{\hat{M}} - z_{1-\frac{\alpha}{2}} \cdot SE \left(\ell_m \cdot \hat{\beta}_{\hat{M}} \right), \ell_m \cdot \hat{\beta}_{\hat{M}} + z_{1-\frac{\alpha}{2}} \cdot SE \left(\ell_m \cdot \hat{\beta}_{\hat{M}} \right) \right) \quad (4.11)$$

with $SE \left(\ell_m \cdot \hat{\beta}_{\hat{M}} \right) = \sigma \sqrt{\left(\ell \left(\mathbf{X}'_{\hat{M}} \cdot \mathbf{V}^{-1} \cdot \mathbf{X}_{\hat{M}} \right)^{-1} \ell' \right)}$ being the standard error of the contrast and $z_{1-\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})^{th}$ percentile of the standard normal distribution. The interval in (4.11) is a marginal interval because it does not depend on any other contrasts $m' \neq m$. When $q > 1$, the confidence intervals are typically adjusted for multiple comparisons, so that overall, the family of confidence intervals has a coverage probability of $100(1 - \alpha)\%$. These are known as simultaneous confidence intervals and their consideration will be outside the scope of this work. When $q=1$, the relationship between the chi-square and normal distributions is often exploited, where the positive square root of a random variable with one degree of freedom results in a normally distributed variable. This allows use of the normal distribution in hypothesis testing and the simultaneous confidence interval will reduce to the marginal confidence interval.

One possible way to incorporate model uncertainty in the hypothesis tests in (4.10) could be to use the corrected variance in place of σ^2 . It is not expected this will improve the testing qualities discussed in Chapter 2 when the sample size is much larger than the GDF. This strategy will be referred to as the corrected method and the relationship between the GDF and adjusted variance results in $Q_{cor} = Q_{\hat{M}} \cdot \frac{N-p_{\hat{M}}}{N-\hat{\mathcal{D}}_{\hat{M}}}$ and $SE_{cor} \left(\ell_m \cdot \hat{\beta}_{\hat{M}} \right) = SE \left(\ell_m \cdot \hat{\beta}_{\hat{M}} \right) \cdot \sqrt{\frac{N-p_{\hat{M}}}{N-\hat{\mathcal{D}}_{\hat{M}}}}$. Substituting these quantities into (4.10) and (4.11) will yield the corrected test statistics and confidence intervals. In this strategy, the quantity $\frac{N-p_{\hat{M}}}{N-\hat{\mathcal{D}}_{\hat{M}}}$ acts as an inflation factor, inflating the statistics for model uncertainty.

Another strategy for performing hypothesis tests and constructing confidence intervals is to use the estimate of the GGLA. Because of the asymptotic normality

of $\hat{\mathbf{d}}(\mathbf{Y})$, a test statistic can be created that tests a similar hypothesis as (4.10). If $\hat{\boldsymbol{\beta}}_{GGLA}$ is the estimate of the fixed-effects parameters computed using (4.7) and (4.9) to estimate $\mathbf{g}(\mathbf{Y})$, the corresponding test statistic:

$$Q_{GGLA} = \hat{\sigma}_{cor}^2 \left(\mathbf{L} \cdot \hat{\boldsymbol{\beta}}_{GGLA} - \mathbf{h}_o \right)' \cdot (\mathbf{L} \cdot \boldsymbol{\Sigma} \cdot \mathbf{L}')^{-1} \cdot \left(\mathbf{L} \cdot \hat{\boldsymbol{\beta}}_{GGLA} - \mathbf{h}_o \right). \quad (4.12)$$

Under the null hypothesis, this test statistic will have an asymptotic χ^2 distribution with q degrees of freedom.

Due to the asymptotic normality of the GGLA estimates, confidence intervals can be created in a similar fashion to (4.11). Replacing $\hat{\boldsymbol{\beta}}_{\hat{M}}$ and $SE\left(\ell_m \cdot \hat{\boldsymbol{\beta}}_{\hat{M}}\right)$ in (4.11) with the appropriate estimates from the GGLA model yields:

$$\left(\ell_m \cdot \hat{\boldsymbol{\beta}}_{\hat{M}} - z_{1-\frac{\alpha}{2}} \cdot \sigma_{cor} \sqrt{\ell_m \cdot \boldsymbol{\Sigma} \cdot \ell'_m}, \ell_m \cdot \hat{\boldsymbol{\beta}}_{\hat{M}} + z_{1-\frac{\alpha}{2}} \cdot \sigma_{cor} \sqrt{\ell_m \cdot \boldsymbol{\Sigma} \cdot \ell'_m} \right) \quad (4.13)$$

Lastly, a test statistic that uses both the generalised least squares estimates and the estimates from the GGLA can be used for hypothesis testing and the construction of confidence intervals. The fixed-effect estimates from the generalised least squares are used because of their well-known qualities and the variance estimate from the GGLA is used because of its ability to provide inflated standard error estimates that adjust for model uncertainty. This strategy, which will be referred to as the combined method, yields a test statistic:

$$Q_{comb} = \hat{\sigma}_{cor}^2 \left(\mathbf{L} \cdot \hat{\boldsymbol{\beta}}_{\hat{M}} - \mathbf{h}_o \right)' \cdot (\mathbf{L} \cdot \boldsymbol{\Sigma} \cdot \mathbf{L}')^{-1} \cdot \left(\mathbf{L} \cdot \hat{\boldsymbol{\beta}}_{\hat{M}} - \mathbf{h}_o \right). \quad (4.14)$$

Because of the robust nature of the generalised least squares estimates, it is assumed that $\left(\mathbf{L} \cdot \hat{\boldsymbol{\beta}}_{\hat{M}} - \mathbf{h}_o \right)$ will have an asymptotic normal distribution with null mean and variance $(\mathbf{L} \cdot \boldsymbol{\Sigma} \cdot \mathbf{L}')$ under the null hypothesis. This method of testing was originally used by *Shen et al.* (2004a) in his assessment of the GGLA for an indepen-

dent response variable. Just as in the previous estimates, a confidence interval can be constructed as:

$$\left(\ell_m \cdot \hat{\beta}_{\hat{M}} - z_{1-\frac{\alpha}{2}} \cdot \sigma_{cor} \sqrt{\ell_m \cdot \Sigma \cdot \ell'_m}, \ell_m \cdot \hat{\beta}_{\hat{M}} + z_{1-\frac{\alpha}{2}} \cdot \sigma_{cor} \sqrt{\ell_m \cdot \Sigma \cdot \ell'_m} \right) \quad (4.15)$$

When $q=1$, the hypothesis tests and confidence intervals described in this section are interchangeable. This means if a confidence interval includes the hypothesized value, the corresponding hypothesis will fail to reject the null hypothesis, and vice versa. Thus, the coverage probability and Type-I error are complements when the hypothesized value is zero. Since the Type-I error is more commonly understood than the coverage probability, this quantity is used in the assessment of the aforementioned methods when the focus parameter is zero. When the focus parameter is non-zero, coverage probabilities are assessed.

Each of these strategies will provide unique results in adjusting for model uncertainty. Both the simulation study and the cervical radiculopathy trial can be used to assess the application of these proposed methods. The following sections detail these analyses.

4.5 Simulation Study

Just as in Chapter 3, the proposed methods are evaluated in the context of the simulation study introduced in Chapter 2. For each combination of simulation parameters, the proof-of-concept and data-driven results were assessed to show how the proposed methods perform in both theoretical and clinical scenarios. The proof-of-concept method uses the variance of the simulated model ($\sigma^2 = 1$) for both the perturbation estimation method and variance estimate whereas the data-driven method uses the corrected variance estimate for both the estimation method and variance estimation. Just as in the previous simulations, the elements of \mathbf{V}_i are assumed to

be known but varied to assess how misspecification of these values affect the results.

The GGLA is related to many different quantities whose behavior has been assessed in the presence of model uncertainty. In this study, the estimate of the focus variable from the GGLA is assessed to determine if the GGLA estimate of the fixed-effects parameter is a sufficient estimator of this parameter. The standard error of the focus variable is also assessed to determine how model uncertainty was incorporated into the variance of of the GGLA. Lastly, the Type-I errors of the hypothesis tests and the coverage probabilities of confidence intervals constructed using each of the three methods that adjust for model uncertainty, and the generalised least squares estimate that does not adjust for model uncertainty, are be assessed for each of the different simulation paradigms. The assessment of each quantity was done with respect to the results from methodology that does not correct for model uncertainty, which was shown in the simulation results in Chapter 2.

The proof-of-concept simulation results for the estimate of the focus variable can be seen in Figure 4.1. Generally speaking, the estimates from the GGLA method closely match the generalised least squares estimates, particularly if there are no other non-zero effects in the model (Figures 4.1a,4.1c). There is some slight divergence from the generalised least squares estimates when non-zero effects are in the model (Figure 4.1b) that is pervasive regardless of the other simulation parameters, but this divergence tapers with an increased sample size (Figure 4.1d). The data-driven method produces similar results. These results, as well as the extended results for the proof-of-concept and data-driven methods can be seen in Appendices D.1 and D.2.

Although the mean values are comparable between the estimation methods, the standard deviation of GGLA estimates is higher than the generalised least squares estimates for reasonable values of the intraclass correlation. The difference between standard deviations is a function of whether there are other non-zero parameters in the model, so that in the presence of non-zero variables, the standard deviation increases.

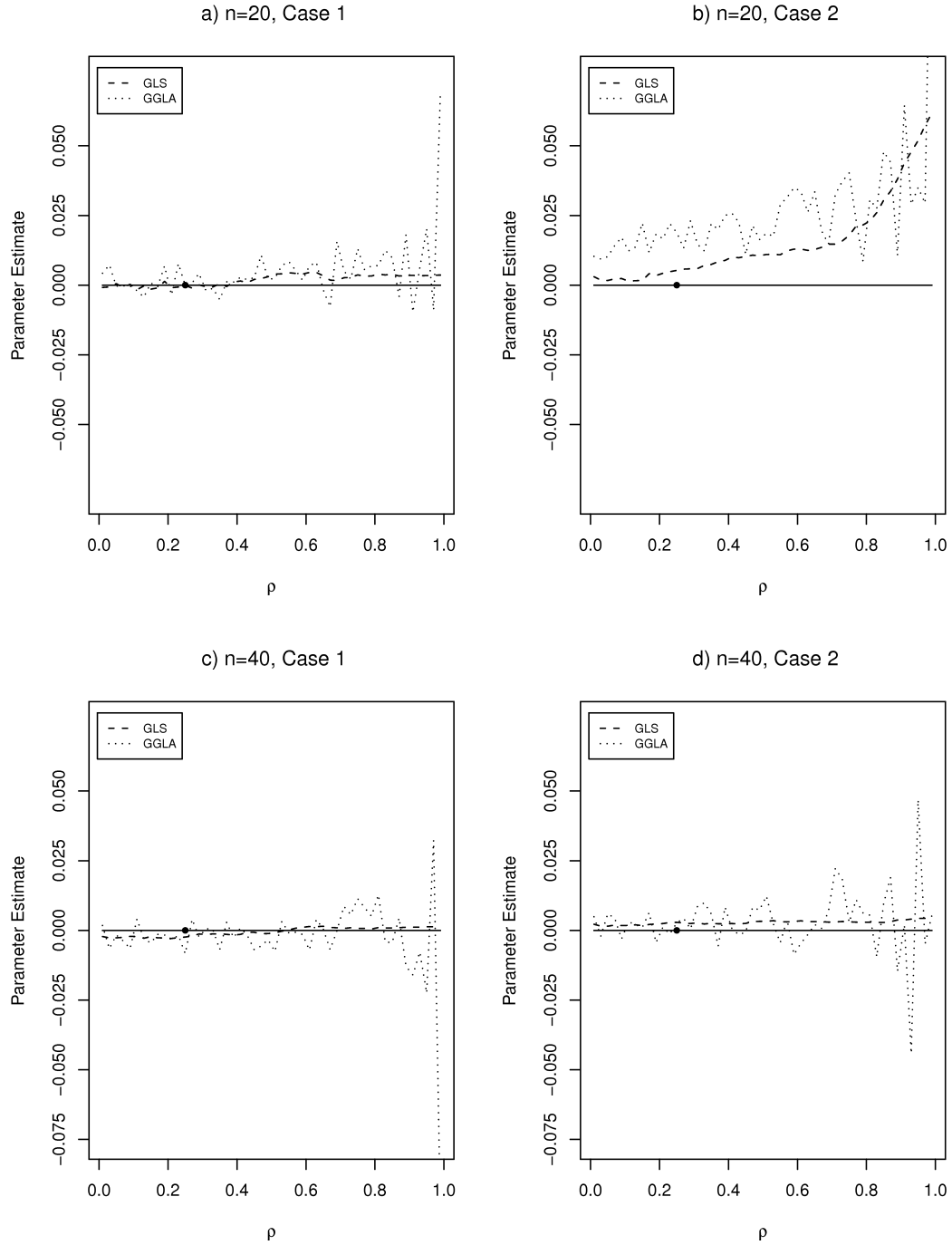


Figure 4.1: Proof-of-concept simulation results showing the mean GGLA estimate of the focus parameter ($\rho=0.25$). The generalised least squares estimates are displayed as the dashed line, while the GGLA estimates are the dotted line. The GGLA estimates mirrors the generalised least squares estimates

This characteristic is exemplified by Figure 4.2 and implies that the GGLA method produces estimates that are typically farther away from the mean value than the generalised least squares estimates. When taken individually, this could severely affect the Type-I error rates and coverage probabilities that use the GGLA estimates, even if the standard error was adjusted appropriately. More importantly, this casts doubt on the interpretability of the parameter estimates, particularly if these estimates are producing results that differ greatly in magnitude than is expected. The extended simulation results can be found in Appendix D.3. The data-driven method produces very similar results, so much so that they do not need to be demonstrated here. All of the data-driven results can be found in Appendix D.4.

The inflation of the standard error of the focus variable using the GGLA and corrected methods are shown in Figure 4.3 for the proof-of-concept simulation. Because the total sample size was much larger than the degrees of freedom, the corrected and generalised least squares estimates are virtually indistinguishable. The standard error of the focus variable is greater than either the standard error calculated from the generalised least squares or corrected methods regardless of simulation parameters and for reasonable values of the intraclass correlation. The extended proof-of-concept simulation results for the estimate of the standard error of the focus variable for the proof-of-concept simulations can be seen in Appendix D.5.

The standard deviations calculated using the corrected method are one-to-one functions of the GDF, so the profile of the proportion of times this quantity is greater than the unadjusted method was displayed in Chapter 3. The proportion of times the standard errors computed by the GGLA method are larger than the method that does not adjust for model uncertainty can be seen in Figure 4.4. These resemble the shape of the profiles in Chapter 3, but are slightly smaller in magnitude. Just as the plots in Figure 3.2, the proportion of times the standard error computed by the GGLA is greater than the unadjusted standard error increases with sample size

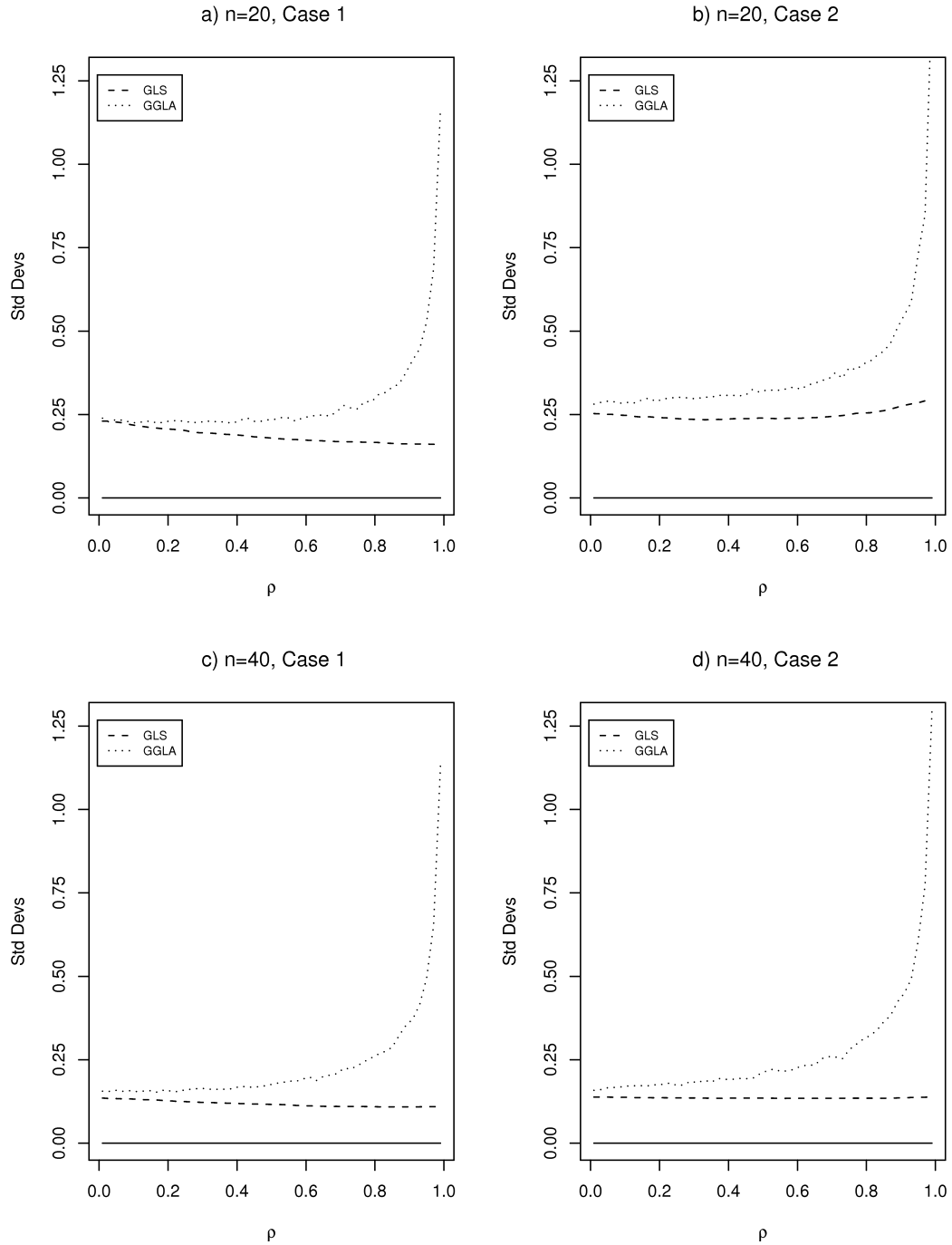


Figure 4.2: Proof-of-concept standard deviation of the parameter estimates from the simulation study ($\rho=0.25$). The dashed line corresponds to the generalised least squares estimates and the dotted lines correspond to the GGLA estimates. The GGLA estimates typically have a higher standard deviation, suggesting estimates with more extreme values.

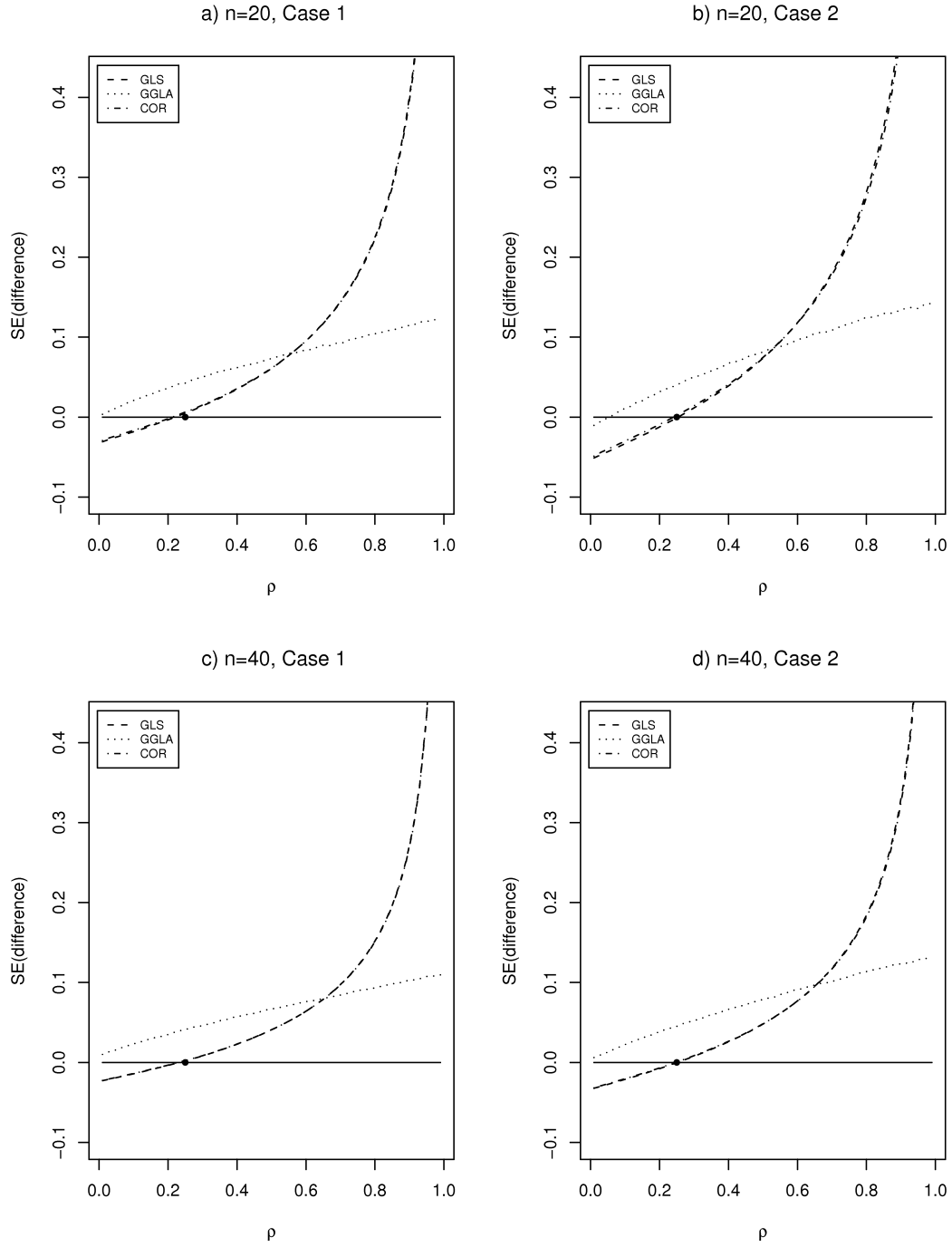


Figure 4.3: Proof-of-concept simulation results showing the standard errors stemming from the GGLA and corrected methods ($\rho=0.25$). The dashed line corresponds to the generalised least squares and the mixed line corresponds to the corrected method. These lines are very close and resemble one mixed line. The GGLA standard error estimates are the dotted line and, for reasonable values of the intraclass correlation, are larger than the standard error.

(Figure 4.4a vs 4.4c, 4.4b vs 4.4d).

The data-driven method shows similar results to the proof-of-concept simulation for small values of the intraclass correlation. However, the results diverge when $\rho = 0.75$. As Figure 4.5a demonstrates, the standard error of the focus variable is approximately equal to or less than the estimate calculated from the generalised least squares estimates. Furthermore, the corrected method holds the same quality when the intraclass correlation is large. This occurs regardless of the model selection metric or simulation paradigm used. The reason for this quality is most likely due to the inability of the corrected variance to sufficiently inflate the estimate of σ^2 for model uncertainty. The simulations also show a huge inflation of the standard error for Case 4 when the AIC metric is used, as seen in Figures 4.5b and 4.5d. It is not clear how or why this inflation arises, but it does not appear when the BIC is used. The extended simulation results for the data-driven method can be seen in Appendix D.7.

The probability that the GDF calculated from the GGLA through a data-driven method echo the overall profile for the proof-of-concept simulations, but have the negative tendencies discussed in the previous paragraph. For space considerations, all of these results can be seen in Appendix D.8.

The Type-I errors of the four methods of constructing hypotheses for the proof-of-concept simulations can be seen in Figure 4.6 for Cases 1 and 2. Since the corrected variance is larger than the adjusted variance for reasonable values of the intraclass correlation, the Type-I error rates are reflective of a test that is more apt to fail to reject the null hypothesis, a characteristic known as being *conservative*. For small sample sizes, this results in Type-I error rates being closer to the nominal value. When the sample size becomes large, the test is often too conservative so that the test that does not adjust for model uncertainty yields Type-I error rates that are closer to nominal values. These characteristics are robust against the simulation parameters,

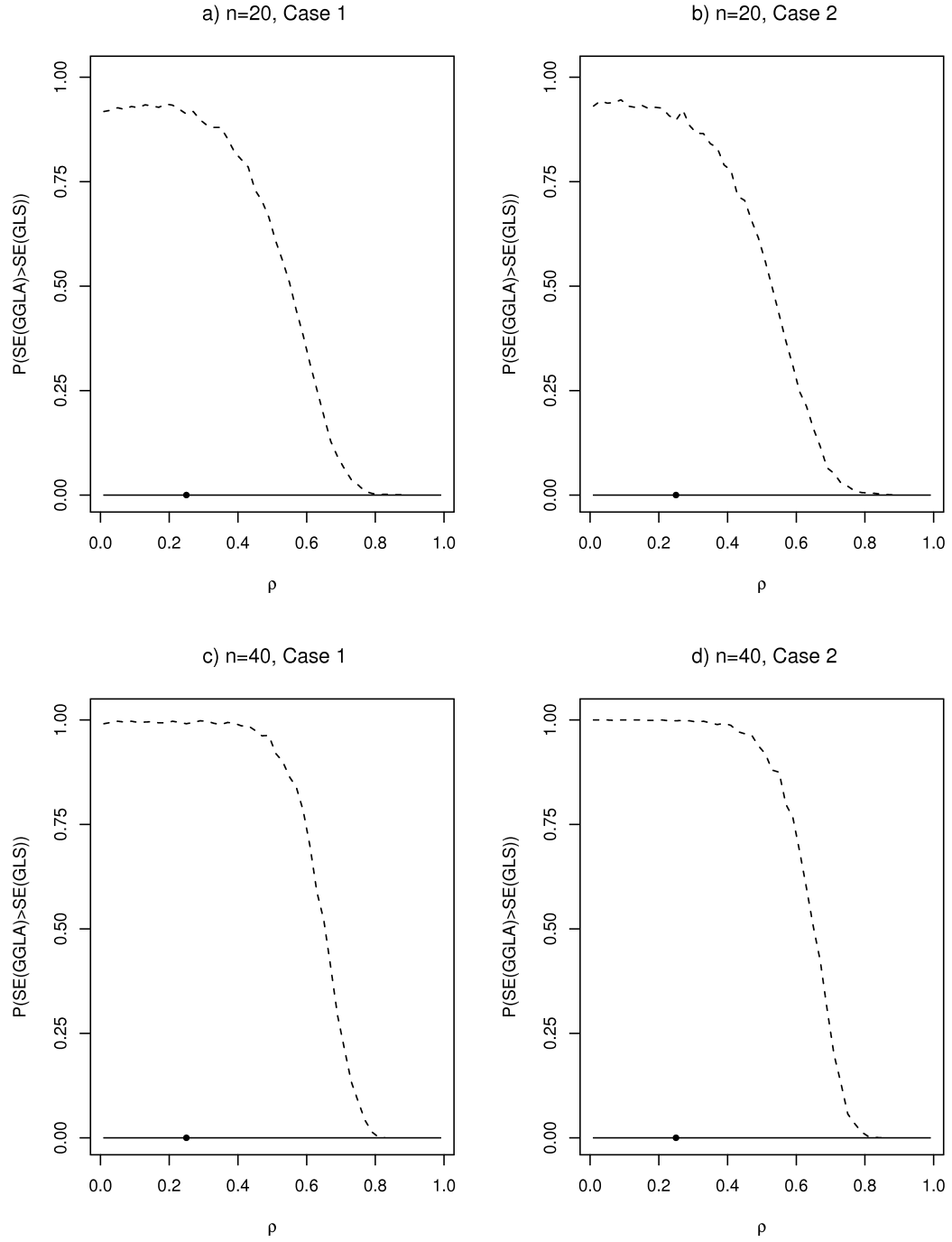


Figure 4.4: [Probabilities that the standard errors from the GGLA method are larger than the GLS estimators for the proof-of-concept simulation ($\rho=0.25$). The dashed line is the probability, which typically overestimates the standard error compared to the GLS estimates.

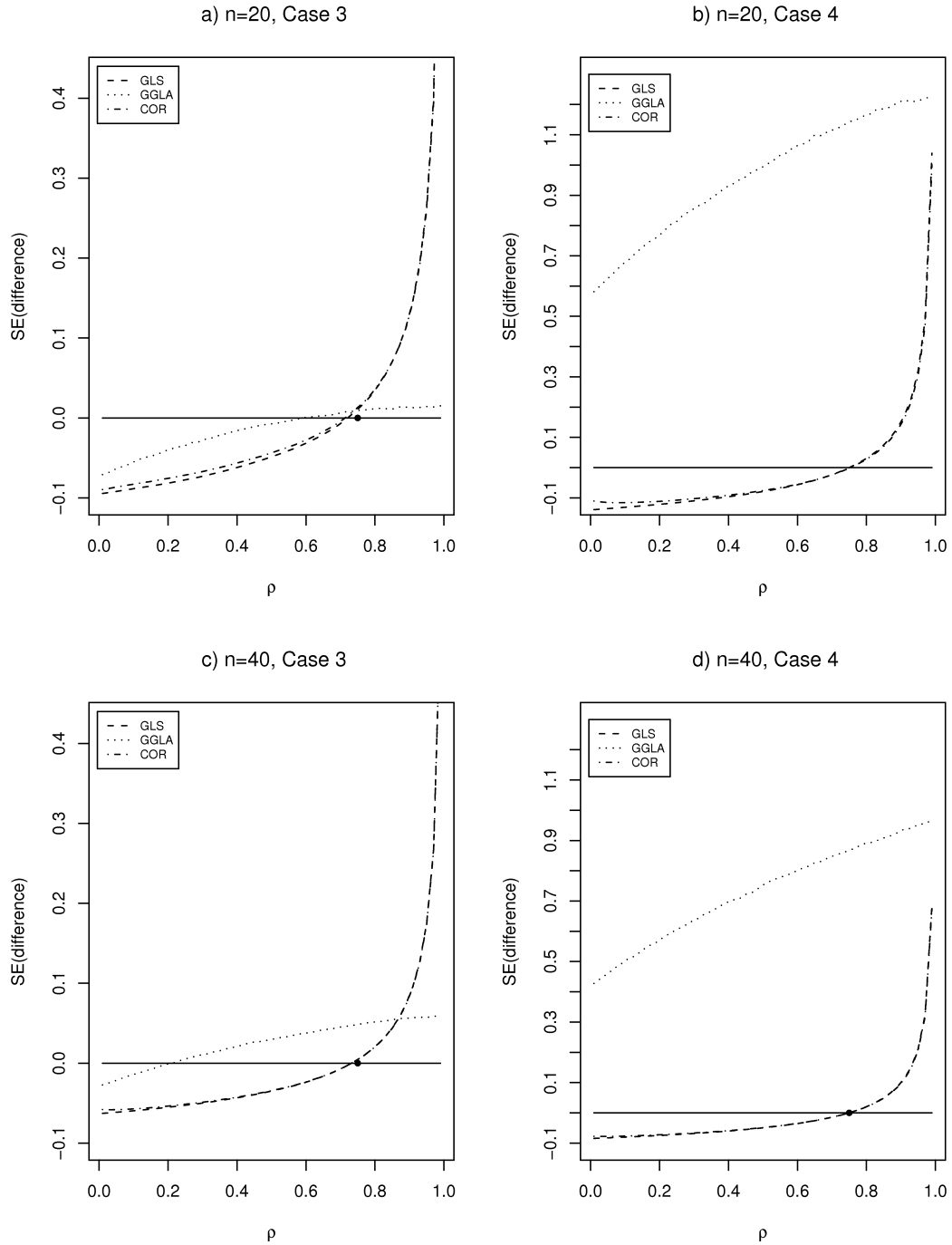


Figure 4.5: [Data-driven simulation results showing the standard errors stemming from the GGLA and corrected methods ($\rho=0.75$). The dashed line corresponds to the generalised least squares and the mixed line corresponds to the corrected method. These lines are very close and resemble one mixed line. The GGLA standard error estimates are the dotted line.

including the value of the focus parameter so that the coverage probabilities are closer to nominal value. The extended simulation results can be seen in Appendix D.9.

When the focus parameter is truly zero, the hypothesis tests using the GGLA method have a tendency to be *liberal*, or having a tendency to reject the null hypothesis more often than expected. As Figure 4.5 demonstrates, this is not due a standard error that lacks a proper adjustment for model uncertainty; rather, the estimates of the focus parameter are more varied about zero, resulting in estimates of the focus parameter that are larger in magnitude than the generalised least squares estimates. This quality extends to Cases 3 and 4, where the true value of the focus parameter is non-zero. However, in these cases, the coverage probabilities from the GGLA method are smaller than the coverage probability using the generalised least squares method without adjusting for model uncertainty.

The combined method exhibits similar characteristics to the corrected method when the focus parameter is zero. It is more conservative than the generalised least squares tests that do not adjust for model uncertainty, regardless of the sample size. It displays a conservative tendency for large sample sizes that could be potentially limit its practical usefulness (Figures 4.6c, 4.6d). When the focus parameter is non-zero and the number of blocks is 20, the combined method yields coverage probabilities that are closer to the nominal value than any of the other methods. When the number of blocks increases to 40, the unadjusted, corrected, and even GGLA methods have coverage probabilities closer to the nominal value. These characteristics are robust against reasonable values of the model selection metric and simulation paradigm chosen. It is generally robust against true value of the intraclass correlation, although the coverage probability profiles of the corrected method and combined method eventually intersect, leading to a slightly different recommendation of a superior method if the intraclass correlation value used for the analysis differs from the truth by a moderate amount.

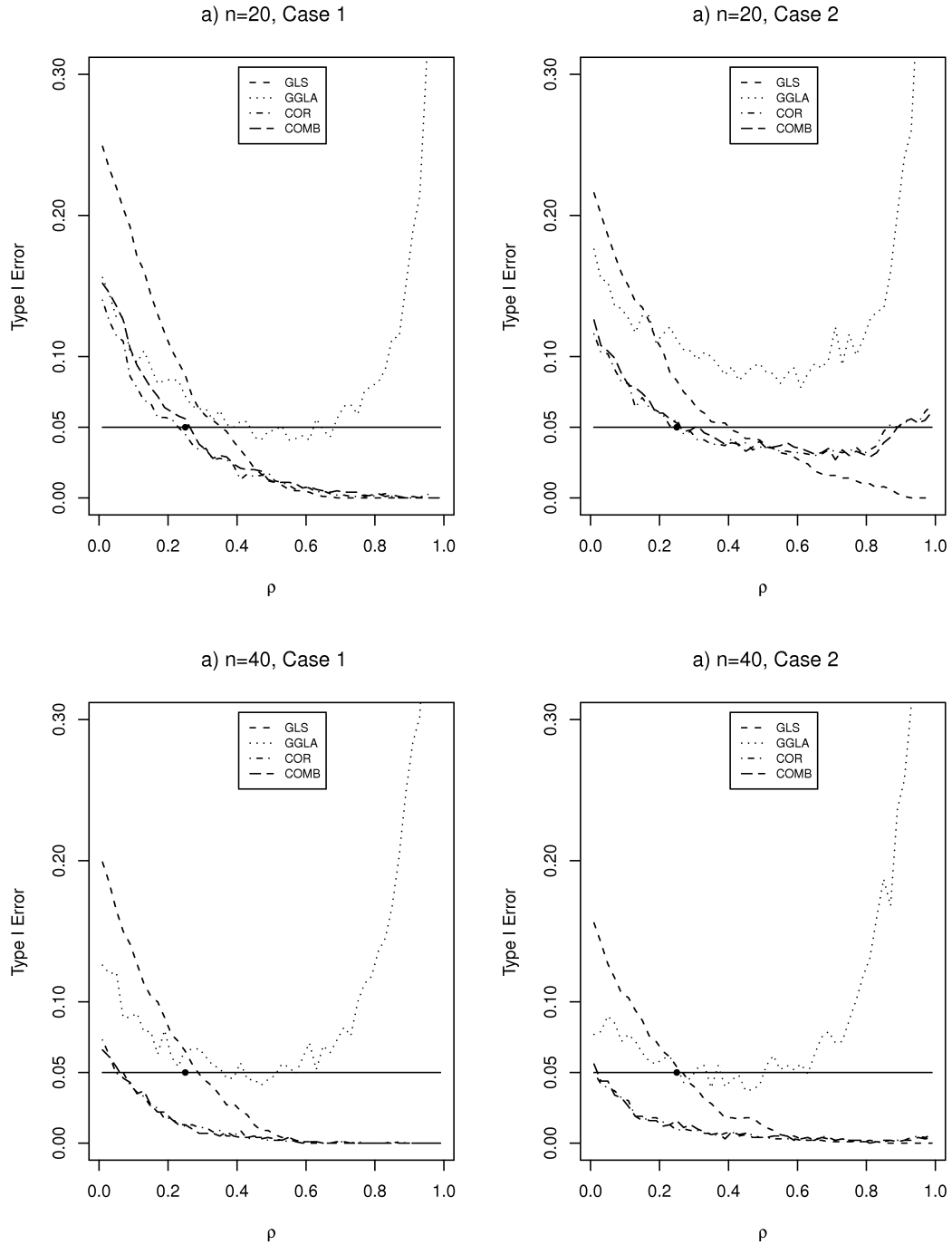


Figure 4.6: Proof-of-concept simulation results showing the Type-I error rates for the unadjusted method (short dashed line), corrected method (mixed line), GGLA method (dotted line) and combined method (long dashed line) ($\rho=0.25$). Methods with Type-I error rates below 0.05 for a given value of the intraclass correlation are conservative, while those that are above are liberal.

In addition to the results discussed previously, the Type-I error rates and coverage probabilities were calculated from the four different methods for the data-driven technique. The results, which can be seen in Figure 4.7, have both similarities and marked differences from the proof-of-concept simulations. The GGLA method of testing and constructing hypotheses still exhibits poor Type-I errors and coverage probabilities when compared to the nominal value. The tests and confidence intervals using the corrected variance performed close to nominal or conservative testing qualities, just as in the proof-of-concept simulations. However, the combined method is liberal when there are 20 blocks (Figure 4.7a,4.7b) and conservative when the number of blocks is 40 (Figure 4.7c,4.7d). This differs from the proof-of-concept results (Figure 4.5), where the combined method was more conservative than the generalised least squares method regardless of the number of blocks. Because of the limited block sizes studied, it remains to be seen if the combined method is superior to the generalised least squares method for any block size.

The other marked difference between the proof-of-concept and data-driven methods lies in the unexplained results displayed in Figure 4.7. As expected, these results made for very wide coverage probabilities, so much so that the coverage probability was 1 for the entire coverage probability profile.

4.6 Application to the Cervical Radiculopathy Trial

The strategies discussed in Section 4.4 can be used to adjust the analysis of the cervical radiculopathy trial, presented in Chapters 2 and 3, for model uncertainty. Using all of the parameters in the final model gives a better insight into the behavior of the GGLA estimate of the fixed-effect parameters in LMMs. Rather than focus on a single parameter, this analysis focuses on all four parameters of the hypothesis set and the eight remaining covariates in the testing set and will demonstrate the effect of using the GGLA in approximation.

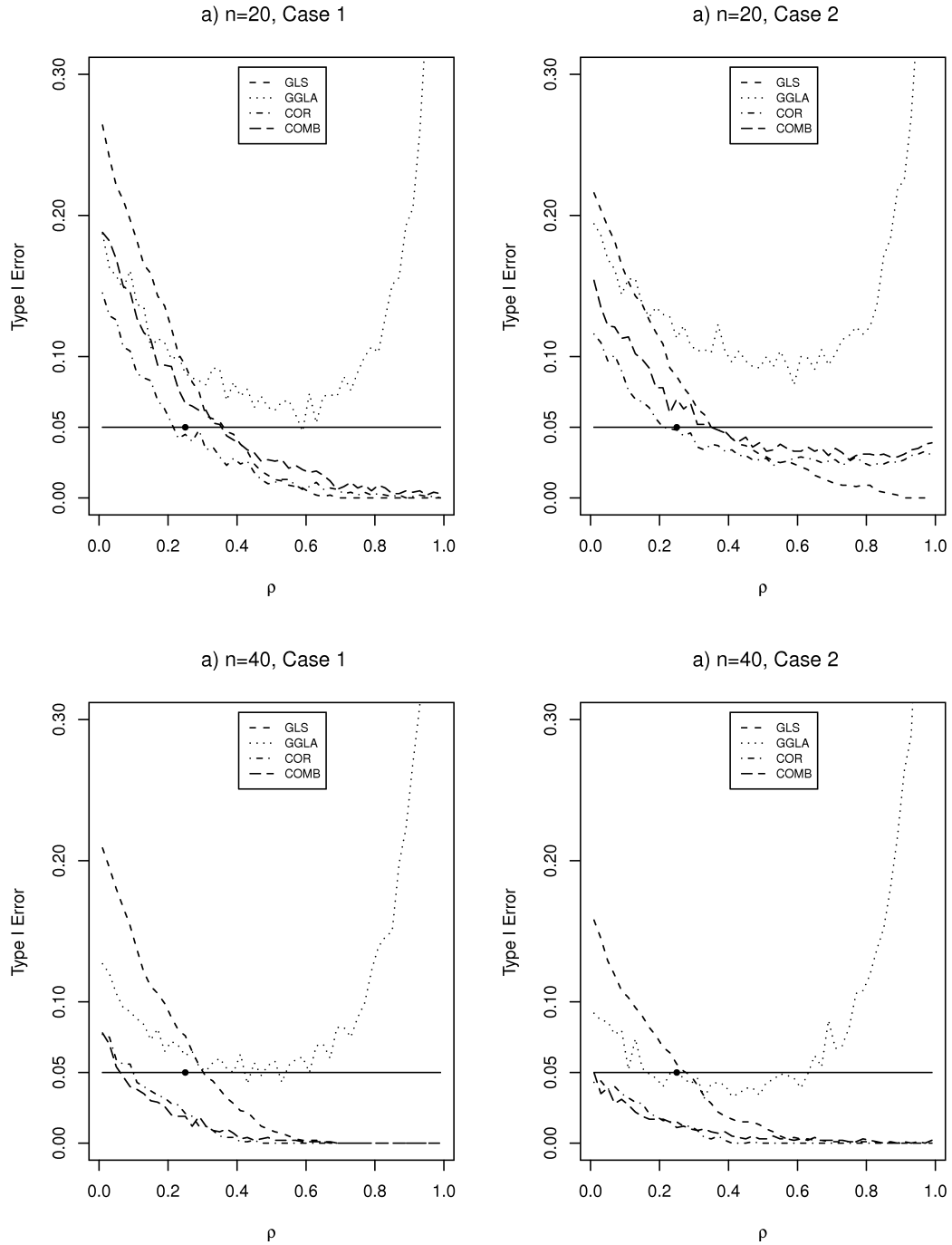


Figure 4.7: Data-driven simulation results showing the Type-I error rates for the unadjusted method (short dashed line), corrected method (mixed line), GGLA method (dotted line) and combined method (long dashed line) ($\rho=0.25$). Methods with Type-I error rates below 0.05 for a given value of the intraclass correlation are conservative, while those that are above are liberal.

Variable	Estimate	SE	Corrected Method			<i>d</i>
			<i>p</i> -value	95% CI		
Intercept	-0.20	3.90	0.959	(-7.85, 7.45)		0.01
Time	-0.97	1.01	0.339	(-2.96, 1.02)		0.19
Treatment Group	0.59	1.37	0.666	(-2.09, 3.27)		0.09
Treatment Group×Time	0.15	1.30	0.706	(-3.04, 2.06)		0.07
BMI	0.07	0.10	0.480	(-0.12, 0.26)		0.14
Previous Symptoms	1.21	1.29	0.349	(-1.32, 3.74)		0.19
FABQ	0.15	0.03	< 0.001	(0.09, 0.21)		0.90 *
Work Related Injury	-0.55	1.66	0.739	(-3.79, 2.69)		0.07
Baseline Pain	0.06	0.07	0.364	(-0.07, 0.16)		0.19
Bothersome Numbness	-2.01	1.38	0.145	(-4.71, 0.69)		0.29
Number of Treatments	-0.60	0.30	0.044	(-1.18, -0.02)		0.40 *
Baseline NDI	0.52	0.08	< 0.001	(0.37, 0.67)		1.33 *

Table 4.1: Multivariable analysis of the cervical radiculopathy data using the corrected method. The parameters estimates with a p -value < 0.05 are flagged by an asterisk (*).

The corrected variance of 29.46 was used in the corrected method to adjust hypothesis tests and confidence intervals for model uncertainty. The results from the corrected method can be seen in Table 4.1. As expected, the standard errors of the parameter estimates from the corrected method are larger than the estimates from the method without accounting from model uncertainty. These standard errors are 1.10 times larger than the values calculated without adjusting for model uncertainty, which mirrors the ratio of the square root of the corrected variance to the adjusted variance. The inflation of the standard errors results in larger p -values and effect sizes as well as wider confidence intervals. All of the variables flagged as significant (p -value < 0.05) from the originally analysis are similarly flagged in the corrected analysis.

Independent from the calculation of the GDF, $D=200$ Monte-Carlo simulations were performed for the perturbation method to estimate the GGLA model using a perturbation size of 0.5. The corrected variance of 29.46 was used for the value of σ_*^2 . The result of the analysis from the GGLA and combined methods can be seen

in Table 4.2. The estimates of the fixed-effect parameters vary around the values computed from the generalised least squares estimates, particularly for parameters with effect sizes near zero. This variation is so large that the interpretation of the effect of the corresponding variable on NDI often changes, making the interpretability unreliable. An example of this is through the variable indicating whether the injury was work related. With just using the generalised least squares estimates, the fixed-effect parameter estimate is positive, but using the GGLA method, the fixed-effect parameter estimate is negative. A cursory experimentation with different seed values in the Monte Carlo simulation offer a clear demonstration of this effect. Conversely, when the effect size of a variable is large, the GGLA estimate of the fixed effect is fairly robust to the change in seed and consistent in its estimation of the mean value. The variation of the estimates is not surprising in light of the simulation results discussed in the previous section. However, it is surprising that the estimates with high effect sizes demonstrate a robustness against the variation.

The standard errors of the parameter estimates from the GGLA method are larger than both the the estimates from the method without accounting from model uncertainty and the corrected method. The standard errors from the GGLA method are between 1.29 and 1.72 times larger than the unadjusted method. This results in higher p -values, wider confidence intervals and smaller effect sizes than the aforementioned methods. The results from the hypothesis test changed for the effect of the number of treatments so that this variable is no longer flagged as an important predictor of NDI. This is due decrease in the estimate of this variable, by almost 66%, and the inflated variance due to adjusting for model uncertainty.

The results of the combined method, which can also be seen in Table 4.2, are similar to that of the corrected method. Since the standard errors are the same as the GGLA method, the combined method has larger p -values and effect sizes and wider confidence intervals than either the unadjusted or corrected method. The parameter

Variable	GGLA Method				Combined Method			
	Estimate	SE	P-value	95% CI	d	Estimate	SE	p -value
Intercept	-2.47	4.57	0.588	(-11.43, 6.48)	0.11	-0.20	4.57	0.965
Time	-2.43	1.43	0.089	(-5.23, 0.36)	0.34	-0.97	1.43	0.497
Treatment Group	0.69	1.84	0.706	(-2.92, 4.31)	0.07	0.59	1.84	0.747
Treatment Group \times Time	-0.32	2.04	0.875	(-4.31, 3.67)	0.03	-0.49	2.04	0.811
BMI	0.14	0.12	0.222	(-0.09, 0.37)	0.24	0.07	0.12	0.564
Previous Symptoms	2.13	1.52	0.161	(-0.85, 5.10)	0.28	1.21	1.52	0.424
FABQ	0.11	0.04	0.001	(0.03, 0.54)	0.54 *	0.04	0.04	< 0.001
Work Related Injury	0.78	2.08	0.709	(-3.30, 4.86)	0.07	-0.55	2.08	0.792
Baseline Pain	0.06	0.10	0.499	(-0.11, 0.23)	0.13	0.06	0.10	0.466
Bothersome Numbness	-2.16	1.91	0.258	(-5.90, 1.58)	0.11	-2.01	1.91	0.292
Number of Treatments	-0.21	0.39	0.588	(-0.97, 0.11)	0.22	-0.60	0.39	0.121
Baseline NDI	0.50	0.12	< 0.001	(0.31, 0.70)	1.00 *	0.52	0.12	< 0.001
								(0.32, 0.7)

Table 4.2: Multivariable analysis of the cervical radiculopathy data using the GGLA and combined method. The parameters estimates with a p -value < 0.05 are flagged by an asterisk (*).

Quantity	GLS Method	Corrected Method	GGLA Method	Combined Method
Omnibus P-value	0.873	0.894	0.930	0.945
Treatment Group (2 wks)	14.8	14.8	15.9	14.8
95% CI	(12.3,17.2)	(12.2, 17.3)	(12.3, 19.6)	(11.1, 18.4)
Control Group (2 wks)	14.2	14.2	15.2	14.2
95% CI	(11.7, 16.6)	(11.6, 16.7)	(11.6, 18.9)	(10.5, 17.8)
Difference (2 wks)	0.6	0.6	0.7	0.6
95% CI	(-1.8, 3.0)	(-2.0, 3.1)	(-2.9, 4.3)	(-3.0, 4.2))
Treatment Group (4 wks)	13.3	13.3	13.2	13.3
95% CI	(10.9, 15.8)	(10.8, 15.9)	(9.5, 16.8)	(9.6, 16.8)
Control Group (4 wks)	13.2	13.2	12.8	13.2
95% CI	(10.8, 15.6)	(10.7, 15.8)	(9.2, 16.4)	(9.6, 16.8)
Difference (4 wks)	0.1	0.1	0.4	0.1
95% CI	(-2.3, 2.5)	(-2.4, 2.7)	(-3.3, 4.0)	(-3.5, 3.7)

Table 4.3: Omnibus test for the effect of cervical traction and confidence intervals for the estimated NDI for each treatment group at 2 and 4 weeks using all methods.

estimates that come from the generalised least squares method are more representative of the true effects and make for much more meaningful inference. The increase in the standard errors was large enough to make the parameter estimate corresponding to the number of treatments no longer significantly different from zero (p -value=0.121), leading to a different conclusion than either the unadjusted or corrected variance.

Lastly, the omnibus tests for a test of no overall treatment effect and the confidence intervals for the NDI for the treatment groups at the 2 and 4 week follow-up visits for each of the four methods can be seen in Table 4.3. All of the proposed methods have a p -value that is higher than the GLS methods, which is reflective of the inflated variance used in these methods. The GGLA method reports nominally higher parameter estimates for the 2-week predictions compared to the methods using the generalised least squares estimates. The GGLA and combined methods have wider confidence intervals than the GLS and corrected method. However, all of the methods are in agreement that cervical traction shows little evidence in improving NDI scores.

4.7 Discussion and Conclusion

This chapter extended the concept of accounting for model uncertainty introduced in Chapter 3 and to a vector of functions, or statistics, of the response variable. The variance of these statistics are potentially greater for an unknown model than if the true model were known *a priori*. A simple way to adjust for model uncertainty is to use the generalised least squares estimates of the fixed-effects and corrected variance discussed in the previous chapter.

A separate model, known as the general global linear approximation (GGLA), was introduced to account for model uncertainty due to its attractive qualities to inflate the variance for inflexible models. This model provides unbiased estimates for the vector of statistics that is being estimated and will have variance no smaller than estimates from the generalised least squares estimates. This method was used to incorporate model uncertainty in the fixed-effect parameter estimates. The estimates of the GGLA method were calculated using information from the calculation of the GDF, thus enabling a similar estimation method.

The proof-of-concept simulation results show that, even though the method produces estimates that are unbiased over many simulations, the individual estimates are inflated by an order of magnitude that limits their usefulness in hypothesis tests, confidence intervals, and more importantly, interpretability. However, the variance of these quantities were appreciably inflated. Of the four methods proposed, the corrected and combined methods were able to effectively make the Type-I errors and coverage probabilities closer to the nominal values. Due to the more extreme estimates produced from the GGLA method, the method solely based on the GGLA failed to do achieve these goals.

The data-driven methodology diverged from the conclusions demonstrated by the proof-of-concept simulations. When only the parameters of the perturbation method were derived from the data, the GGLA and combined methods failed to demonstrate

an ability to adjust for model uncertainty in hypothesis tests and Type-I errors. The most robust method when using a purely data-driven approach to accounting for model uncertainty in inference and model selection is the corrected method, as the estimates of the Type-I error and coverage probability were closer to the nominal value than all of the other methods

When applied to the cervical radiculopathy trial, the proposed methods showed another quality not assessable from the simulation study. For parameters with large effect sizes, the sensitivity of the GGLA estimates decreased so that they were an accurate reflection of the relationship between the variable and NDI. The methods all estimated similar conclusions regarding the lack of efficacy of the cervical traction treatment and provided similar confidence intervals for the NDI each of the treatment groups at both 2 and 4 weeks follow-up.

The proposed methodology made use of asymptotic theory to come to certain conclusions regarding the form of the hypothesis tests and confidence intervals. However, the simulation results discussed in Chapter 2 and the theoretical conclusions made in Chapters 3 and 4 show that as the sample size increases, the effects of model uncertainty will become less pronounced due to the consistent or efficient nature of model selection procedures. Thus, asymptotic theory does not really have an application when model uncertainty is most prevalent. Exact, or at least non-asymptotic, methods can be explored to address this situation and produce a better theoretic background and provide a relevant foundation for inference and the construction of confidence intervals.

The estimates for the parameters in the GGLA were estimated in a manner consistent with the estimation of the GDF. Superior estimates may be found by finding optimal estimates based on a loss function, such as the squared-error loss. These estimates may be able to correct the extreme values of the GGLA estimates and poor performance when applied with a data-driven method. This strategy will be topic of

future endeavors.

CHAPTER V

Discussion

Statistical knowledge has become so commonplace in the medical research community that almost all researchers and clinicians have some appreciation of statistical foundations. Underlying all of these foundations is a statistical model, from which estimation, inference and prediction are drawn. However, little research has been done exploring the effects of uncertainty in the statistical model, particularly in linear mixed-effects models. This work explored and addressed some of these issues.

One particular instance where model uncertainty is implied is in model selection, where data driven methods are used to recommend the models that balance between biased estimates and inflated variances. A statistical model for model selection in linear mixed-effect models was proposed. This model enabled fixed-effect selection, random-effect selection, covariance structure selection, or any combination of these, to be assessed mathematically. From this model, the expected values of the estimates of the fixed-effects, residual variance and variance of the fixed-effects was explored. While the generalised least square estimates of the fixed-effect parameters were unbiased in the presence of model uncertainty, the two variance terms are potentially biased in the presence of model uncertainty. Hypothesis tests and confidence intervals calculated from these estimates diverged from their nominal qualities so they were more likely to suggest a difference from a pre-specified value. The divergence of the

hypothesis tests and confidence intervals from their stated qualities can have serious clinical consequences.

The generalized degrees of freedom were used to correct the residual variance for model uncertainty. This quantity was proposed by *Zhang et al.* (2012) and measures a model's flexibility to perturbations in the response variable. It was re-purposed to adjust for model uncertainty by substituting it for the model degrees of freedom in the generalised least squares estimate of the residual variance so that inflexible, poor-fitting, models will have higher variance estimates. The inflated variance estimated, referred to as the corrected variance, was meant to correct the liberal qualities of the hypothesis tests and the overly narrow confidence intervals that result when using the adjusted variance.

An extension of the general global linear approximation used by *Shen et al.* (2004a) to adjust linear regression models was proposed to perform a similar adjustment for linear mixed-effects models. This method enabled the unbiased estimation of an arbitrary vector of statistics and had a variance that is related to the model fit: both important qualities for adjusting for model uncertainty. Estimation of the parameters in this model was based on the conclusions from generalized degrees of freedom so that a similar estimation method could be used.

The perturbation method of estimation was used for both the generalized degrees of freedom and the general global linear approximation. This method used a Monte Carlo simulation to create a pseudo-response variable. A statistic that uses the pseudo response variable and is conditional on the observed responses was used to estimate each of the respective quantities. Since the generalized degrees of freedom and parameters of the general global linear approximation were able to be estimated using easily tractable and estimable functions of the response variable, this method was well suited for that task.

These methods were evaluated using two different simulation paradigms, each with

multiple simulation parameters. The simulation paradigms differed only by the standard deviation used in the perturbation method. As a proof-of-concept simulation, the value of the standard deviation used in creating the pseudo-response vector was the value used in simulating the error distribution. A data-driven method was also used to create the pseudo-response vector, with the generalised least squares estimate of the standard deviation used to for the calculation of the GDF and the corrected variance used for the general global linear approximation.

Overall, the results of these methods were mixed. The proof-of-concept simulations showed promise for both the GDF calculation and general global linear approximation. The GDF values were higher than the average dimension of the model, which translates to an inflation of the residual variance due to model uncertainty. When this corrected variance was used in testing hypotheses and construction of confidence intervals, the Type-I errors and coverage probabilities tended to be closer to their respective nominal value than the generalised least squares results that were unadjusted for model uncertainty.

The fixed-effects estimated from the general global linear approximation were an unbiased estimate of the fixed-effect parameters, but larger values in magnitude when compared to the generalised least squares estimates. Even though the standard errors of these estimates were larger than the those produced by the generalised least squares, the testing qualities were poorer due to the extreme values in the estimates themselves. However, when the generalised least squares estimates of the fixed-effects with the standard errors from the general global linear approximation, the testing qualities improved, particularly for small samples.

The data-driven simulations provided less than ideal results. For small values of the true intraclass correlation, the generalized degrees of freedom were able to inflate the corrected variance. When the true intraclass correlation values were large, this was not true; the corrected variance was equal to or even smaller than the

adjusted variance. Any method using the general global linear approximation did not perform well using the data driven method, although the combined method was typically superior to the unadjusted method for small samples.

Additionally, the data from a clinical trial evaluating the efficacy of the cervical traction in treating cervical radiculopathy was used to assess the proposed methodology. The corrected variance showed an inflation compared to the adjusted variance. The testing methods all inflated the standard errors, resulting in higher p -values, wider confidence intervals and smaller effect sizes compared to a method that does not adjust for model uncertainty. This trial also demonstrated the robustness of the estimates of the general global linear approximation with large effect sizes and the sensitivity of the estimates with small effect sizes. Except for the effect of the number of treatments on the Neck Disability Index, the proposed methods came to similar conclusions compared to the original study.

Even though the results in evaluating the proposed methods are mixed, this work does make an impact in the field of quantifying model uncertainty. No other known work has attempted quantify and incorporate model uncertainty when there is a relationship between response variables. Since the active researchers in this field are relatively few, progress toward a comprehensive solution to this issue will be slow to occur as knowledge and insight slowly accumulates. As such, this research is expected to significantly impact the field and lead to further development of this topic.

As with any work, practical concerns limited the breadth and scope of topics. Some of these limitations are pertinent only to the separate methods and were discussed in the previous chapters. Others are more general and warrant a general discussion that brings in qualities of the entire work. One general limitation of the simulation results in this study is the large number of combinations that were able to be changed. Simulation parameters, such as total number of covariates, block size, error variance of the response and time independent covariates were held fixed due

strictly to space and time concerns. It has been assumed that the results discussed here extend to differences in these parameters, however, these need to be shown empirically and are left to future work.

Relatedly, the elements of the covariance matrix were assumed to be known, up to a constant. This is a huge limitation that will form the basis of future work. In using the generalized degrees of freedom, an assumption was made that model the estimate of the intraclass correlation value, which would be used in a clinical setting, were robust against model uncertainty. There is no basis behind this assumption, as this problem was outside the scope of this work. If this assumption does not hold, or does not hold for any covariance structure, novel methodologies must be developed to propose a solution to this problem.

Another possible future direction of this work is to extend the scope of potential models was beyond linear mixed-effect models. Fitting models without assuming a normal distribution on the error structure is becoming more and more common. Methods to adjust these models when the response variable is not considered independent have not been developed and remains a potential path of future work. Relatedly, there is no scientific reason to suggest that linear models are superior to other models such as non-linear mixed-effect models or generalized linear mixed-effect models. Thus, methods can be developed to include an adjustment for possibly not using the correct *type* of model for an analysis.

This work motivated the concept of model uncertainty through model selection procedures. However, this is not a requisite procedure when adjusting for model uncertainty. Except in extremely well-controlled experiments in laboratory sciences such as chemistry and physics, the models typically used in an analysis are not known and only considered an approximation. Thus, the methods developed in this paper, as well as some of the other works referenced here, are able to be used even when a model selection procedure is not performed.

APPENDICES

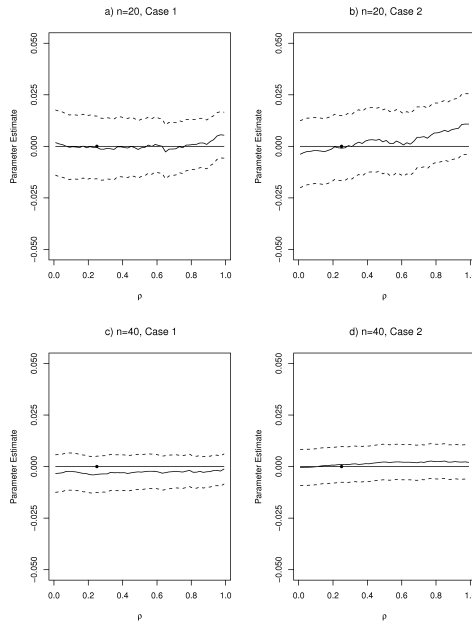
APPENDIX A

Extended Simulation Results

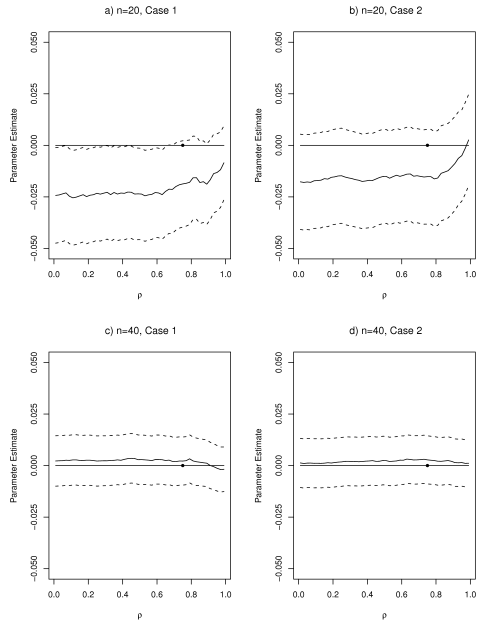
A.1 Fixed Effect Estimation

A.1.1 No Parsimony Correction

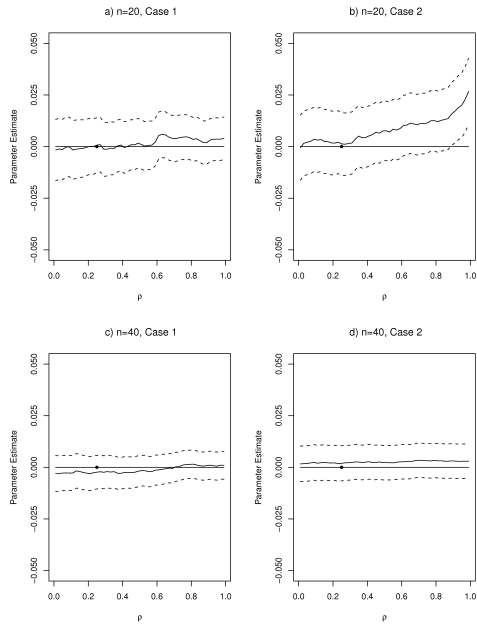
A.1.1.1 Cases 1 and 2



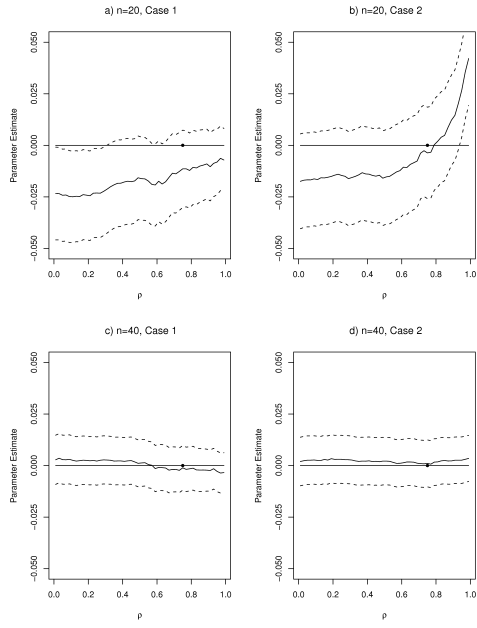
AIC : $\rho=0.25$



AIC : $\rho=0.75$

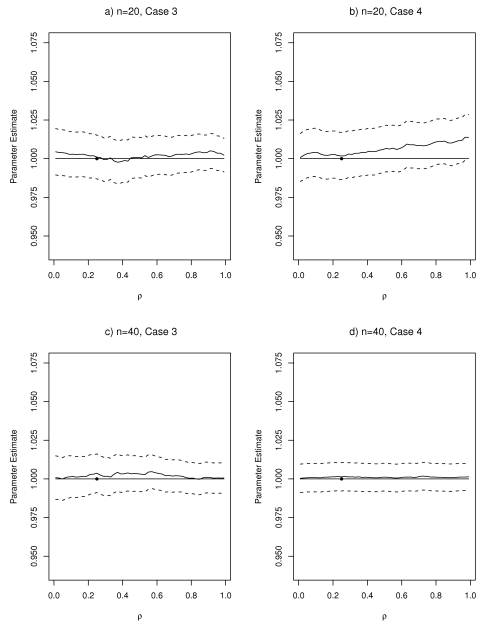


BIC : $\rho=0.25$

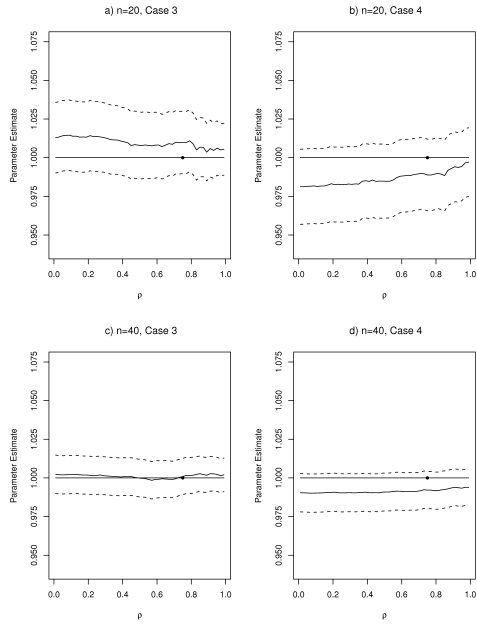


BIC : $\rho=0.75$

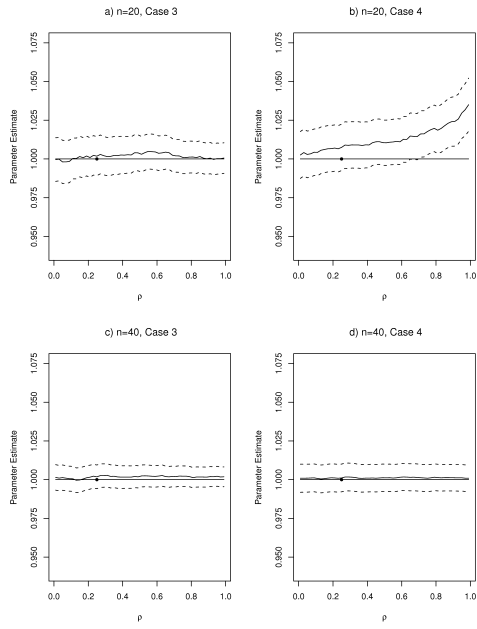
A.1.1.2 Cases 3 and 4



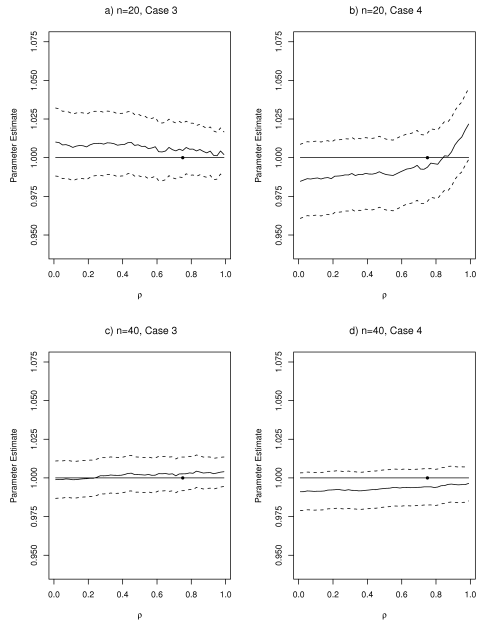
AIC : $\rho=0.25$



AIC : $\rho=0.75$



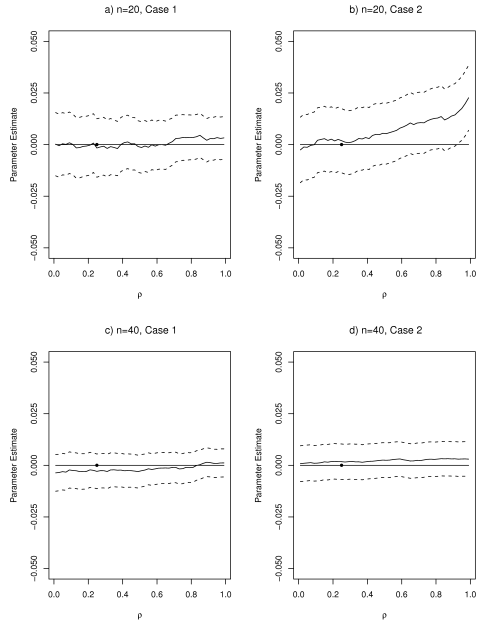
BIC : $\rho=0.25$



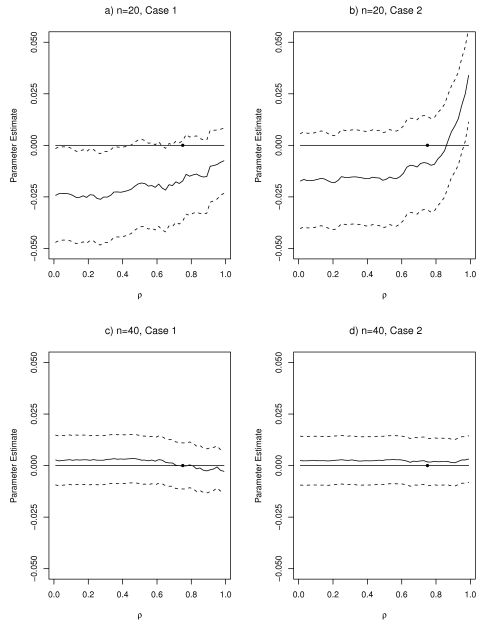
BIC : $\rho=0.75$

A.1.2 With Parsimony Correction

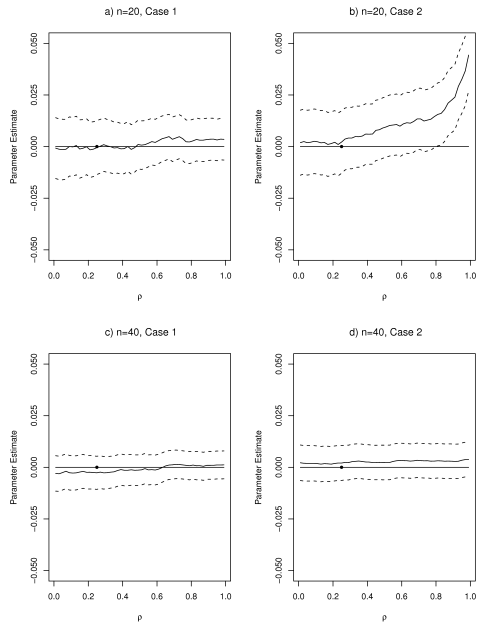
A.1.2.1 Cases 1 and 2



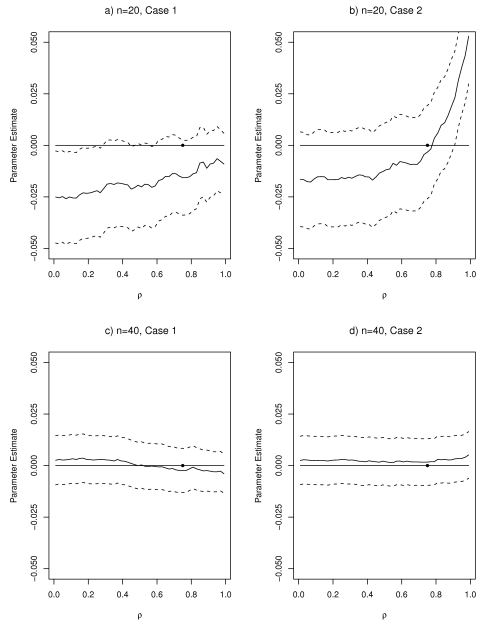
AIC : $\rho=0.25$



AIC : $\rho=0.75$

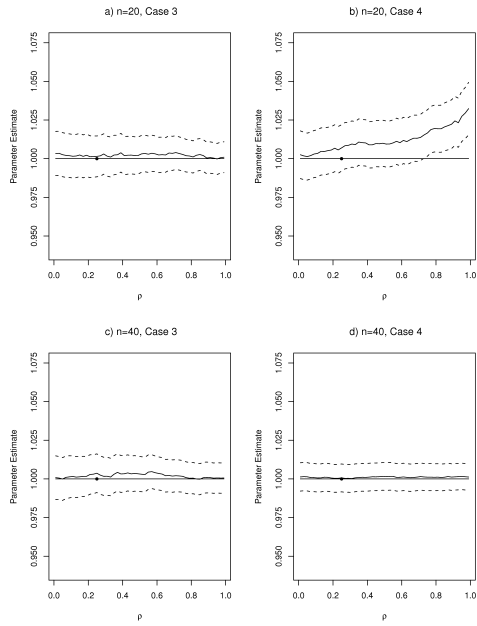


BIC : $\rho=0.25$

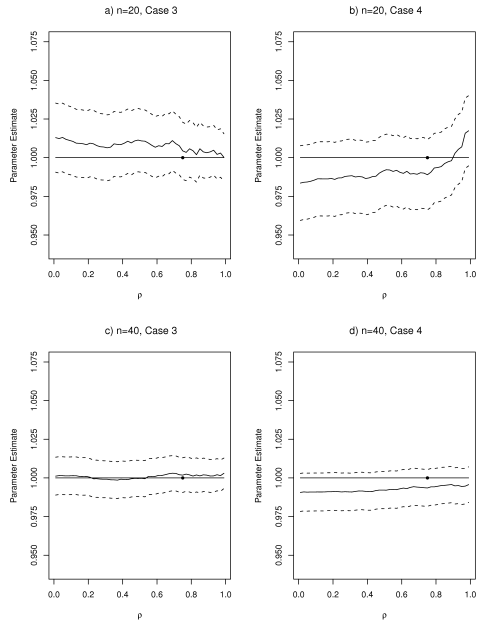


BIC : $\rho=0.75$

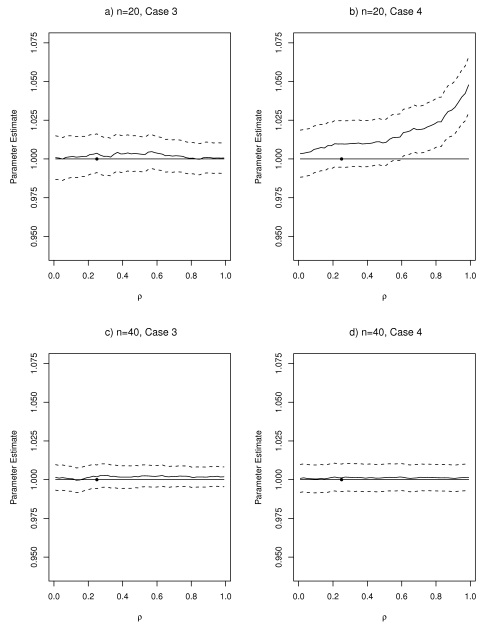
A.1.2.2 Cases 3 and 4



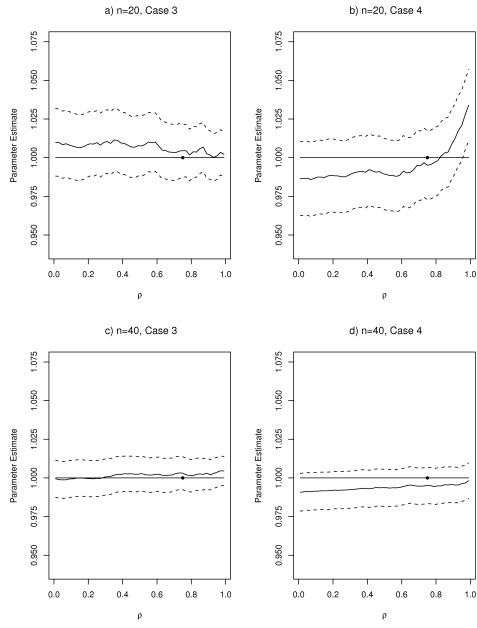
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

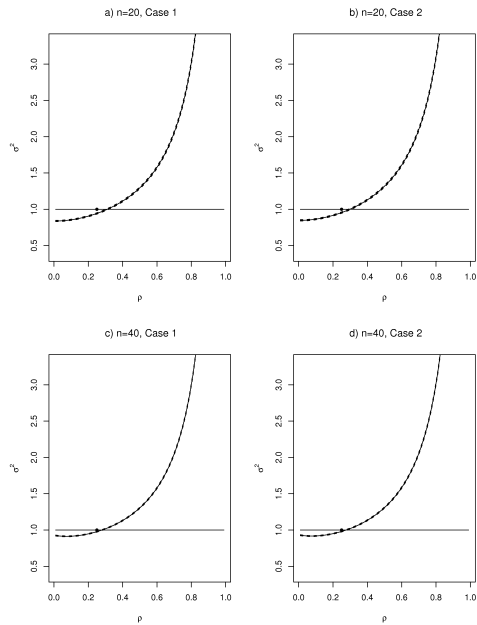


BIC : $\rho=0.75$

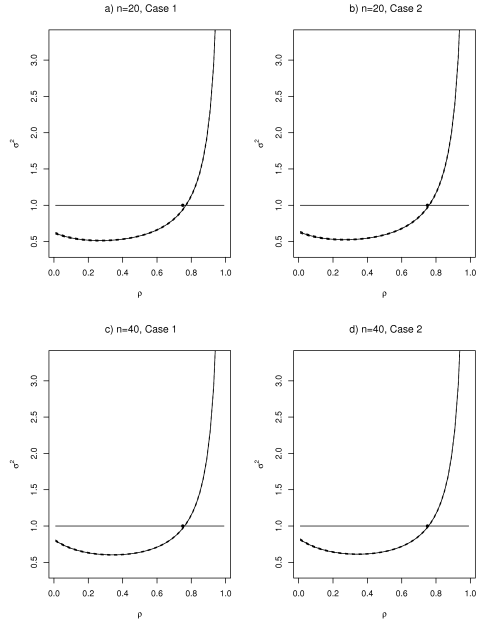
A.2 Residual Variance Estimation

A.2.1 No Parsimony Correction

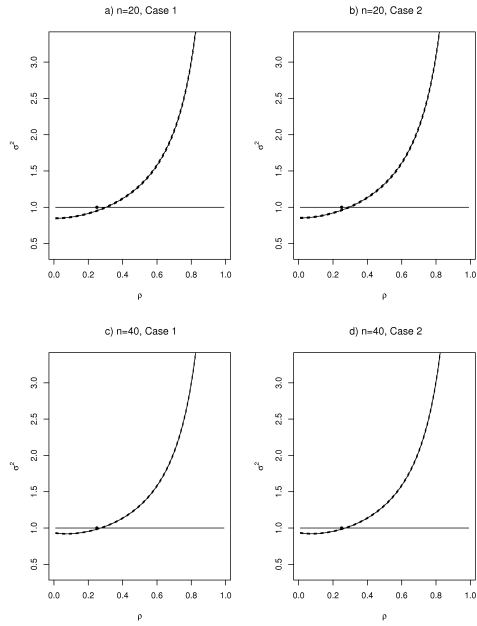
A.2.1.1 Cases 1 and 2



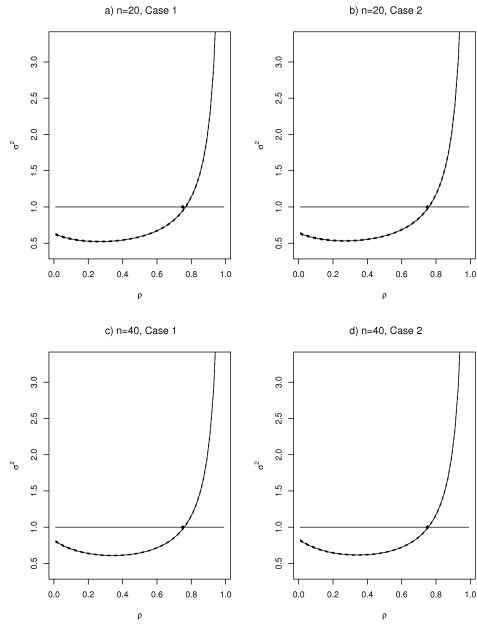
AIC : $\rho=0.25$



AIC : $\rho=0.75$

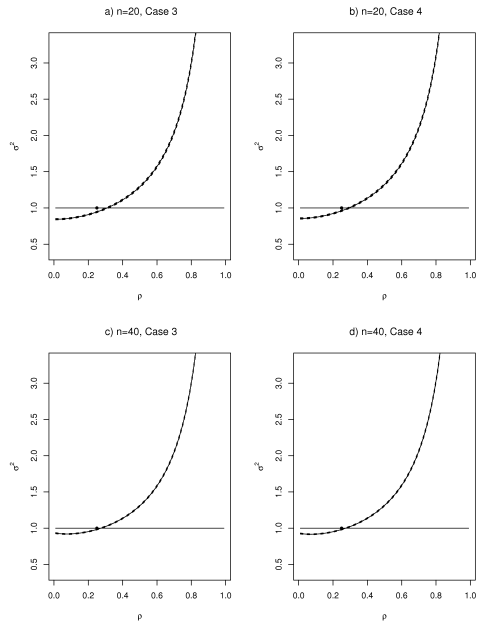


BIC : $\rho=0.25$

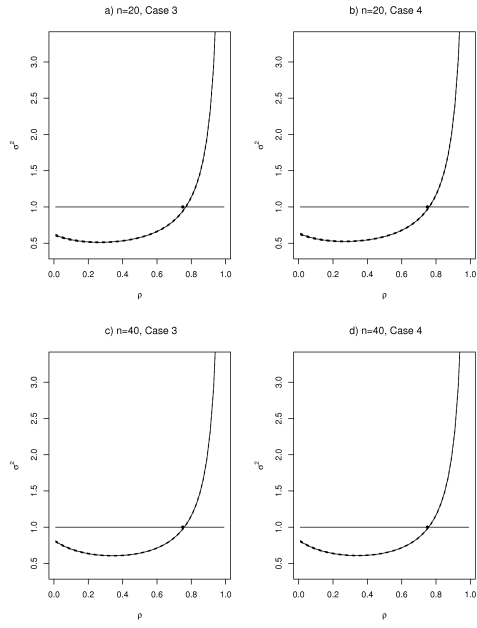


BIC : $\rho=0.75$

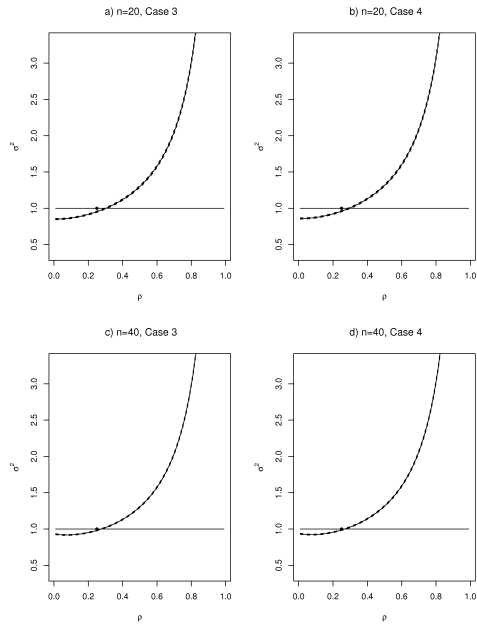
A.2.1.2 Cases 3 and 4



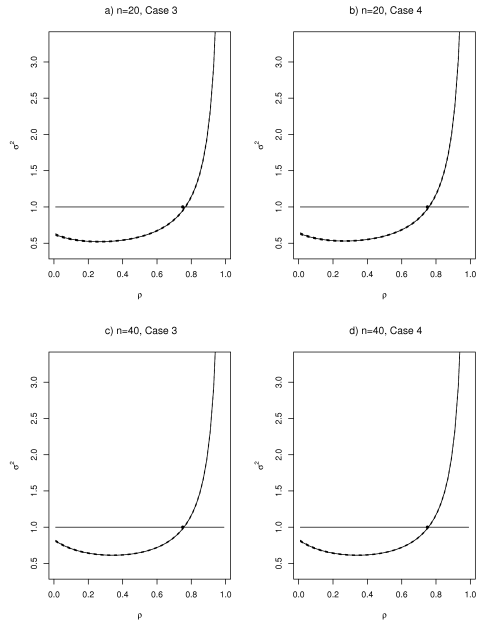
AIC : $\rho=0.25$



AIC : $\rho=0.75$



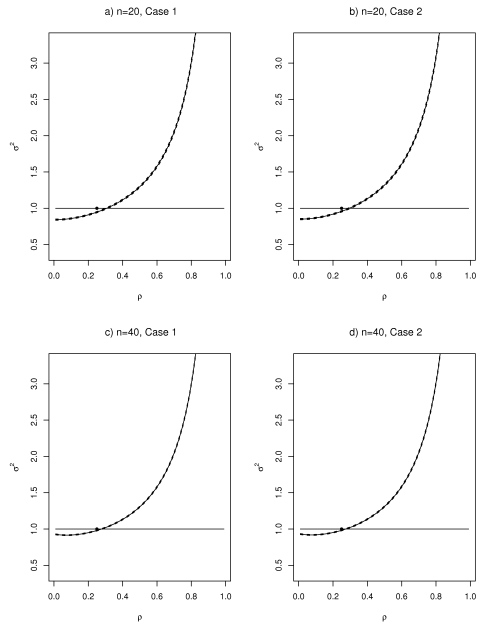
BIC : $\rho=0.25$



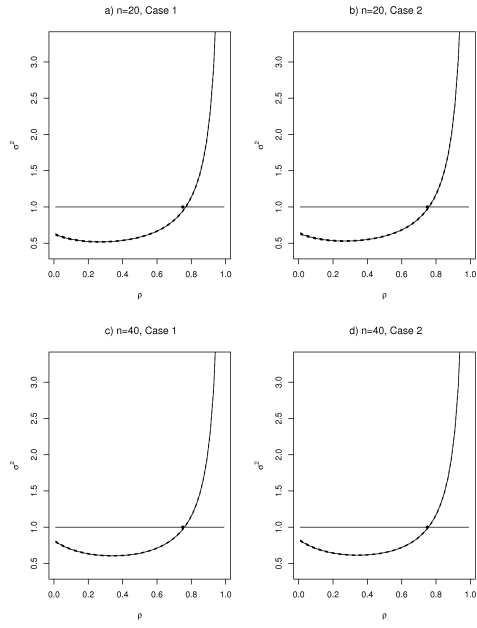
BIC : $\rho=0.75$

A.2.2 With Parsimony Correction

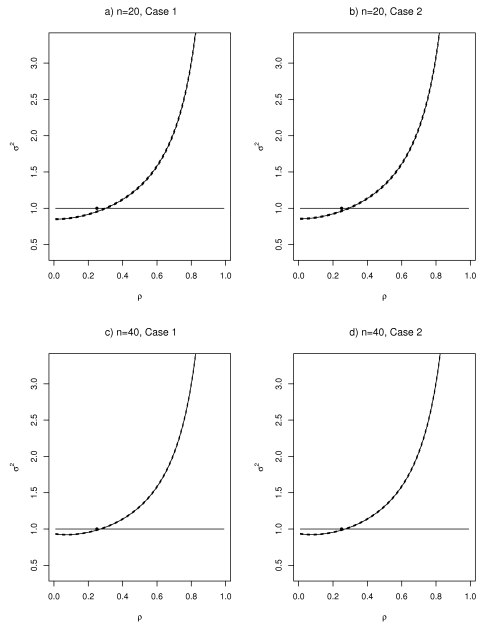
A.2.2.1 Cases 1 and 2



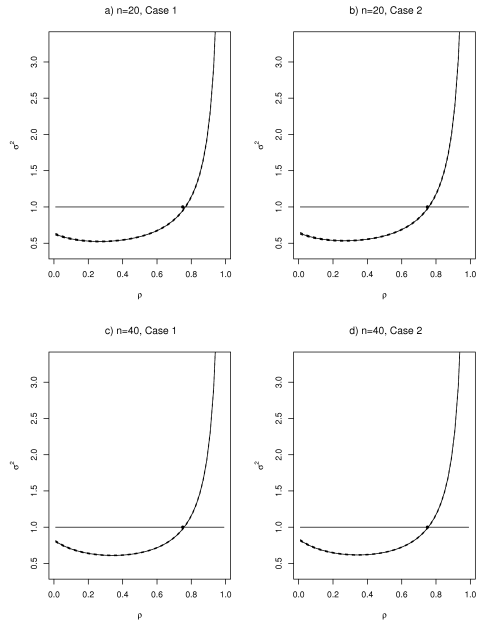
AIC : $\rho=0.25$



AIC : $\rho=0.75$

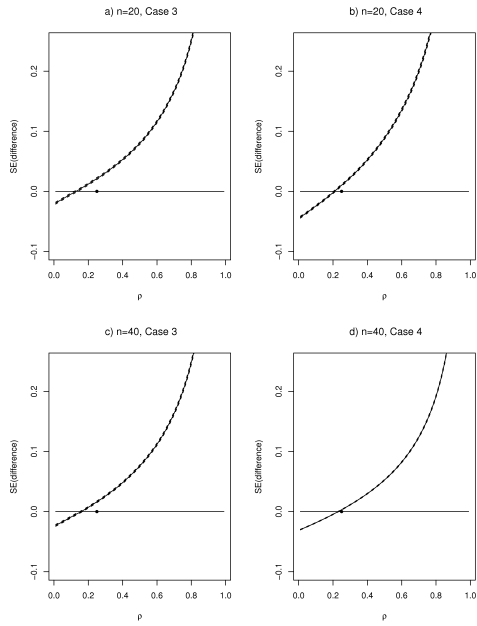


BIC : $\rho=0.25$

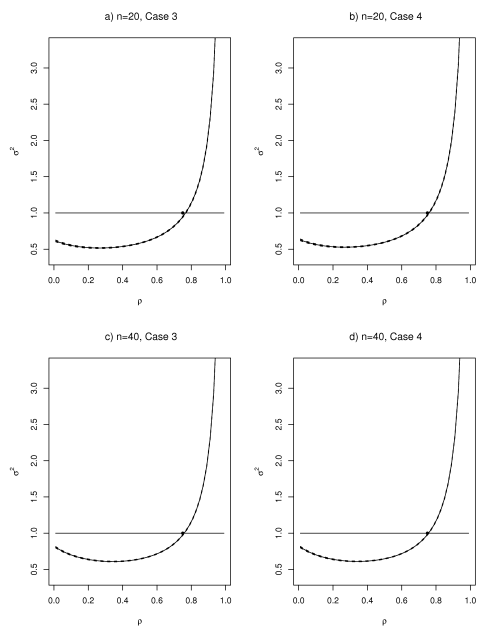


BIC : $\rho=0.75$

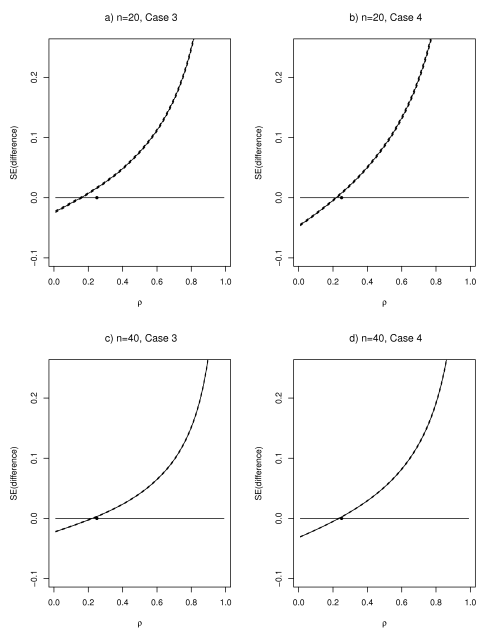
A.2.2.2 Cases 3 and 4



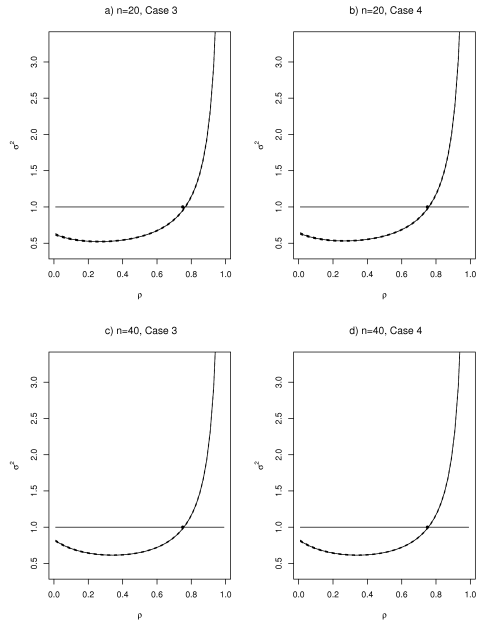
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

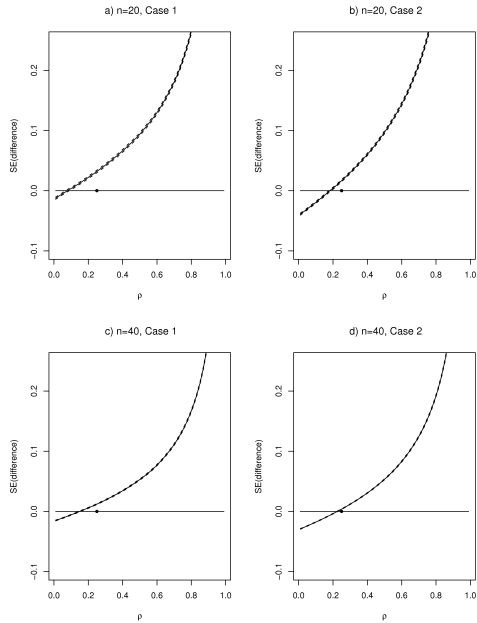


BIC : $\rho=0.75$

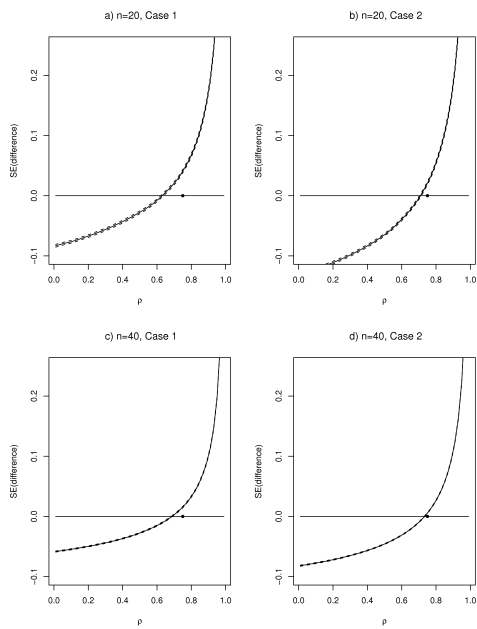
A.3 Standard Error Estimation

A.3.1 No Parsimony Correction

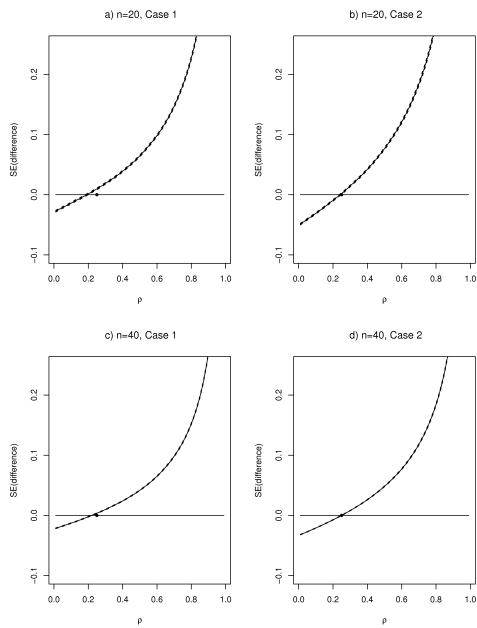
A.3.1.1 Cases 1 and 2



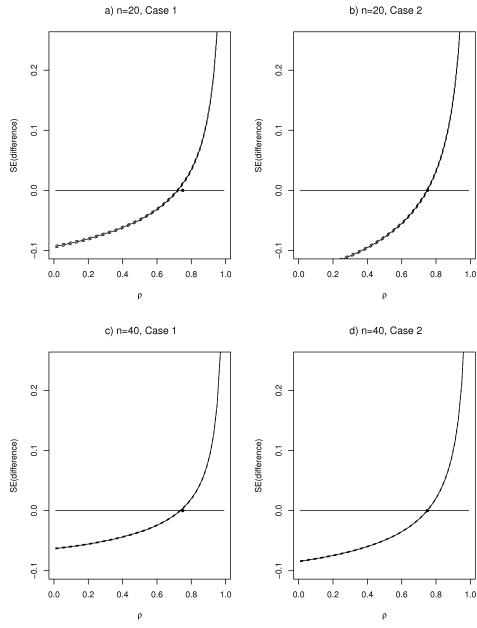
AIC : $\rho=0.25$



AIC : $\rho=0.75$

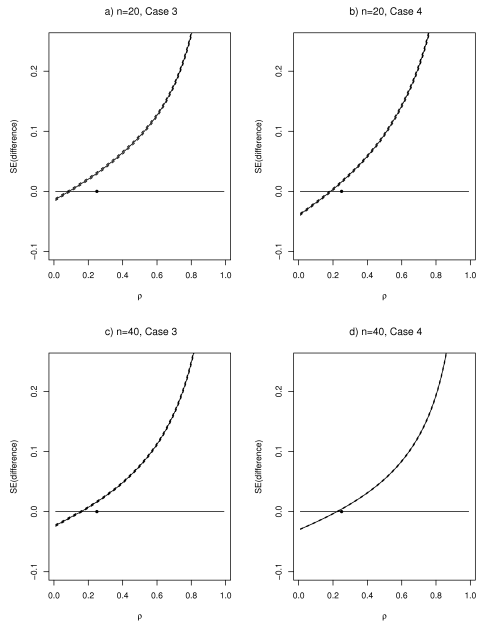


BIC : $\rho=0.25$

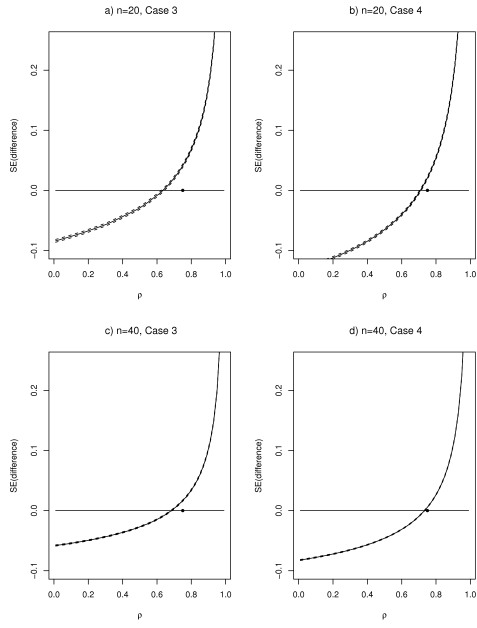


BIC : $\rho=0.75$

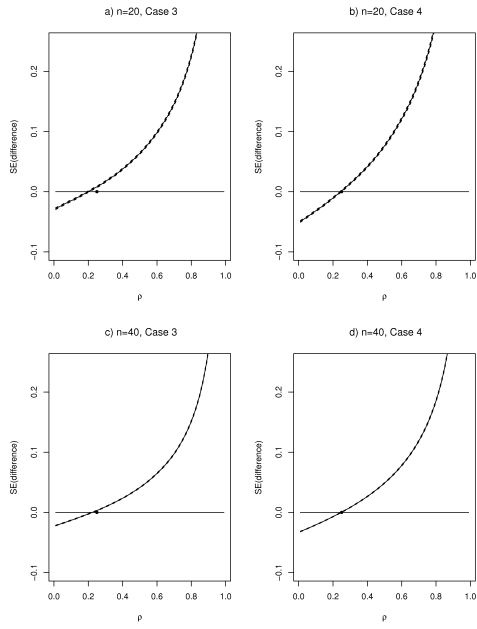
A.3.1.2 Cases 3 and 4



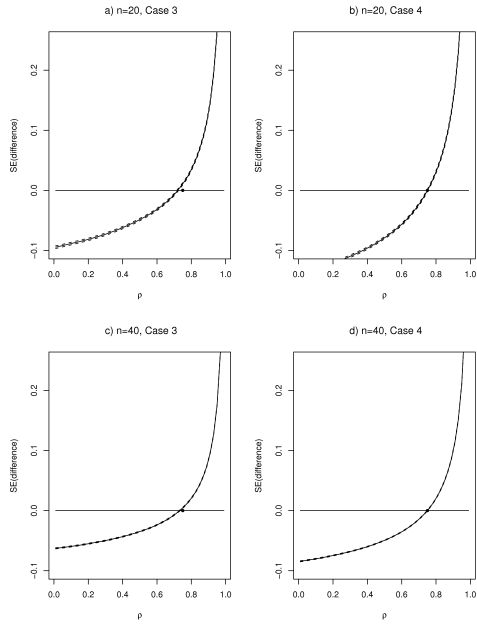
AIC : $\rho=0.25$



AIC : $\rho=0.75$



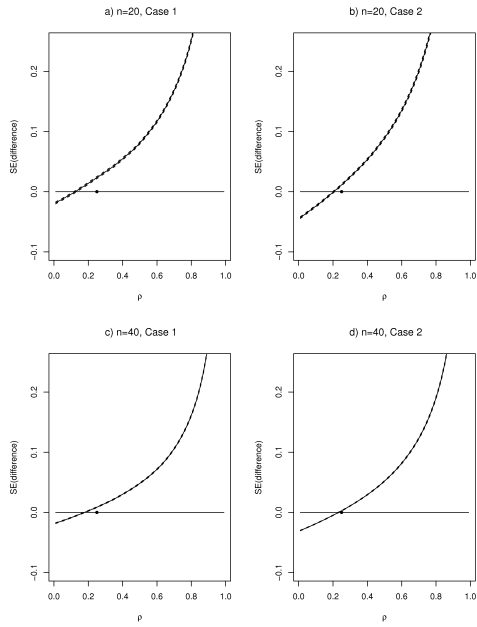
BIC : $\rho=0.25$



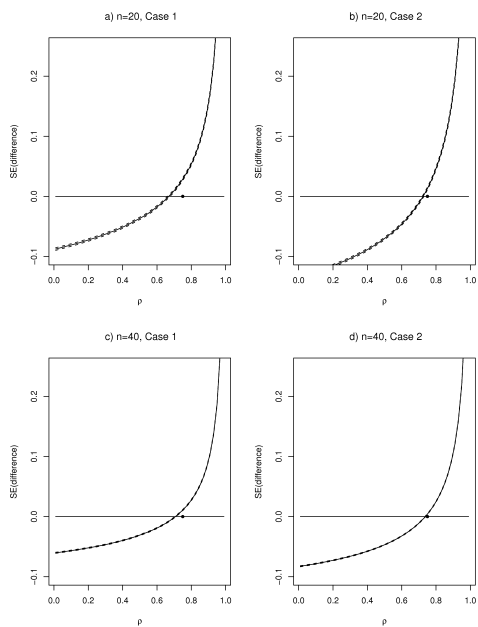
BIC : $\rho=0.75$

A.3.2 With Parsimony Correction

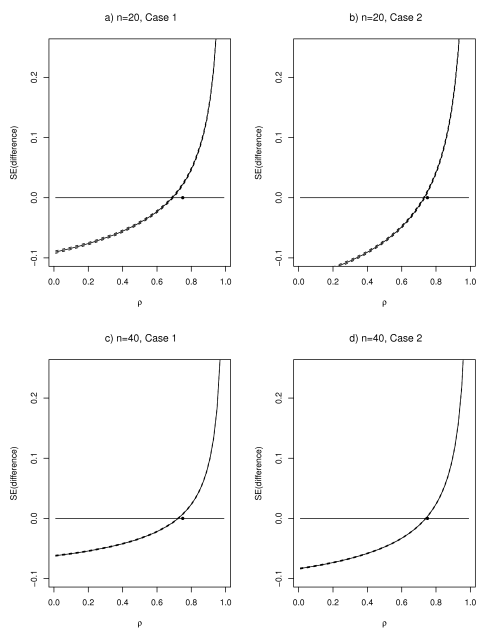
A.3.2.1 Cases 1 and 2



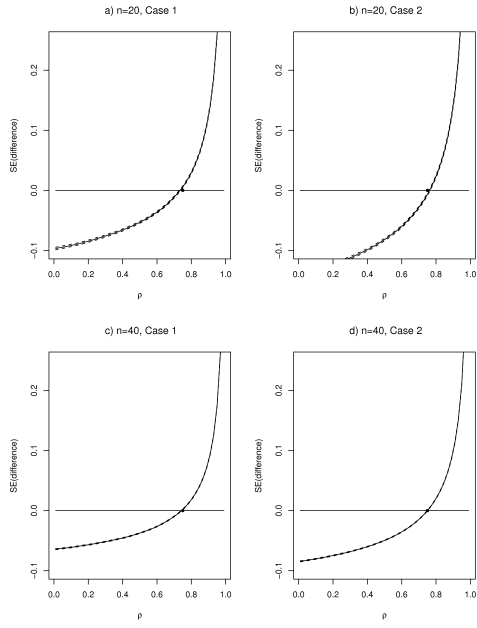
AIC : $\rho=0.25$



AIC : $\rho=0.75$

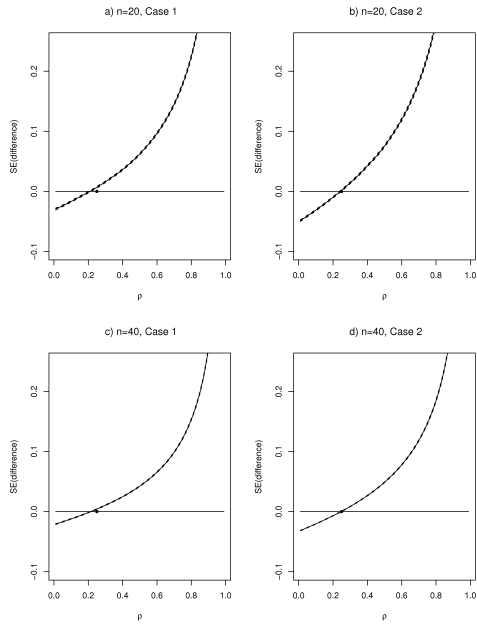


BIC : $\rho=0.25$

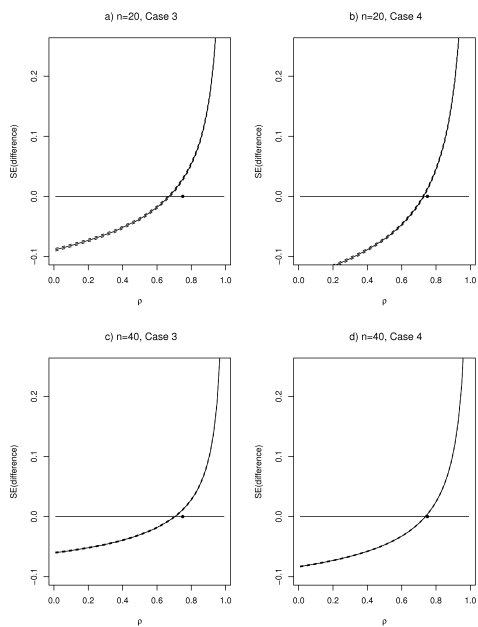


BIC : $\rho=0.75$

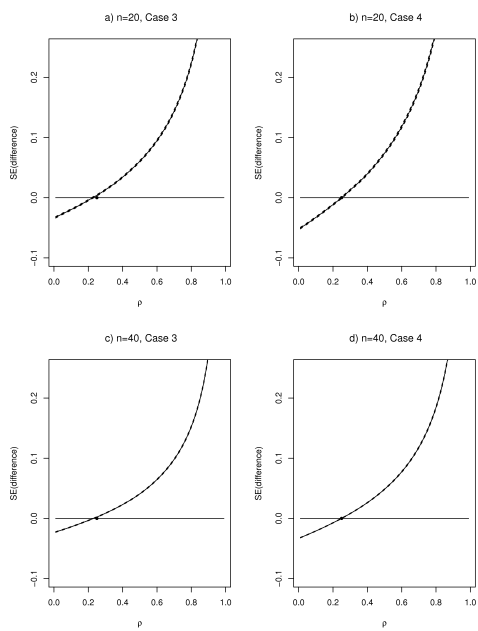
A.3.2.2 Cases 3 and 4



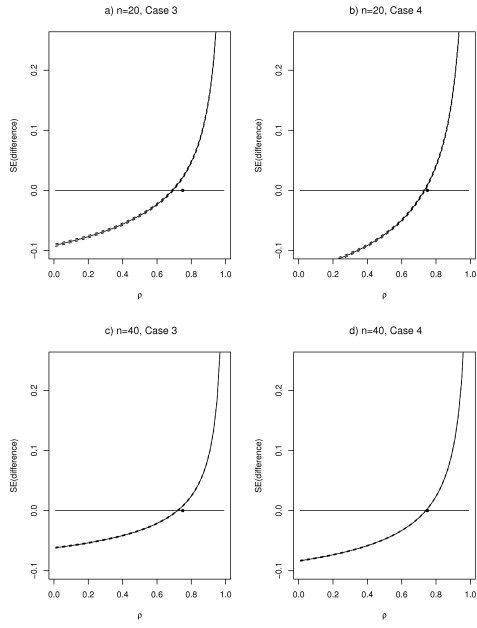
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

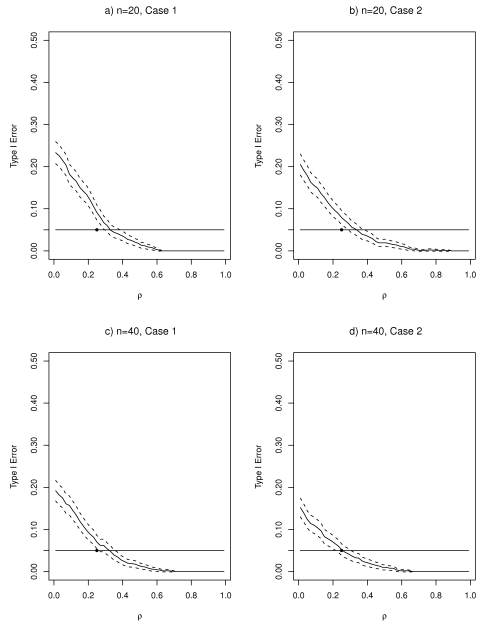


BIC : $\rho=0.75$

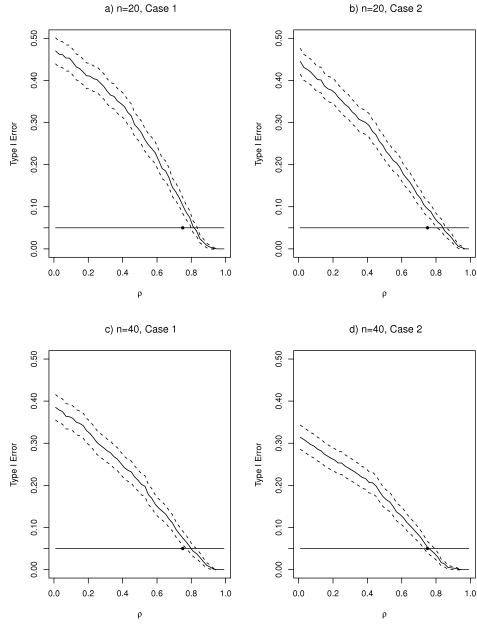
A.4 Type-I Error and Coverage Probability

A.4.1 No Parsimony Correction

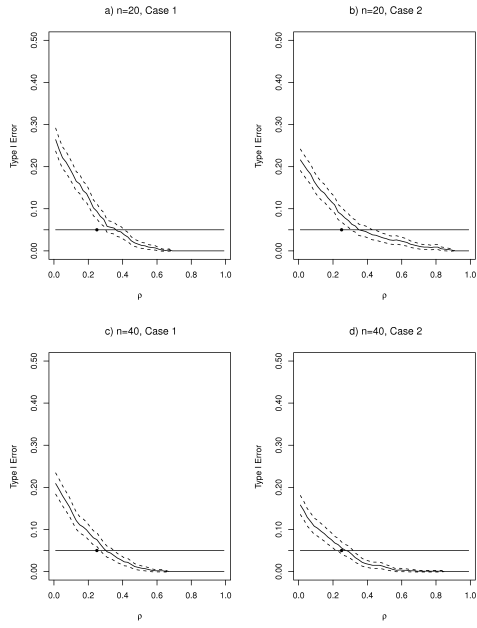
A.4.1.1 Cases 1 and 2



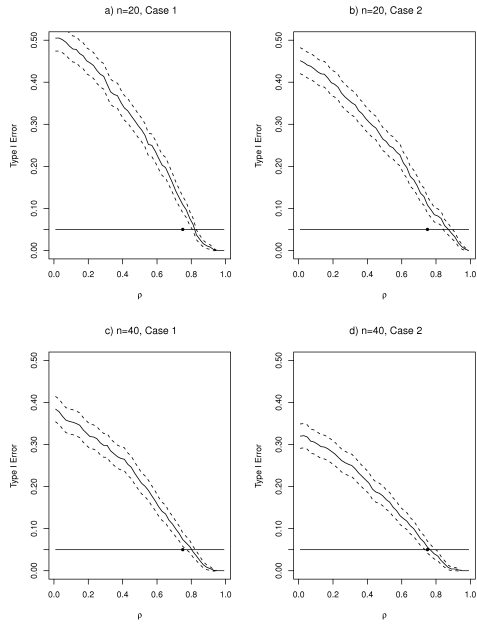
AIC : $\rho=0.25$



AIC : $\rho=0.75$

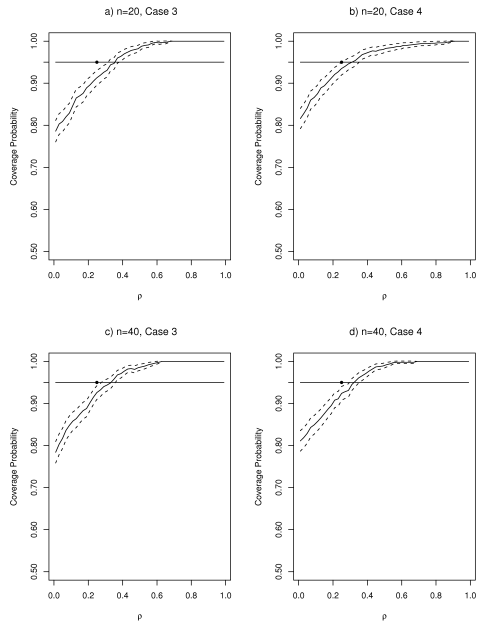


BIC : $\rho=0.25$

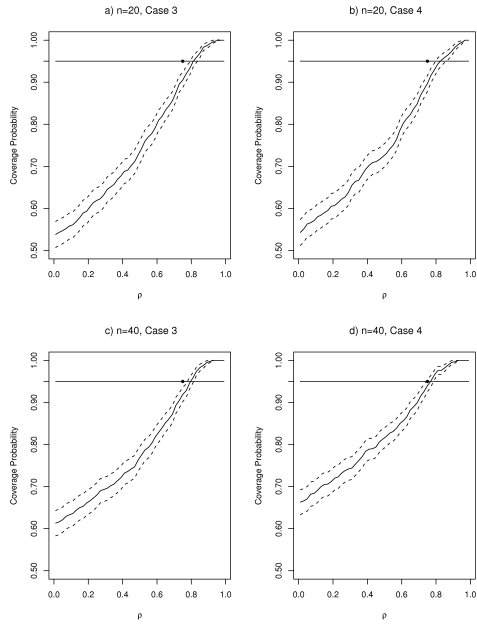


BIC : $\rho=0.75$

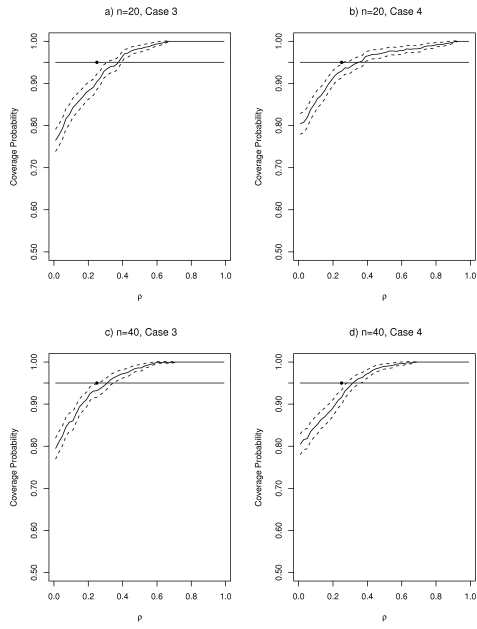
A.4.1.2 Cases 3 and 4



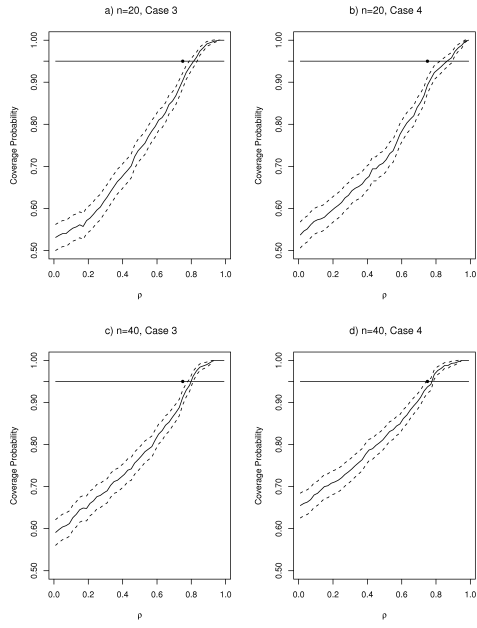
AIC : $\rho=0.25$



AIC : $\rho=0.75$



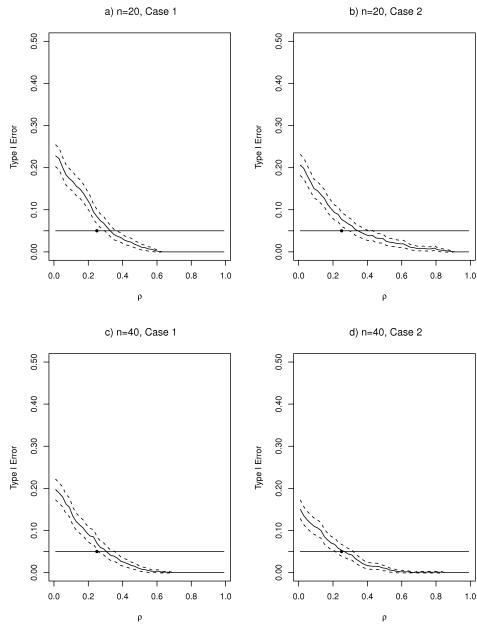
BIC : $\rho=0.25$



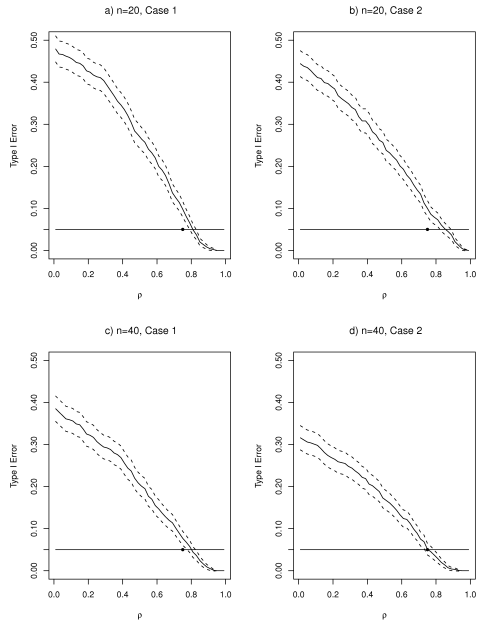
BIC : $\rho=0.75$

A.4.2 With Parsimony Correction

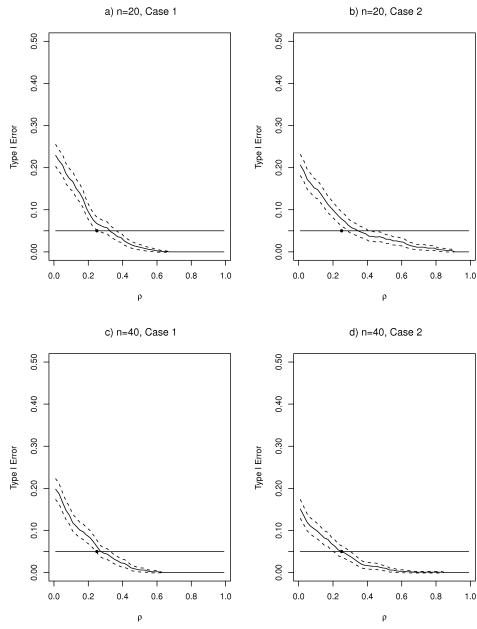
A.4.2.1 Cases 1 and 2



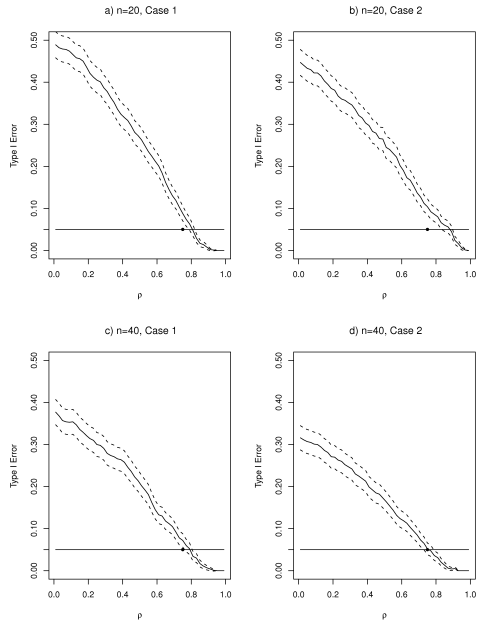
AIC : $\rho=0.25$



AIC : $\rho=0.75$

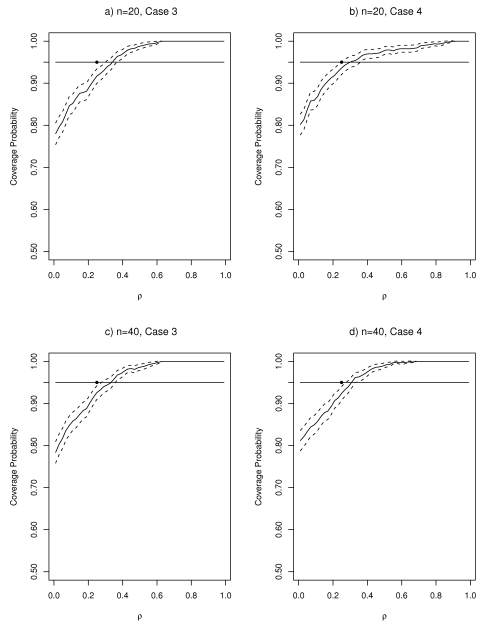


BIC : $\rho=0.25$

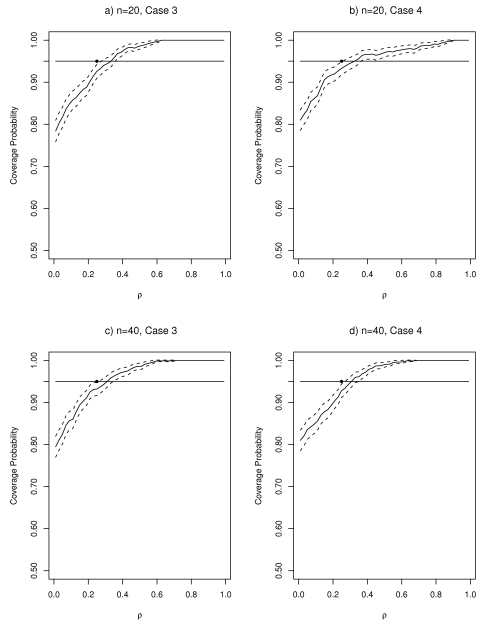


BIC : $\rho=0.75$

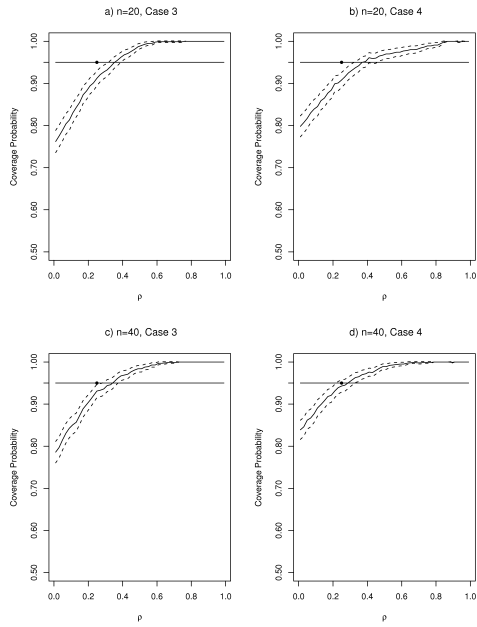
A.4.2.2 Cases 3 and 4



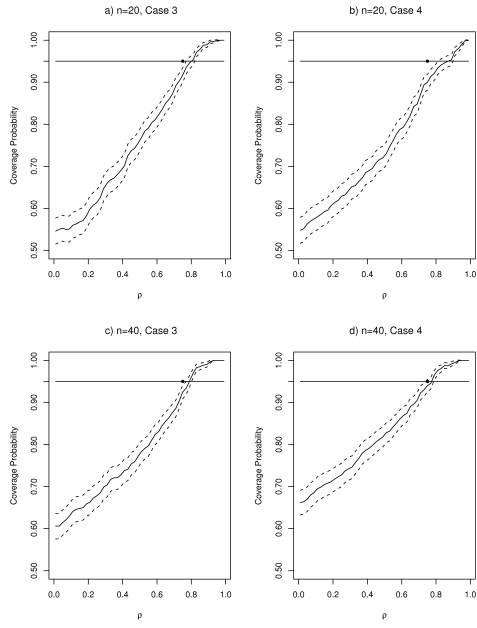
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$



BIC : $\rho=0.75$

APPENDIX B

Chapter 3 Proofs

Prior to detailing the results of Theorem 3.1 and Lemma 3.2, it is necessary to discuss the assumptions needed for the proof. After all partitioning, the statistical model can be written as

$$\mathbf{Y} = \mathbf{X}_1 \cdot \boldsymbol{\beta}_1 + \mathbf{X}_{2,M_o} \cdot \boldsymbol{\beta}_{2,M_o} + \mathbf{X}_{2,M_o^C} \cdot \boldsymbol{\beta}_{2,M_o^C} + \boldsymbol{\epsilon} = \mathbf{B}^* \cdot \boldsymbol{\eta} + \mathbf{W}^* \cdot \boldsymbol{\gamma} + \boldsymbol{\epsilon} \quad (\text{B.1})$$

with $\mathbf{B}^* = (\mathbf{X}_1, \mathbf{X}_{2,M_o})$ and $\mathbf{W}^* = \mathbf{X}_{2,M_o^C}$. This partitioning implies that $\boldsymbol{\beta} = (\boldsymbol{\eta}', \boldsymbol{\gamma}')'$. Recall that $\text{var}(\boldsymbol{\epsilon}) = \sigma^2 \cdot \mathbf{V}$ so that (B.1) can be written as:

$$\mathbf{Y}_t = \mathbf{B} \cdot \boldsymbol{\eta} + \mathbf{W} \cdot \boldsymbol{\gamma} + \mathbf{u}. \quad (\text{B.2})$$

where $\mathbf{V}^{-\frac{1}{2}} \cdot \mathbf{V}^{-\frac{1}{2}} = \mathbf{V}^{-1}$ so that $\mathbf{Y}_t = \mathbf{V}^{-\frac{1}{2}} \cdot \mathbf{Y}$, $\mathbf{W} = \mathbf{V}^{-\frac{1}{2}} \cdot \mathbf{W}^*$, $\mathbf{B} = \mathbf{V}^{-\frac{1}{2}} \cdot \mathbf{B}^*$ and $\mathbf{u} = \mathbf{V}^{-\frac{1}{2}} \cdot \boldsymbol{\epsilon}$. Since it is assumed that the elements of \mathbf{W}^* and \mathbf{W} are unrelated to the response vectors \mathbf{Y} and \mathbf{Y}_t , it can be assumed that $\boldsymbol{\gamma} = \mathbf{0}$. The error term in (B.2), \mathbf{u} , has variance $\sigma^2 \cdot \mathbf{I}$ so that ordinary least squares estimates can be used for all estimation.

The design matrices \mathbf{B} and \mathbf{W} are set up so \mathbf{W} and \mathbf{Y} are perpendicular to each other, which implies that $\mathbf{Y}_t' \cdot \mathbf{W} = \mathbf{0}$. The matrix \mathbf{B} contains all vectors not perpendicular to \mathbf{Y} . Thus, the projection of \mathbf{B} onto \mathbf{Y}_t is a non-zero vector ($\mathbf{B} \cdot$

$(\mathbf{B}' \cdot \mathbf{B})^{-1} \mathbf{B}' \cdot \mathbf{Y}_t \neq \mathbf{0}$). For any matrix \mathbf{A} , define $\mathbf{P}_\mathbf{A} = \mathbf{A} \cdot (\mathbf{A}' \cdot \mathbf{A})^{-1} \mathbf{A}'$ so that $\mathbf{P}_\mathbf{A}$ is the commonly defined projection matrix. This matrix is both symmetric and idempotent. From the previous definition of \mathbf{W} and \mathbf{B} , the following is true: $\mathbf{P}_\mathbf{B} \cdot \mathbf{W} = \mathbf{P}_\mathbf{W} \cdot \mathbf{B} = \mathbf{0}$.

The proof of Theorem 2.1.1 will use the aforementioned information directly. The proof of Theorems 2.1.2 and 2.1.3 stem from the transformation used in (B.2) and are simple extension of the proof found in *Shen et al.* (2004a). However, a detailed explanation of that proof is as follows.

B.1 Proof of Theorem 3.1 and Lemma 3.2

Even though the distribution of the error terms in (B.1) are not independent and identically distributed, the transformation used to get (B.2) results in independent and identically distributed error terms. Thus, the proof is for Theorems 2.1.2 and 2.1.3 have been discussed by *Shen et al.* (2004a). However, a discussion of this proof follows to ensure completeness and uniformity in notation.

Under the generalized least squares estimates, the estimate of the vector of fixed-effects is:

$$\begin{aligned} \hat{\beta} = \begin{pmatrix} \hat{\eta} \\ \hat{\gamma} \end{pmatrix} &= \left[\begin{pmatrix} \mathbf{B}' \\ \mathbf{W}' \end{pmatrix} \cdot \begin{pmatrix} \mathbf{B} & \mathbf{W} \end{pmatrix} \right]^{-1} \cdot \begin{pmatrix} \mathbf{B}' \cdot \mathbf{Y}_t \\ \mathbf{W}' \cdot \mathbf{Y}_t \end{pmatrix} \\ &= \begin{bmatrix} \mathbf{B}' \cdot \mathbf{B} & \mathbf{B}' \cdot \mathbf{W} \\ \mathbf{W}' \cdot \mathbf{B} & \mathbf{W}' \cdot \mathbf{W} \end{bmatrix}^{-1} \cdot \begin{pmatrix} \mathbf{B}' \cdot \mathbf{Y}_t \\ \mathbf{W}' \cdot \mathbf{Y}_t \end{pmatrix} \end{aligned} \quad (\text{B.3})$$

The inverse of the partitioned matrix in (B.3) can be solved using well-known methods found in any standard multivariate analysis or linear algebra text. Solving

this inverse yields:

$$\begin{aligned} & \begin{bmatrix} \mathbf{B}' \cdot \mathbf{B} & \mathbf{B}' \cdot \mathbf{W} \\ \mathbf{W}' \cdot \mathbf{B} & \mathbf{W}' \cdot \mathbf{W} \end{bmatrix}^{-1} = \\ & \begin{bmatrix} (\mathbf{B}' \cdot \mathbf{Q}_{\mathbf{W}} \cdot \mathbf{B})^{-1} & -(\mathbf{B}' \cdot \mathbf{Q}_{\mathbf{W}} \cdot \mathbf{B})^{-1} \cdot \mathbf{B}' \cdot \mathbf{W} \cdot (\mathbf{W}' \cdot \mathbf{W})^{-1} \\ -(\mathbf{W}' \cdot \mathbf{Q}_{\mathbf{B}} \cdot \mathbf{W})^{-1} \cdot \mathbf{W}' \cdot \mathbf{B} \cdot (\mathbf{B}' \cdot \mathbf{B})^{-1} & (\mathbf{W}' \cdot \mathbf{Q}_{\mathbf{B}} \cdot \mathbf{W})^{-1} \end{bmatrix} \end{aligned} \quad (\text{B.4})$$

Combining (B.2), using $\boldsymbol{\gamma}=\mathbf{0}$, (B.3) and (B.4) yields:

$$\begin{aligned} \begin{pmatrix} \hat{\boldsymbol{\eta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} &= \begin{bmatrix} (\mathbf{B}' \cdot \mathbf{Q}_{\mathbf{W}} \cdot \mathbf{B})^{-1} \cdot \mathbf{B}' \cdot \mathbf{Y} - (\mathbf{B}' \cdot \mathbf{Q}_{\mathbf{W}} \cdot \mathbf{B})^{-1} \cdot \mathbf{B}' \cdot \mathbf{W} \cdot (\mathbf{W}' \cdot \mathbf{W})^{-1} \cdot \mathbf{W} \cdot \mathbf{Y}_{\mathbf{t}} \\ (\mathbf{W}' \cdot \mathbf{Q}_{\mathbf{B}} \cdot \mathbf{W})^{-1} \cdot \mathbf{W}' \cdot \mathbf{Y} - (\mathbf{W}' \cdot \mathbf{Q}_{\mathbf{B}} \cdot \mathbf{W})^{-1} \cdot \mathbf{W}' \cdot \mathbf{B} \cdot (\mathbf{B}' \cdot \mathbf{B})^{-1} \cdot \mathbf{B}' \cdot \mathbf{Y}_{\mathbf{t}} \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{B}' \cdot \mathbf{Q}_{\mathbf{W}} \cdot \mathbf{B})^{-1} \cdot \mathbf{B}' \cdot \mathbf{Q}_{\mathbf{W}} \cdot \mathbf{Y} \\ (\mathbf{W}' \cdot \mathbf{Q}_{\mathbf{B}} \cdot \mathbf{W})^{-1} \cdot \mathbf{W}' \cdot \mathbf{Q}_{\mathbf{B}} \cdot \mathbf{Y} \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{B}' \cdot \mathbf{Q}_{\mathbf{W}} \cdot \mathbf{B})^{-1} \cdot \mathbf{B}' \cdot \mathbf{Q}_{\mathbf{W}} \cdot (\mathbf{B} \cdot \boldsymbol{\eta} + \mathbf{u}) \\ (\mathbf{W}' \cdot \mathbf{Q}_{\mathbf{B}} \cdot \mathbf{W})^{-1} \cdot \mathbf{W}' \cdot \mathbf{Q}_{\mathbf{B}} \cdot (\mathbf{B} \cdot \boldsymbol{\eta} + \mathbf{u}) \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{\eta} + (\mathbf{B}' \cdot \mathbf{Q}_{\mathbf{W}} \cdot \mathbf{B})^{-1} \cdot \mathbf{B}' \cdot \mathbf{Q}_{\mathbf{W}} \cdot \mathbf{u} \\ (\mathbf{W}' \cdot \mathbf{Q}_{\mathbf{B}} \cdot \mathbf{W})^{-1} \cdot \mathbf{W}' \cdot \mathbf{Q}_{\mathbf{B}} \cdot \mathbf{u} \end{bmatrix} \end{aligned} \quad (\text{B.5})$$

The elimination of a term in $\hat{\boldsymbol{\gamma}}$ is due to the identity $\mathbf{Q}_{\mathbf{B}} \cdot \mathbf{B} = (\mathbf{I} - \mathbf{P}_{\mathbf{B}}) \cdot \mathbf{B} = \mathbf{0}$.

After a model selection procedure is used, (B.5) can be written as:

$$\hat{\boldsymbol{\beta}}_{\hat{M}} = \begin{pmatrix} \hat{\boldsymbol{\eta}}_{\hat{M}} \\ \hat{\boldsymbol{\gamma}}_{\hat{M}} \end{pmatrix} = \begin{bmatrix} \boldsymbol{\eta} + (\mathbf{B}' \cdot \mathbf{Q}_{\mathbf{W}_{\hat{M}}} \cdot \mathbf{B})^{-1} \cdot \mathbf{B}' \cdot \mathbf{Q}_{\mathbf{W}_{\hat{M}}} \cdot \mathbf{u} \\ (\mathbf{W}_{\hat{M}}' \cdot \mathbf{Q}_{\mathbf{B}} \cdot \mathbf{W}_{\hat{M}})^{-1} \cdot \mathbf{W}_{\hat{M}}' \cdot \mathbf{Q}_{\mathbf{B}} \cdot \mathbf{u} \end{bmatrix} \quad (\text{B.6})$$

where $\mathbf{W}_{\hat{M}}$ contains only columns of \mathbf{W} that were recommended by the model selection procedure. As the sample size n gets large, the consistent or overconsistent

property of the model selection procedure will mean that $\mathbf{W}_{\hat{M}} \rightarrow \mathbf{0}$, resulting in $\mathbf{Q}_{W_{\hat{M}}} = (\mathbf{I} - \mathbf{P}_{W_{\hat{M}}}) \rightarrow \mathbf{I}$. Therefore:

$$\lim_{n \rightarrow \infty} \hat{\boldsymbol{\beta}}_{\hat{M}} = \lim_{n \rightarrow \infty} \begin{pmatrix} \hat{\boldsymbol{\eta}}_{\hat{M}} \\ \hat{\boldsymbol{\gamma}}_{\hat{M}} \end{pmatrix} = \begin{bmatrix} \boldsymbol{\eta} + (\mathbf{B}' \cdot \mathbf{B})^{-1} \cdot \mathbf{B}' \cdot \mathbf{u} \\ \mathbf{0} \end{bmatrix} \quad (\text{B.7})$$

Since \mathbf{u} is null valued, the expectation of of (B.7) yields $E(\lim_{n \rightarrow \infty} \hat{\boldsymbol{\beta}}_{\hat{M}}) = (\boldsymbol{\eta}', \mathbf{0}')'$. Lastly, $\boldsymbol{\beta}_{2, M^c} = \mathbf{0}$ implies that $\mathbf{W} = \mathbf{0}$, which, in addition to (B.6), completes the proof of Theorem 3.1.

This proof does not use the identity $\mathbf{P}_B \cdot \mathbf{W} = \mathbf{P}_W \cdot \mathbf{B} = \mathbf{0}$. Invoking this identity in (B.5) permits the conclusion that $E(\hat{\boldsymbol{\mu}}_{\hat{M}}) = \boldsymbol{\beta}$. However, to remain general against the paradigm that model selection is performed, the conclusion of Theorem 3.1 is as stated. The asymptotic convergence of the efficient and consistent model selection properties were explored by *Shao* (1997) and considered to be in probability for both consistent and efficient model selection; almost sure convergence is achieved for consistent model selection. Thus, the fixed-effect estimates will have this effect as well.

From (B.7), it is apparent that asymptotically $\hat{\boldsymbol{\beta}}_{\hat{M}}$ is a linear function of \mathbf{u} , which has a normal distribution with mean $\mathbf{0}$ and variance $\sigma^2 \cdot \mathbf{I}$. Since $\hat{\boldsymbol{\beta}}_{\hat{M}}$ is a linear function of a normally distributed random variable, it will have a normal distribution, thus completing the proof of Lemma 3.2.

B.2 Proof of Theorem 3.3

In the framework of model selection using information criteria, models are selected that minimize a function that is proportional to $(\mathbf{Y}_t - \hat{\mathbf{Y}}'_{M,t}) \cdot (\mathbf{Y}_t - \hat{\mathbf{Y}}'_{M,t}) + c \cdot p_M$ where $\hat{\mathbf{Y}}_{M,t}$ is the ordinary least squares predicted values which can be denoted by

$\mathbf{P}_{\mathbf{B}, \mathbf{W}_M} \cdot \mathbf{Y}_t$. Equivalently, this criterion can be denoted by:

$$\mathbf{u}' \cdot (\mathbf{I} - \mathbf{P}_{\mathbf{B}, \mathbf{W}_M}) \cdot \mathbf{u} + c \cdot p_M. \quad (\text{B.8})$$

Since the criterion of selecting a particular model is based on fixed characteristics of the model $(\mathbf{P}_{\mathbf{B}, \mathbf{W}_M}, p_M)$, a constant c and a random vector \mathbf{u} , the estimated model can be viewed probabilistically. A dichotomous variable $I(\hat{M}=M)$ can be used to indicate whether model M is the selected model. This variable can be thought of a Bernoulli variable with probability $P(\hat{M}=M)$.

For model M , the ordinary least squares estimate of σ^2 is $\hat{\sigma}_M^2 = \frac{SSE_M}{N-p_M}$ where $SSE_M = SSE_{\mathbf{B}, \mathbf{W}_M}$ is the error sums of squares defined by operand of the logarithm function in (B.8), p_M is the total number of columns of \mathbf{B} and \mathbf{W}_M (assuming both \mathbf{B} and \mathbf{W}_M are full rank), and N is the total number of observations. If the true model is known, then the expected value of the variance estimate is $E(\hat{\sigma}_{M_o}^2) = \frac{SSE_B}{N-p_{M_B}}$. By the definition of expected value,

$$E(\hat{\sigma}_M^2) = \sum_{M: M \subset M_o} \hat{\sigma}_M^2 \cdot P(\hat{M} = M) = \sum_{M: M \subset M_o} \frac{SSE_M}{N - p_M} \cdot P(\hat{M} = M) \quad (\text{B.9})$$

From the decomposition of the sums of squares, it is known that $SSE_M = SSTO - SSR_M$ where $SSTO$ is the total sums of squares and $SSR_M = SSR_{\mathbf{B}, \mathbf{W}_M}$ is the sum of squares regression defined as $\mathbf{Y}_t' \cdot \mathbf{P}_{\mathbf{B}, \mathbf{W}_M} \cdot \mathbf{Y}_t$. Note that $SSTO$ is constant with respect to different models and only a function of \mathbf{Y}_t . The decomposition of (B.9) yields

$$\begin{aligned} E(\hat{\sigma}_M^2) &= \sum_{M: M \subset M_o} (SSTO - SSR_M) \cdot \frac{P(\hat{M} = M)}{N - p_M} \\ &= \sum_{M: M \subset M_o} (SSTO - SSR_{\mathbf{B}} - SSR_{\mathbf{W}_M}) \cdot \frac{P(\hat{M} = M)}{N - p_M} \end{aligned} \quad (\text{B.10})$$

The decomposition of SSR_M is possible because, by definition, \mathbf{P}_B and \mathbf{P}_{W_M} are orthogonal. (B.10) can be rewritten as:

$$\begin{aligned} E(\hat{\sigma}_{\hat{M}}^2) &= \sum_{M: M \subset M_o} (SSTO - SSR_B) \cdot \frac{P(\hat{M} = M)}{N - p_M} - SSR_{W_M} \cdot \frac{P(\hat{M} = M)}{N - p_M} \\ &\geq \sum_{M: M \subset M_o} (SSTO - SSR_B) \cdot \frac{P(\hat{M} = M)}{N - p_M} \end{aligned} \quad (\text{B.11})$$

If the right-hand side of (B.11) is multiplied by $\frac{N - p_{M_B}}{N - p_M}$, it is easy to see that

$$E(\hat{\sigma}_{\hat{M}}^2) \geq E(\hat{\sigma}_{M_o}^2) \cdot \frac{N - p_{M_B}}{N - p_M}$$

and since $\frac{N - p_{M_B}}{N - p_M} \leq 1$,

$$E(\hat{\sigma}_{\hat{M}}^2) \geq E(\hat{\sigma}_{M_o}^2) \quad (\text{B.12})$$

Furthermore, equality in (B.12) can be shown if $\mathbf{W}=\mathbf{0}$, thus completing the proof of Theorem 3.3.

B.3 Proof of Theorem 3.5

Definition 3.4 states that the GDF of a model is $\frac{\partial}{\partial y_{ij}} E(\hat{\mu}_{ij})$. Under the generalized least squares estimators, the predicted values are linear functions of the response variable \mathbf{Y} . The definition of the expected value permits writing the generalized leverage as:

$$h_{ij}^M = \frac{\partial}{\partial \mu_{ij}} E[\hat{\mu}_{ij}(\mathbf{Y})] = \frac{\partial}{\partial \mu_{ij}} \int \hat{\mu}_{ij}(\mathbf{Y}) \cdot p(\mathbf{Y}_i | \mu_i, \mathbf{V}_i) d\mathbf{Y}_i \quad (\text{B.13})$$

where $p(\mathbf{Y}_i | \mu_i, \sigma^2 \cdot \mathbf{V}_i)$ is the distribution of \mathbf{Y}_i , which, under the LMM, is normally distributed with mean $\boldsymbol{\mu}$ and variance $\sigma^2 \cdot \mathbf{V}_i$. If the order of the integration and differentiation are transposed, then:

$$\begin{aligned}
h_{ij}^M &= \int \frac{\partial}{\partial \mu_{ij}} \hat{\mu}_{ij}(\mathbf{Y}) \cdot p(\mathbf{Y}_i | \mu_i, \mathbf{V}_i) d\mathbf{Y}_i \\
&= \int \hat{\mu}_{ij}(\mathbf{Y}) \cdot \frac{\partial}{\partial \mu_{ij}} p(\mathbf{Y}_i | \mu_i, \mathbf{V}_i) d\mathbf{Y}_i \\
&= \int \hat{\mu}_{ij}(\mathbf{Y}) \cdot (2\pi)^{\frac{n_i}{2}} \cdot |\mathbf{V}_i|^{-\frac{1}{2}} \cdot \frac{\partial}{\partial y_{ij}} \exp \left(-\frac{1}{2 \cdot \sigma^2} (\mathbf{Y}_i - \boldsymbol{\mu}_i)' \cdot \mathbf{V}_i^{-1} \cdot (\mathbf{Y}_i - \boldsymbol{\mu}_i) \right) d\mathbf{Y}_i \\
&= \int \hat{\mu}_{ij}(\mathbf{Y}) \cdot (2\pi)^{\frac{n_i}{2}} \cdot |\mathbf{V}_i|^{\frac{1}{2}} \cdot \exp \left(-\frac{1}{2 \cdot \sigma^2} (\mathbf{Y}_i - \boldsymbol{\mu}_i)' \cdot \mathbf{V}_i^{-1} \cdot (\mathbf{Y}_i - \boldsymbol{\mu}_i) \right) \cdot -\frac{1}{2 \cdot \sigma^2} \\
&\quad [(\mathbf{Y}_i - \boldsymbol{\mu}_i)' \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j + \mathbf{e}_j' \cdot \mathbf{V}_i^{-1} \cdot (\mathbf{Y}_i - \boldsymbol{\mu}_i)] \cdot (-1) d\mathbf{Y}_i
\end{aligned} \tag{B.14}$$

Combining terms and using the definition of the expected value leads to the first part of Theorem 3.5:

$$h_{ij}^M = \frac{1}{\sigma^2} \cdot \mathbb{E} [\hat{\mu}_{ij}(\mathbf{Y}) (\mathbf{Y}_i - \boldsymbol{\mu}_i)' \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j]$$

To show the second part of the Theorem, consider the definition of the covariance operator is used. For clarity, the notation denoting the dependency of the response vector on the predicted values is dropped.

$$\begin{aligned}
\frac{1}{\sigma^2} \cdot \text{cov}(\mu_{ij}, (\mathbf{Y}_i - \boldsymbol{\mu}_i)' \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j) &= \frac{1}{\sigma^2} \text{cov}(\mathbf{e}_j \cdot \hat{\boldsymbol{\mu}}_i, \mathbf{Y}' \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j) \\
&= \frac{1}{\sigma^2} \mathbf{e}_j' \cdot \text{cov}(\hat{\boldsymbol{\mu}}_i, \mathbf{Y}) \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j \\
&= \frac{1}{\sigma^2} \mathbf{e}_j' \cdot \mathbb{E} [(\hat{\boldsymbol{\mu}}_i - \mathbb{E}(\hat{\boldsymbol{\mu}}_i)) \cdot (\mathbf{Y}_i - \mathbb{E}(\mathbf{Y}_i)')] \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j
\end{aligned}$$

Expanding the operands of the expectation yields:

$$\frac{1}{\sigma^2} \mathbf{e}_j' \cdot \mathbb{E} [(\hat{\boldsymbol{\mu}}_i \cdot \mathbf{Y}_i' - \mathbb{E}(\hat{\boldsymbol{\mu}}_i) \cdot \mathbf{Y}_i' - \hat{\boldsymbol{\mu}}_i \cdot \mathbb{E}(\mathbf{Y}_i') - \mathbb{E}(\hat{\boldsymbol{\mu}}_i) \cdot \mathbb{E}(\mathbf{Y}_i')) \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j]$$

$$\frac{1}{\sigma^2} \mathbf{e}_j' \cdot \mathbb{E} [(\hat{\boldsymbol{\mu}}_i \cdot \mathbf{Y}_i' - \hat{\boldsymbol{\mu}}_i \cdot \mathbb{E}(\mathbf{Y}_i')) \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j + \mathbb{E}(\hat{\boldsymbol{\mu}}_i) \cdot \mathbb{E}(\mathbf{Y}_i') - \mathbb{E}(\hat{\boldsymbol{\mu}}_i) \cdot \mathbf{Y}_i'] \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j$$

$$\frac{1}{\sigma^2} \mathbf{e}_j' \cdot \mathbb{E} [(\hat{\boldsymbol{\mu}}_i \cdot \mathbf{Y}_i' - \hat{\boldsymbol{\mu}}_i \cdot \boldsymbol{\mu}_i') \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j + \mathbf{0}] = \mathbb{E} [\hat{\mu}_{ij}(\mathbf{Y})(\mathbf{Y}_i - \boldsymbol{\mu}_i)' \cdot \mathbf{V}_i^{-1} \cdot \mathbf{e}_j]$$

This completes the proof of Theorem 3.5.

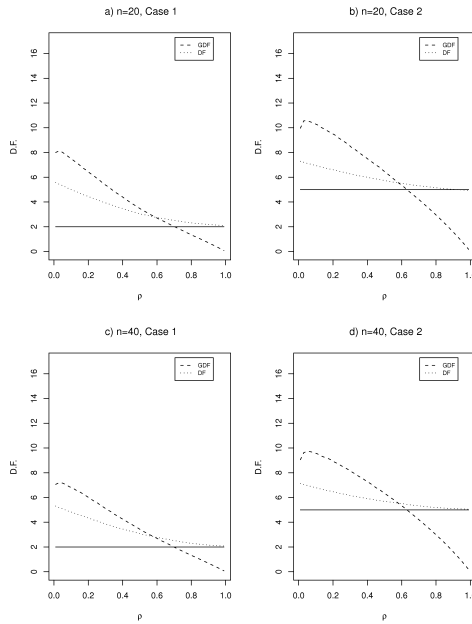
APPENDIX C

Extended GDF Simulation Results

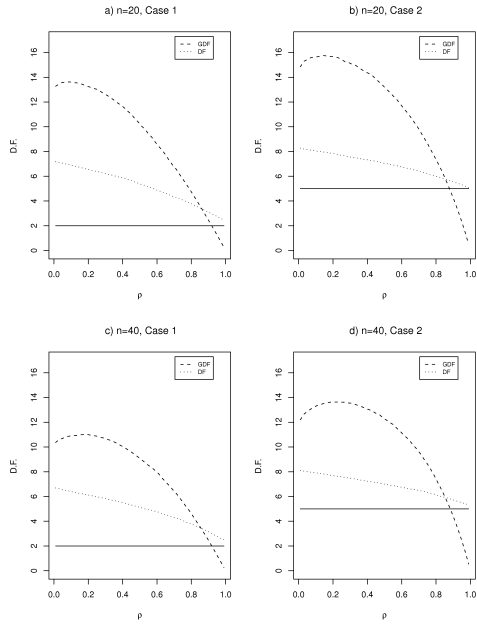
C.1 Generalized Degrees of Freedom-Proof-of-Concept

C.1.1 No Parsimony Correction

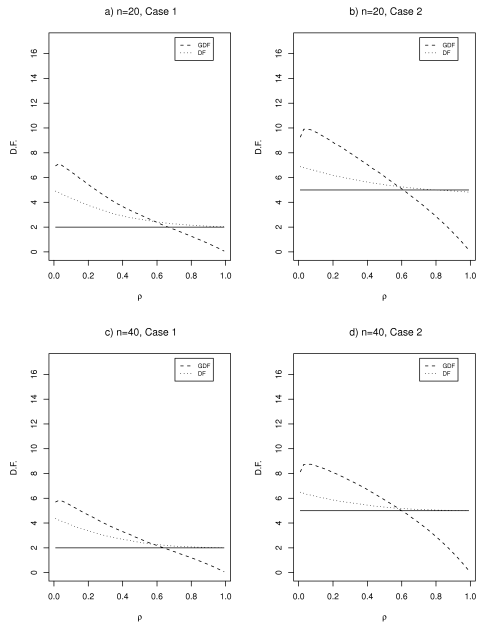
C.1.1.1 Cases 1 and 2



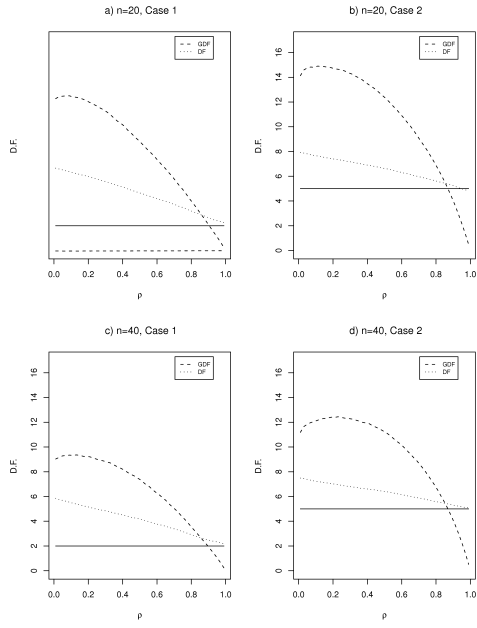
AIC : $\rho=0.25$



AIC : $\rho=0.75$

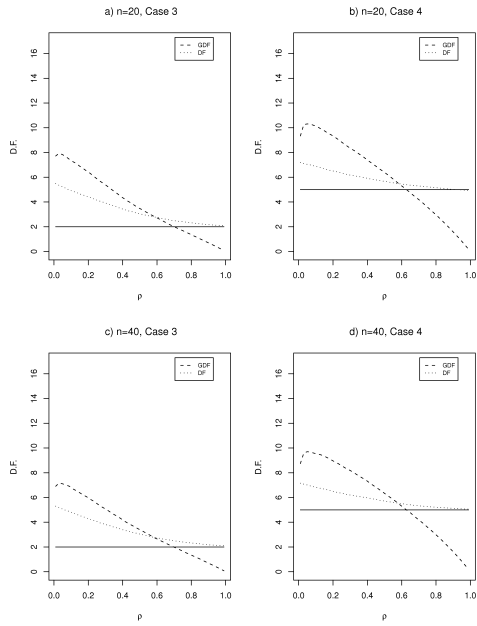


BIC : $\rho=0.25$

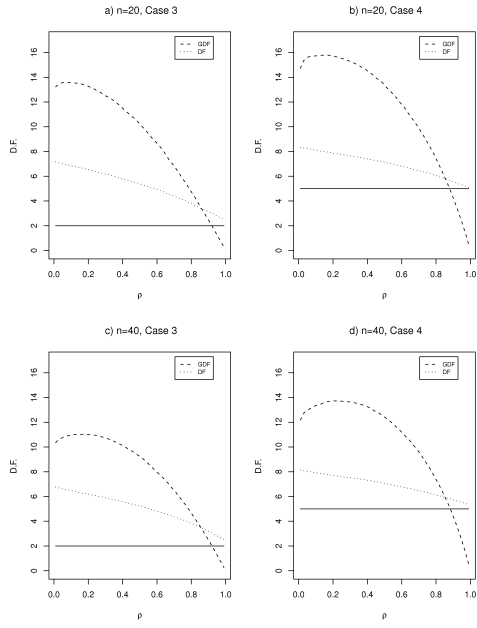


BIC : $\rho=0.75$

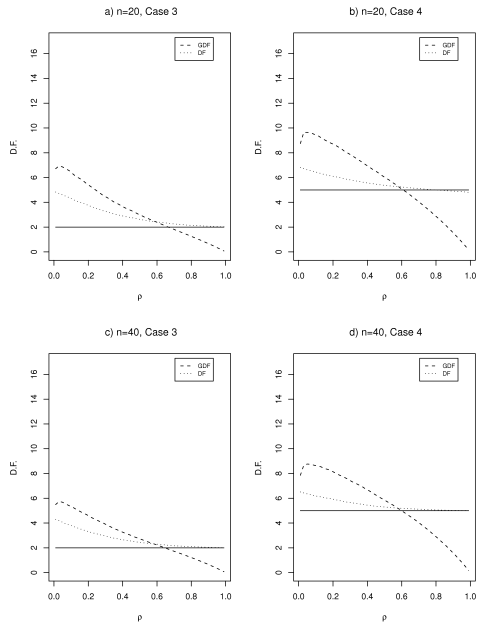
C.1.1.2 Cases 3 and 4



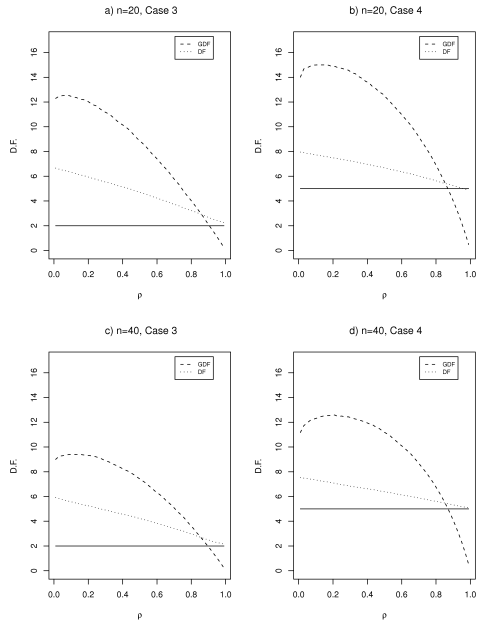
AIC : $\rho=0.25$



AIC : $\rho=0.75$



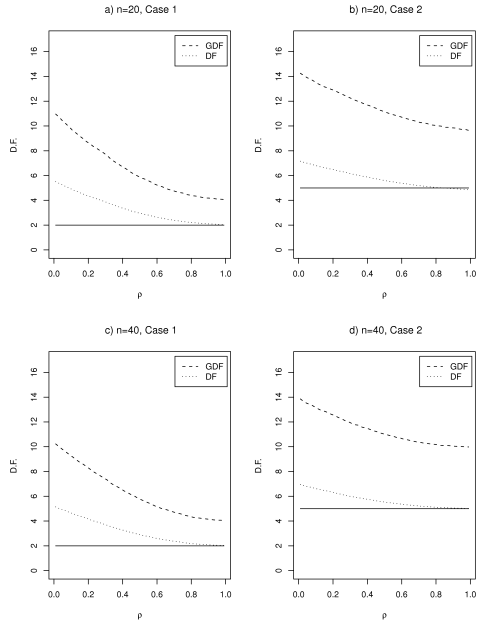
BIC : $\rho=0.25$



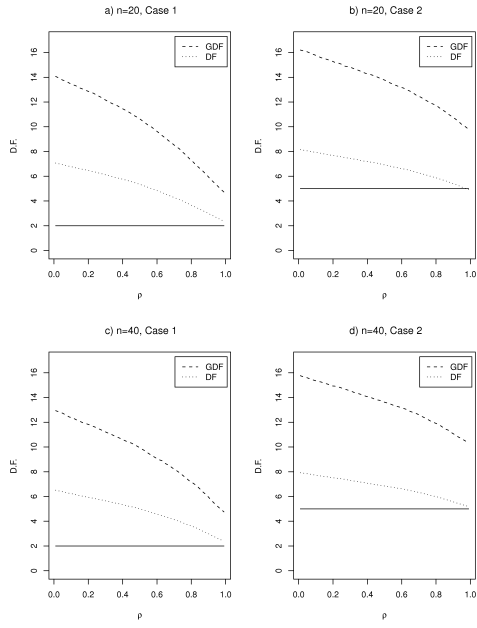
BIC : $\rho=0.75$

C.1.2 With Parsimony Correction

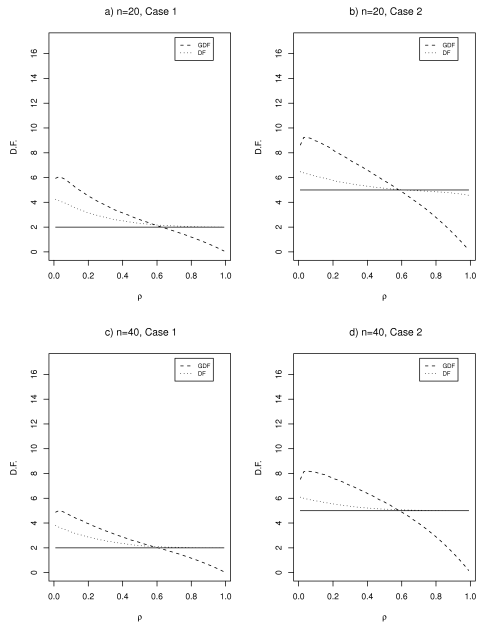
C.1.2.1 Cases 1 and 2



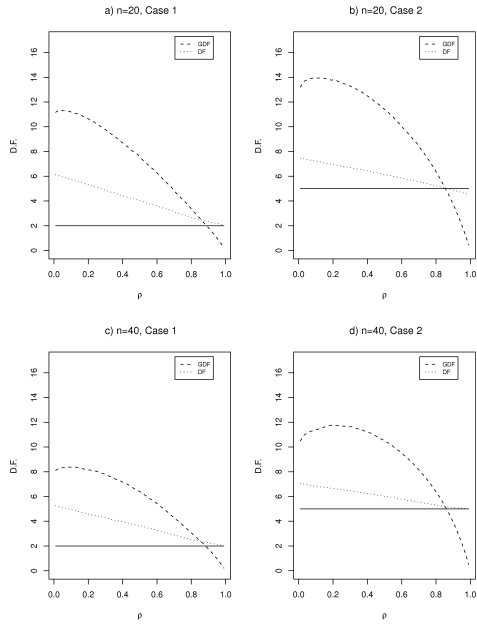
AIC : $\rho=0.25$



AIC : $\rho=0.75$

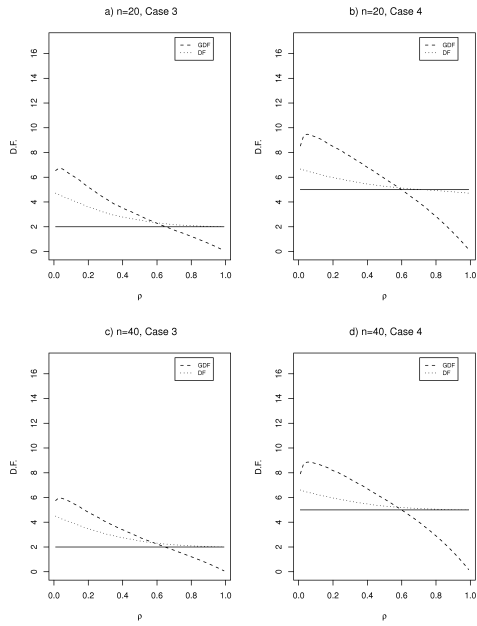


BIC : $\rho=0.25$

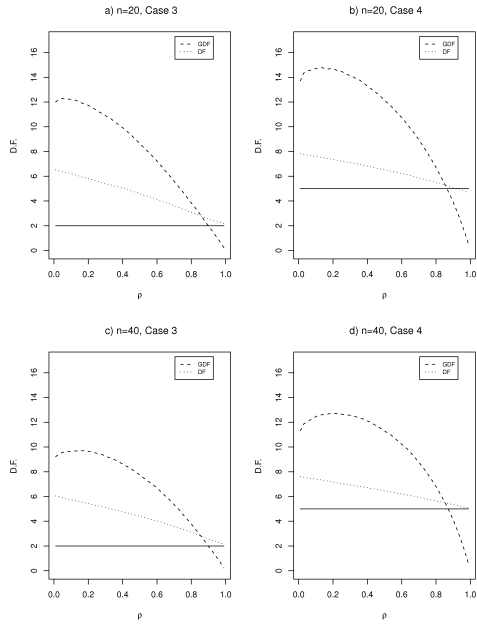


BIC : $\rho=0.75$

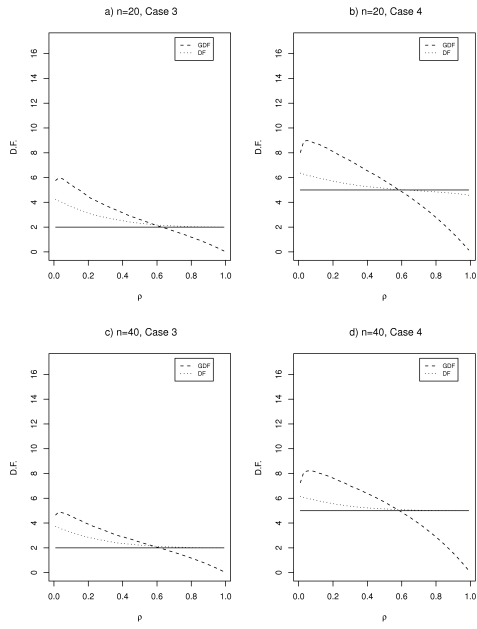
C.1.2.2 Cases 3 and 4



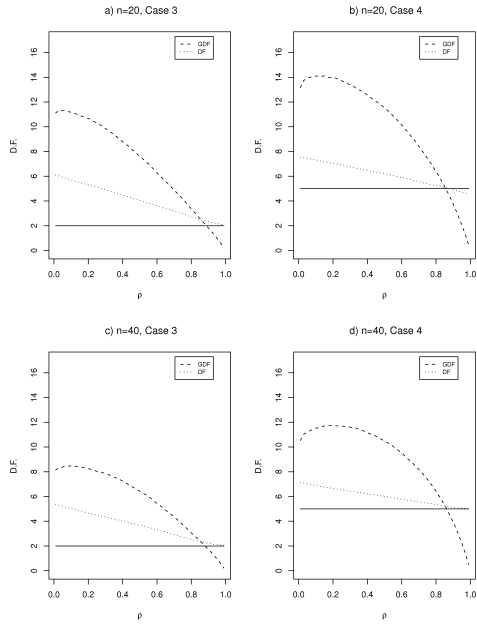
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

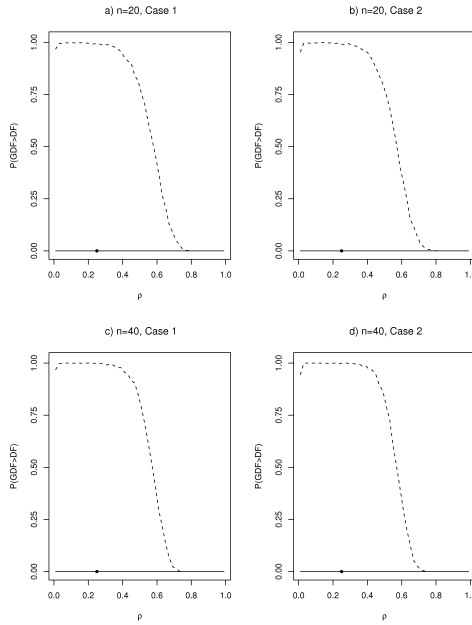


BIC : $\rho=0.75$

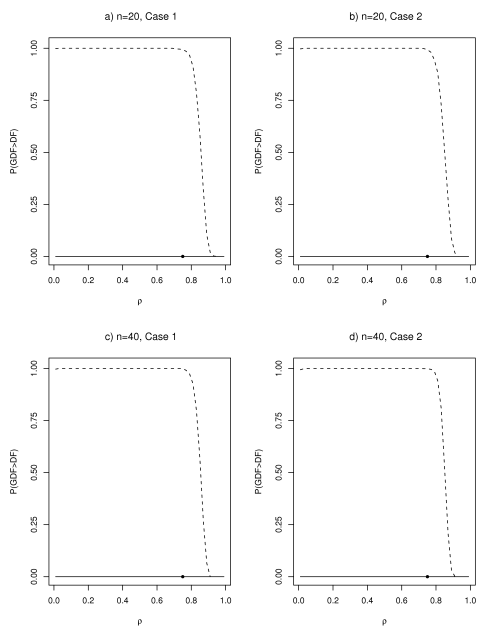
C.2 Relationship between GDF and Model DF – Proof-of-Concept

C.2.1 No parsimony correction

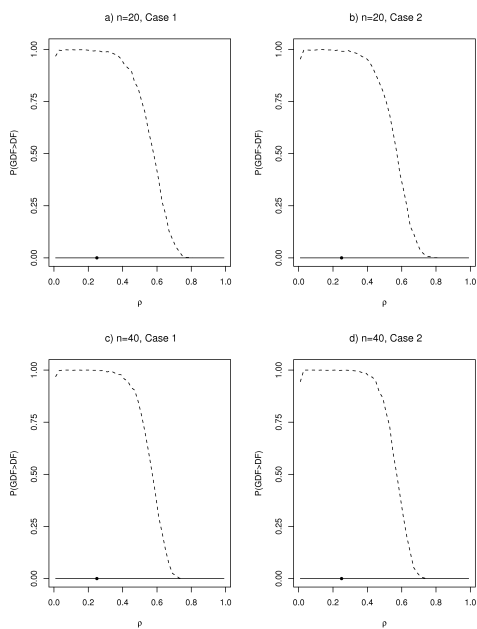
C.2.1.1 Cases 1 and 2



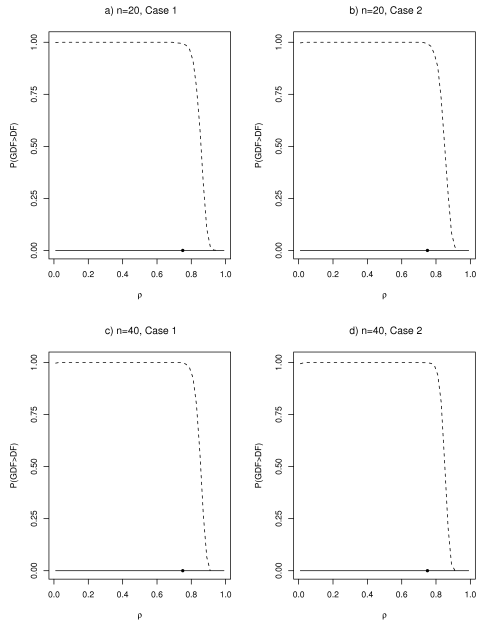
AIC : $\rho=0.25$



AIC : $\rho=0.75$

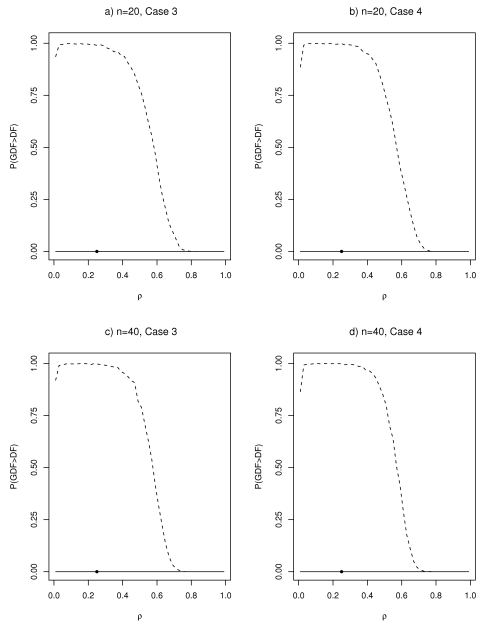


BIC : $\rho=0.25$

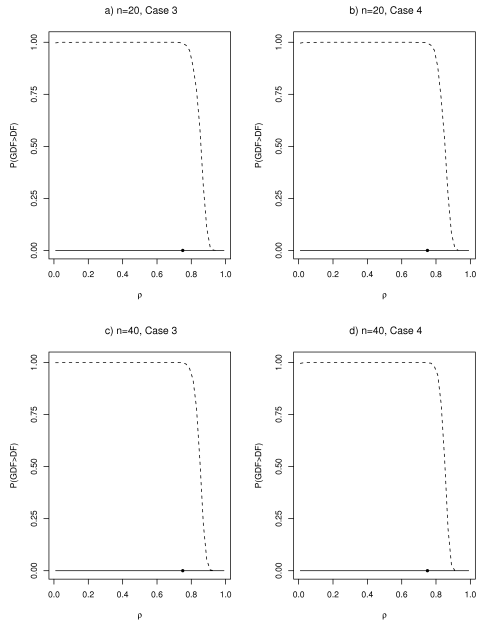


BIC : $\rho=0.75$

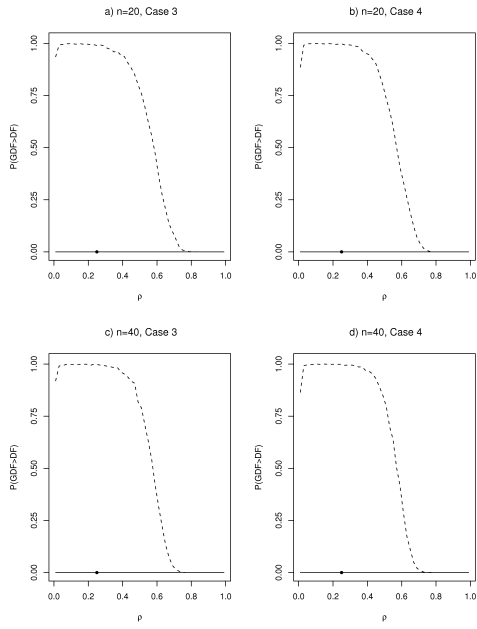
C.2.1.2 Cases 3 and 4



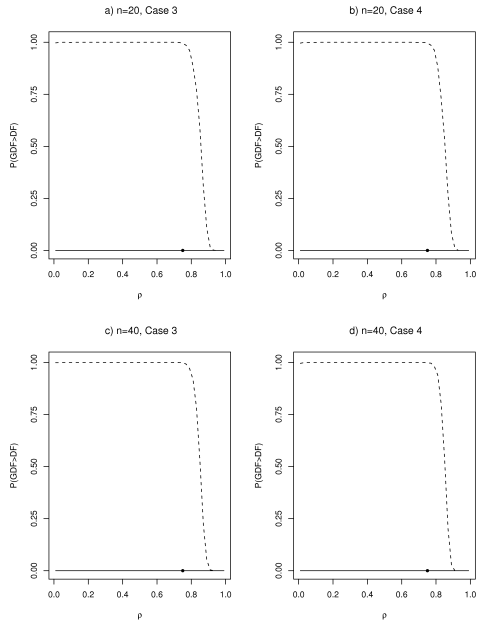
AIC : $\rho=0.25$



AIC : $\rho=0.75$



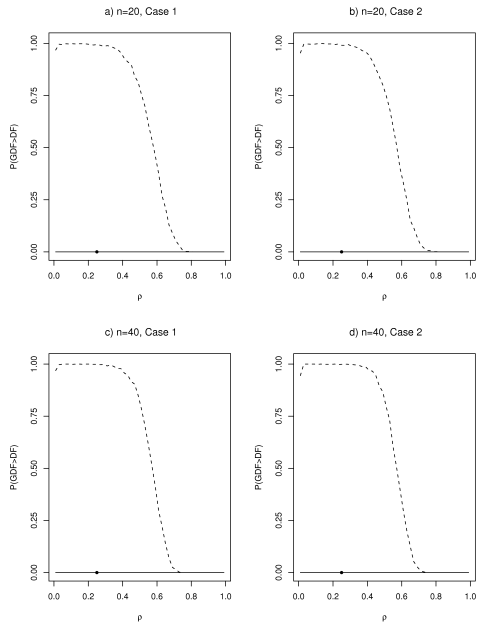
BIC : $\rho=0.25$



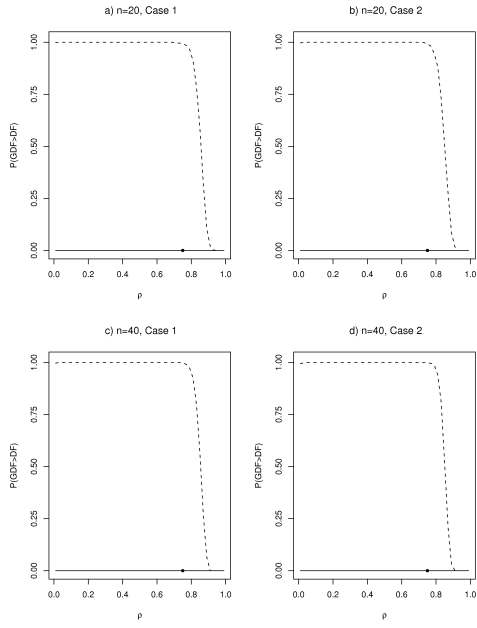
BIC : $\rho=0.75$

C.2.2 With Parsimony Correction

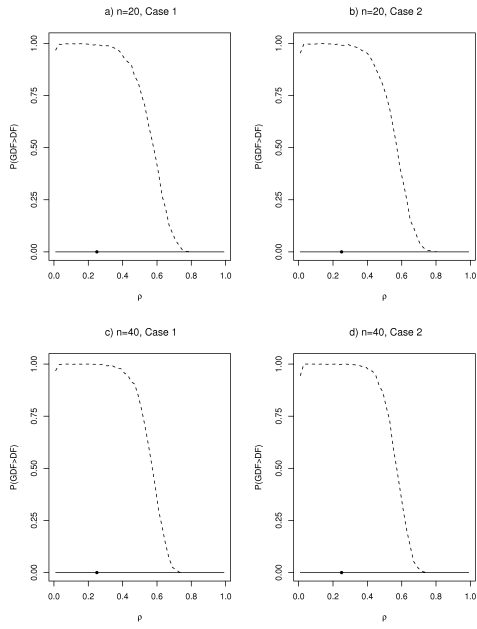
C.2.2.1 Cases 1 and 2



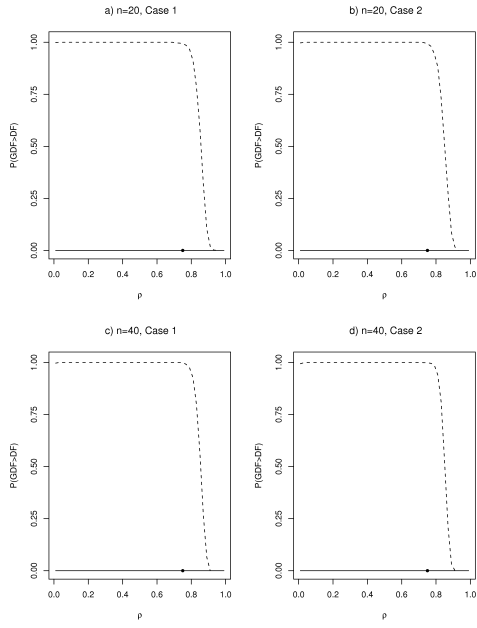
AIC : $\rho=0.25$



AIC : $\rho=0.75$

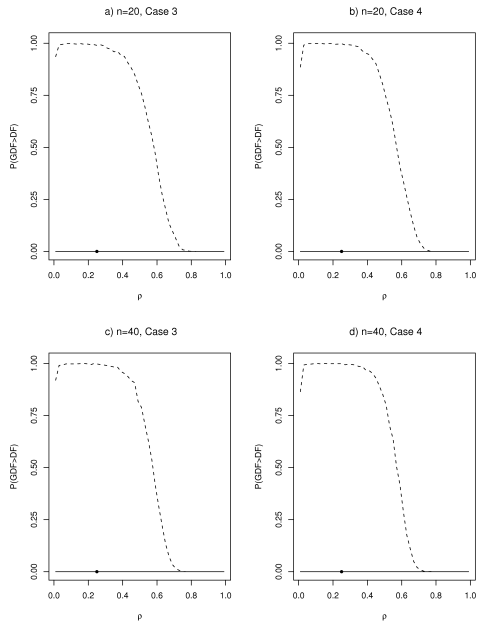


BIC : $\rho=0.25$

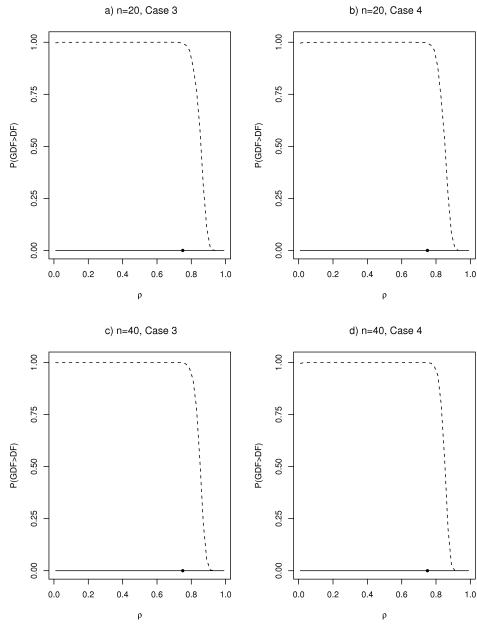


BIC : $\rho=0.75$

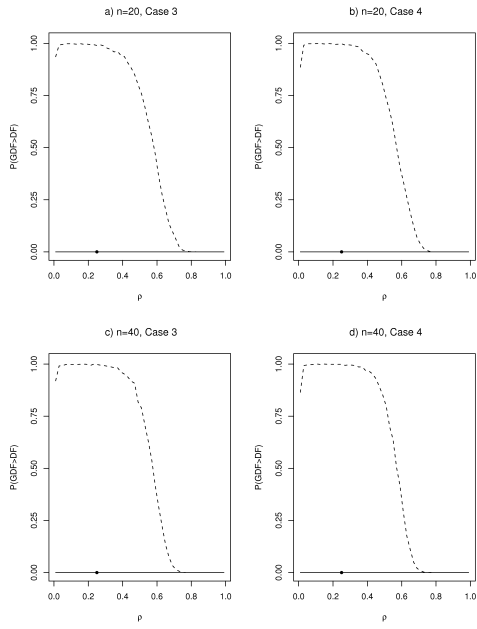
C.2.2.2 Cases 3 and 4



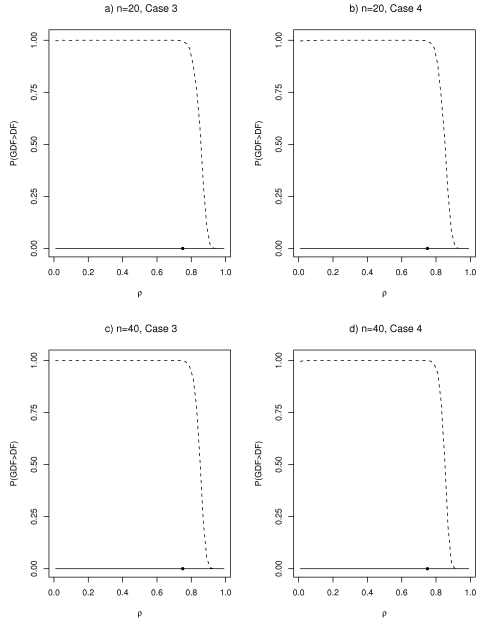
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

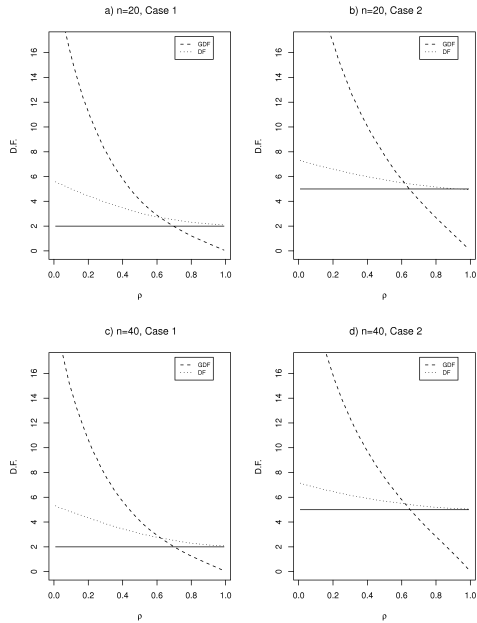


BIC : $\rho=0.75$

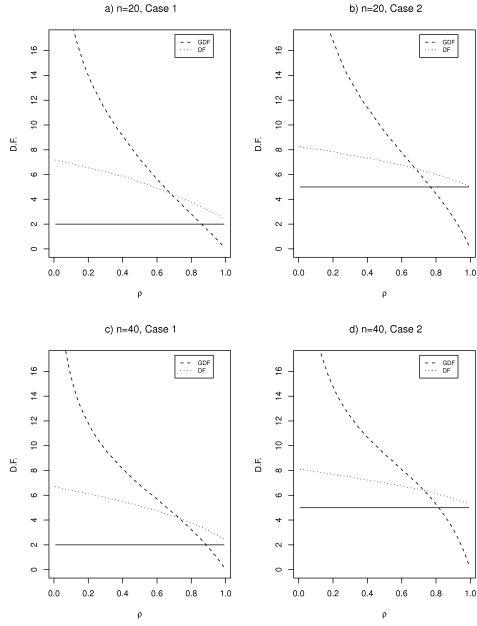
C.3 Generalized Degrees of Freedom-Data-Driven Method

C.3.1 No Parsimony Correction

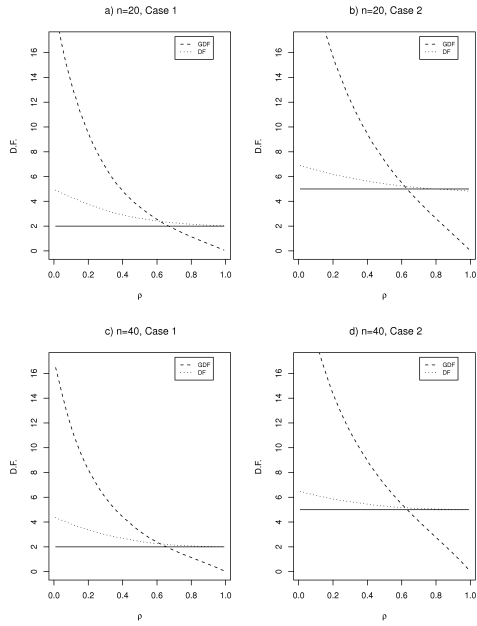
C.3.1.1 Cases 1 and 2



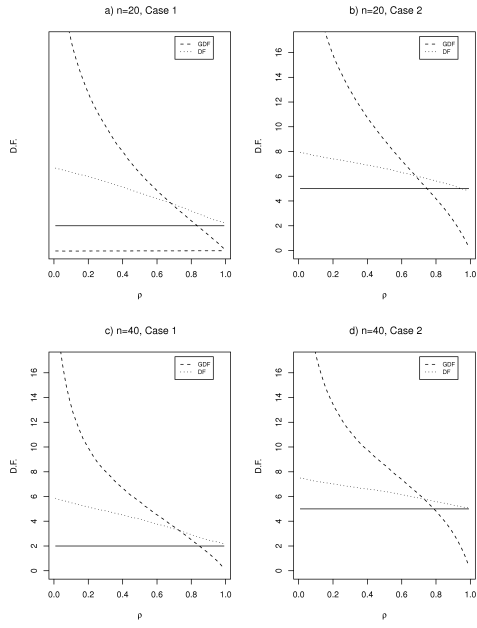
AIC : $\rho=0.25$



AIC : $\rho=0.75$

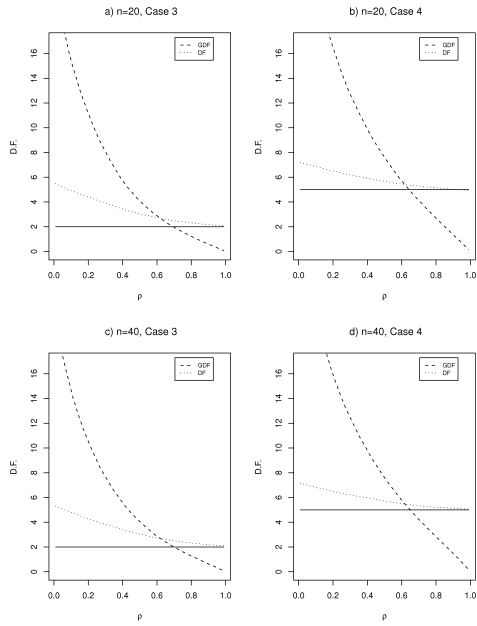


BIC : $\rho=0.25$

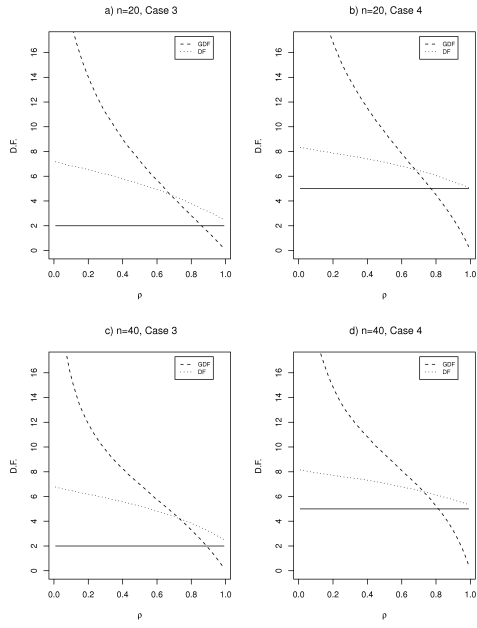


BIC : $\rho=0.75$

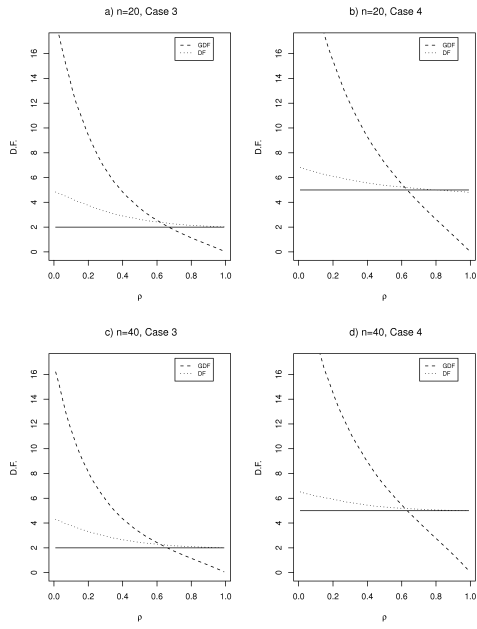
C.3.1.2 Cases 3 and 4



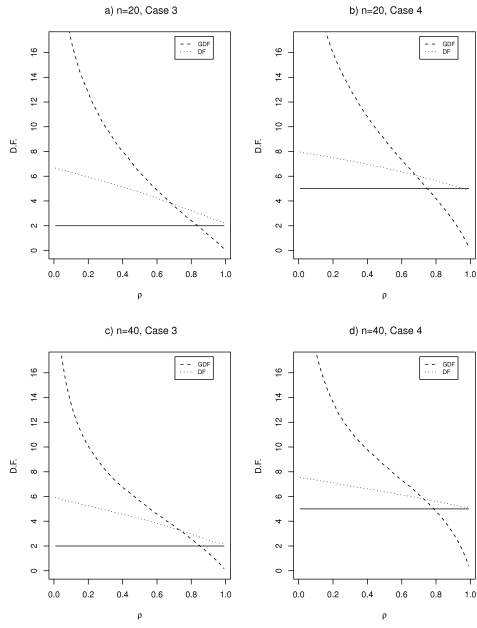
AIC : $\rho=0.25$



AIC : $\rho=0.75$



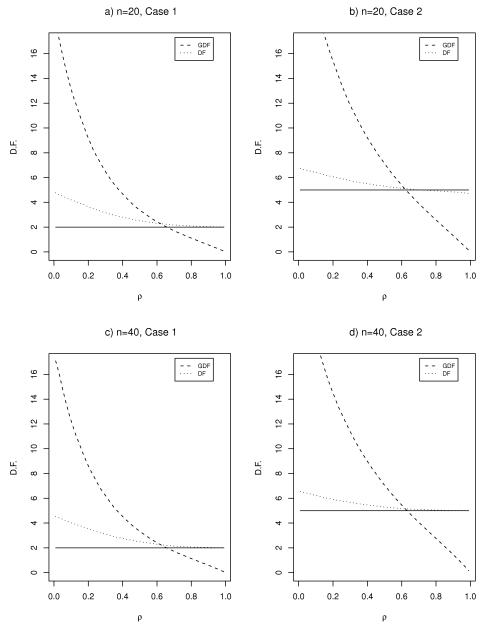
BIC : $\rho=0.25$



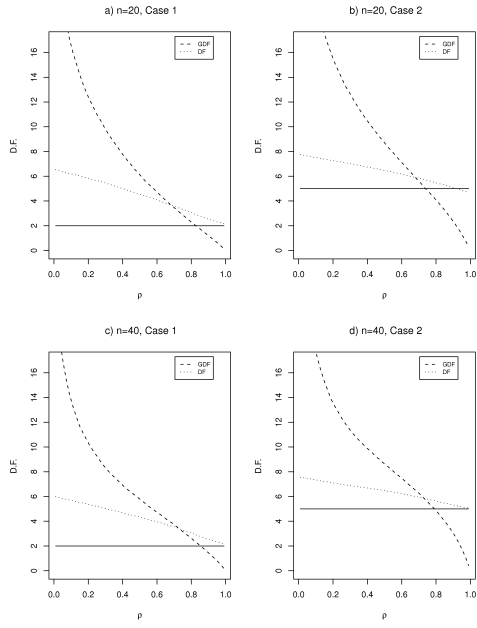
BIC : $\rho=0.75$

C.3.2 With Parsimony Correction

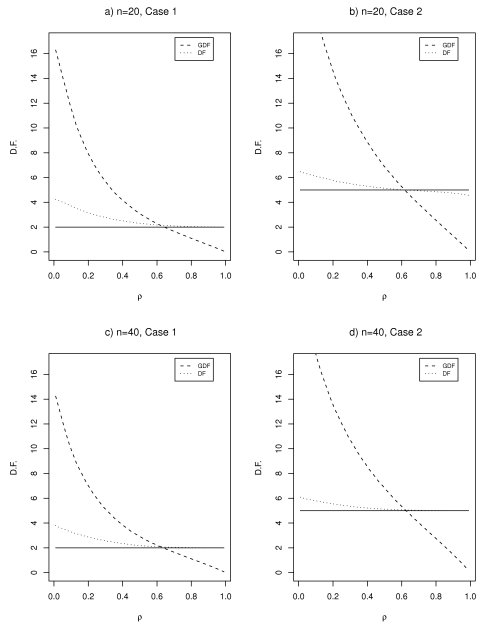
C.3.2.1 Cases 1 and 2



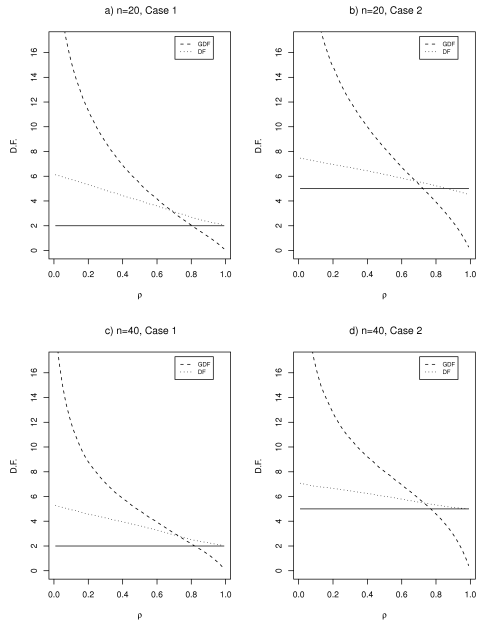
AIC : $\rho=0.25$



AIC : $\rho=0.75$

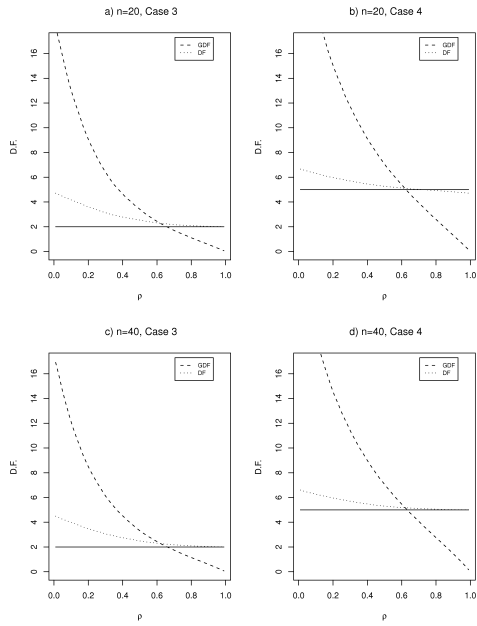


BIC : $\rho=0.25$

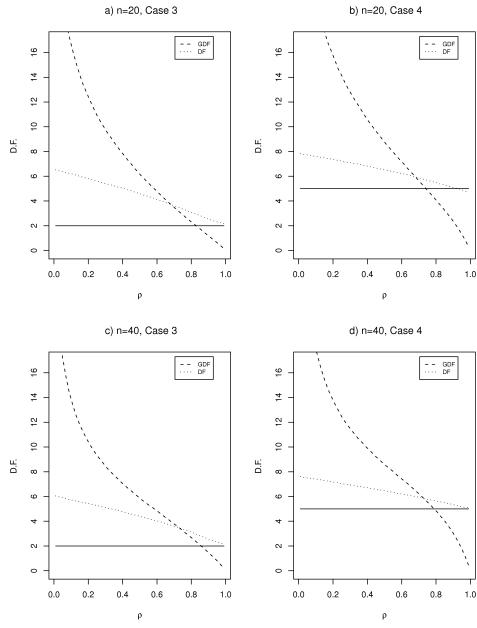


BIC : $\rho=0.75$

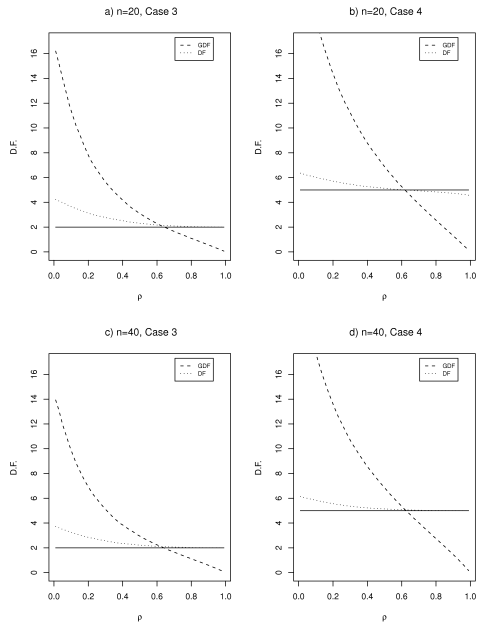
C.3.2.2 Cases 3 and 4



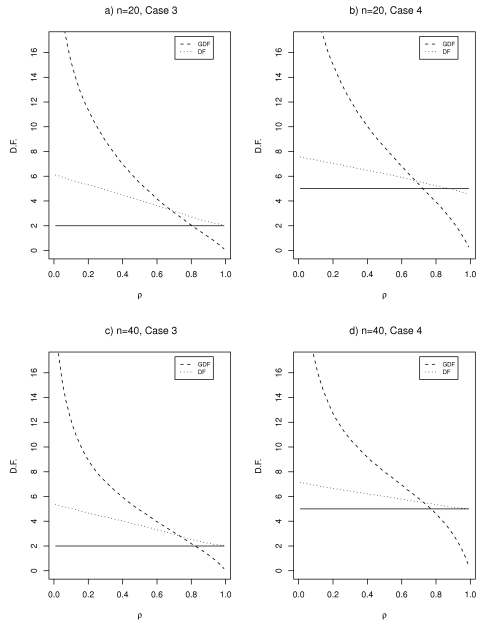
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

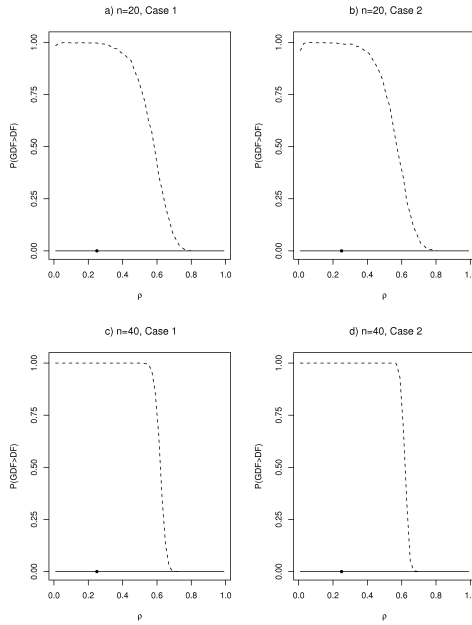


BIC : $\rho=0.75$

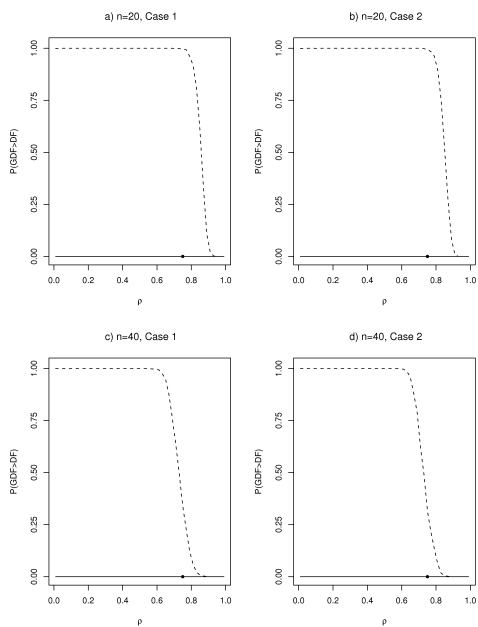
C.4 Relationship between GDF and Model DF – Proof-of-Concept

C.4.1 No parsimony correction

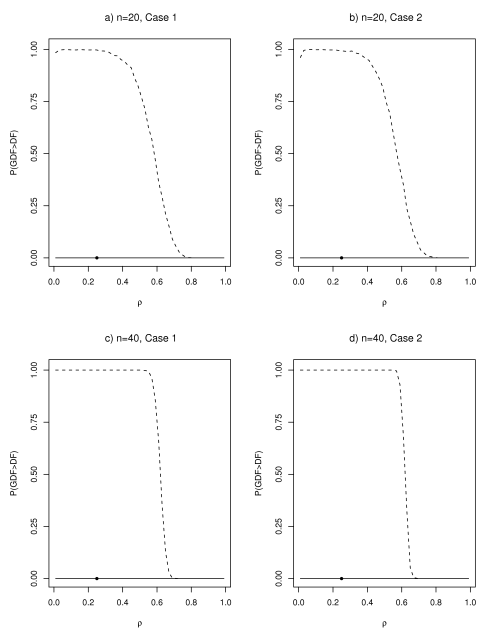
C.4.1.1 Cases 1 and 2



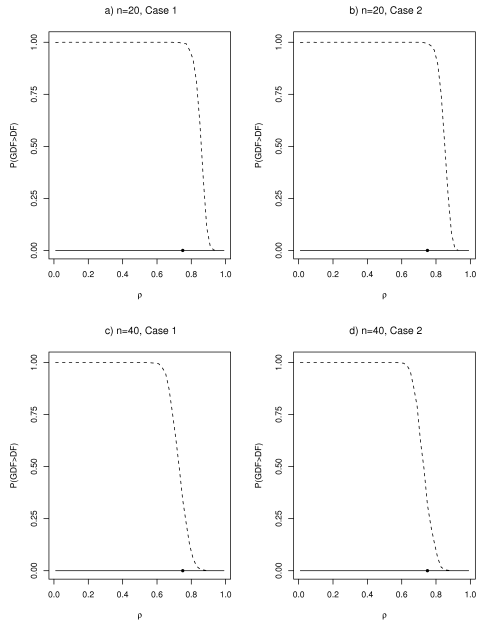
AIC : $\rho=0.25$



AIC : $\rho=0.75$

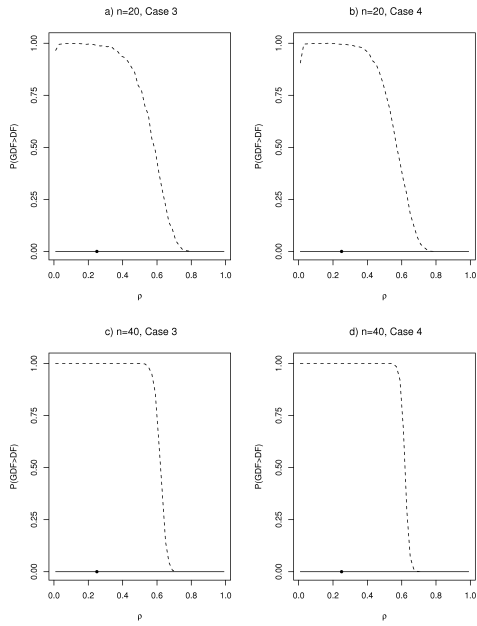


BIC : $\rho=0.25$

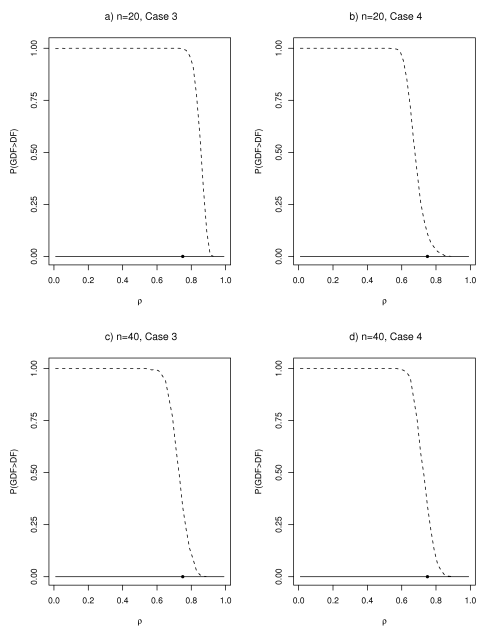


BIC : $\rho=0.75$

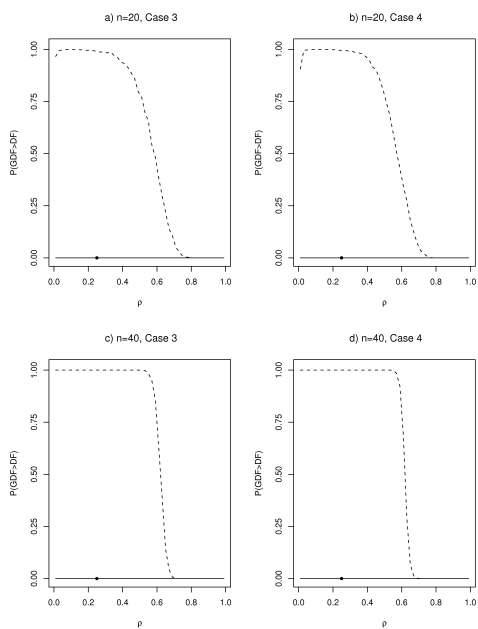
C.4.1.2 Cases 3 and 4



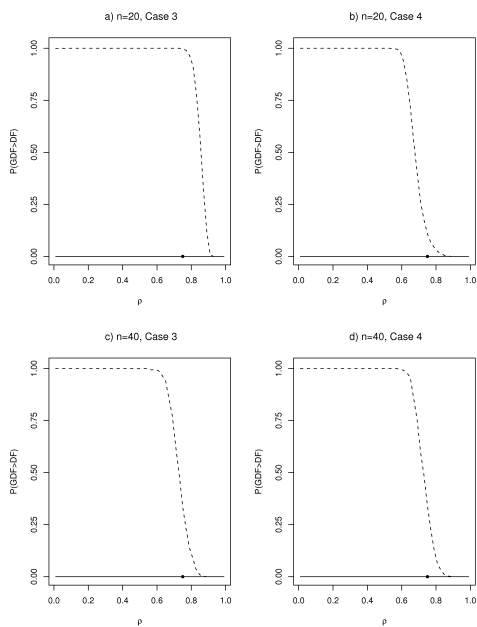
AIC : $\rho=0.25$



AIC : $\rho=0.75$



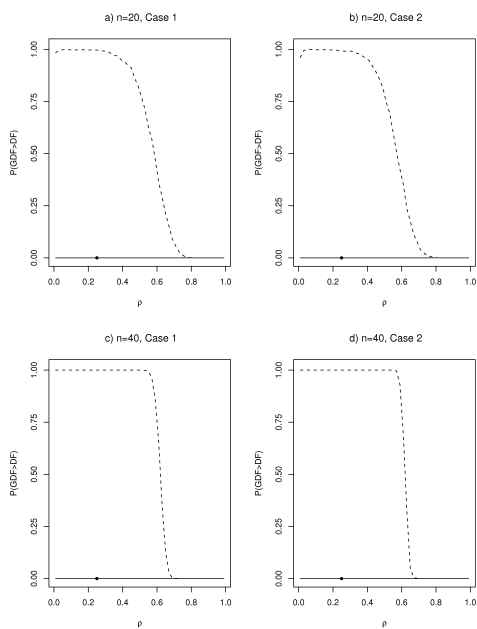
BIC : $\rho=0.25$



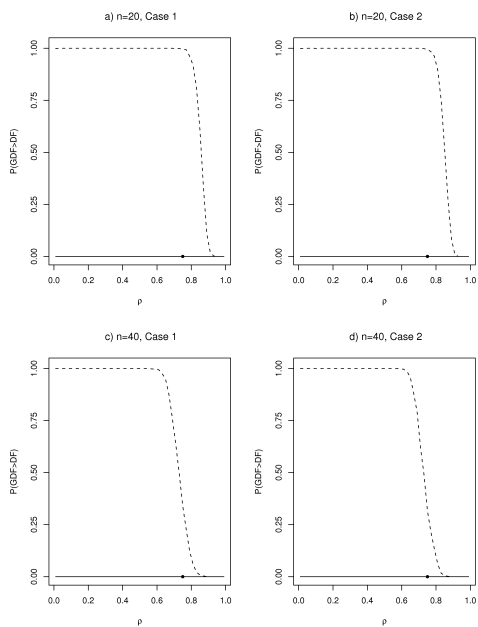
BIC : $\rho=0.75$

C.4.2 With Parsimony Correction

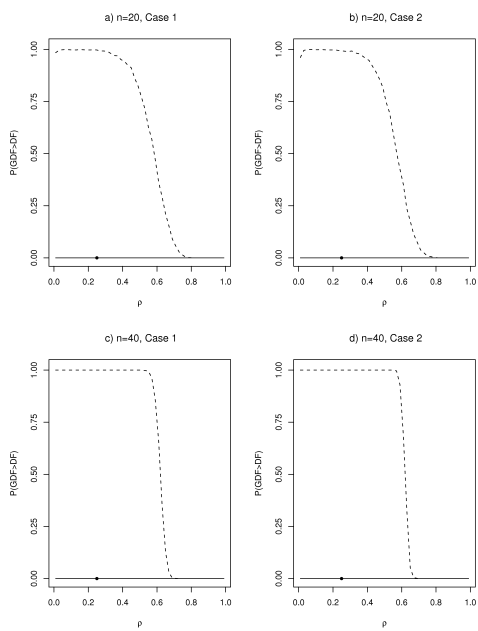
C.4.2.1 Cases 1 and 2



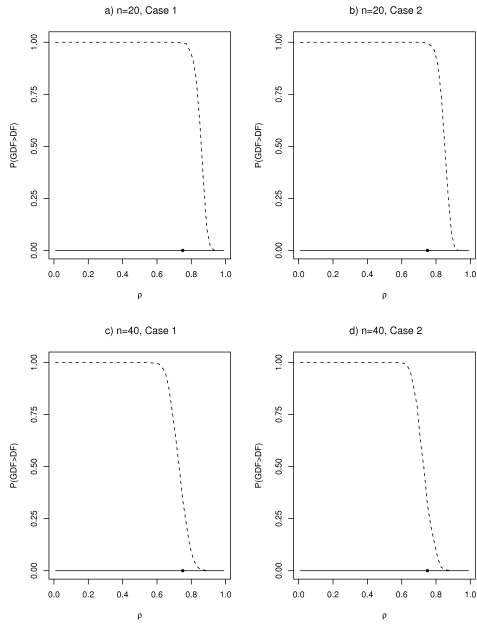
AIC : $\rho=0.25$



AIC : $\rho=0.75$

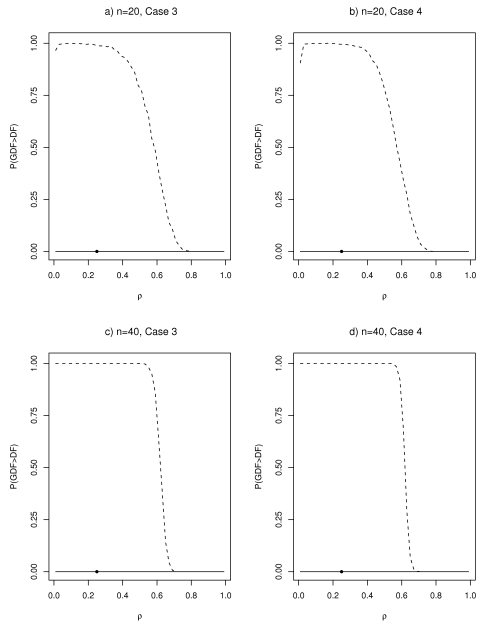


BIC : $\rho=0.25$

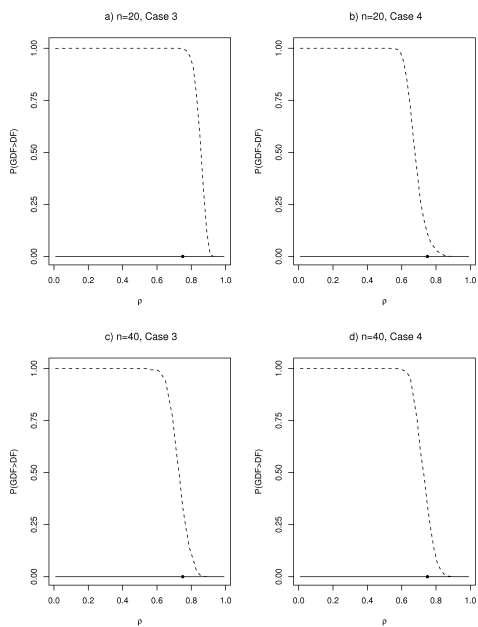


BIC : $\rho=0.75$

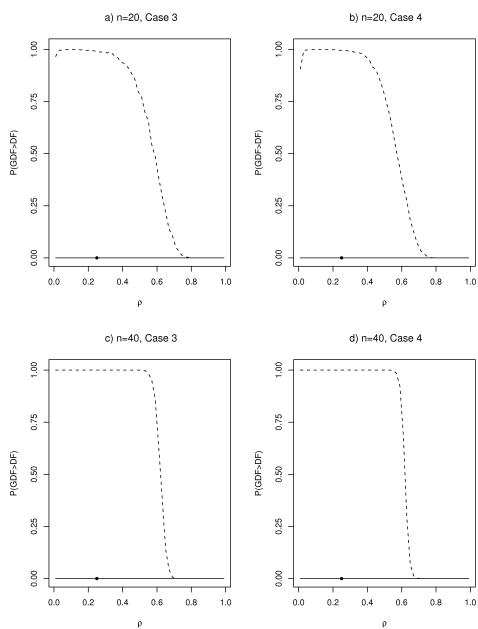
C.4.2.2 Cases 3 and 4



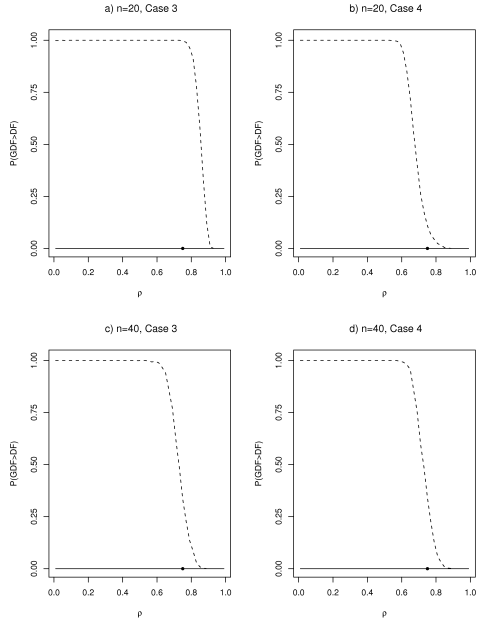
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

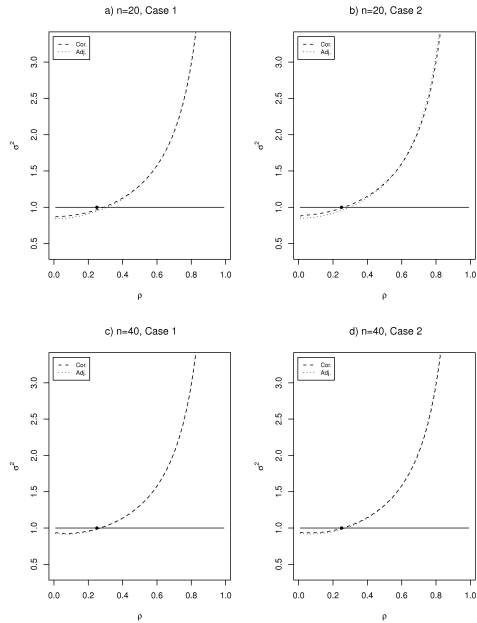


BIC : $\rho=0.75$

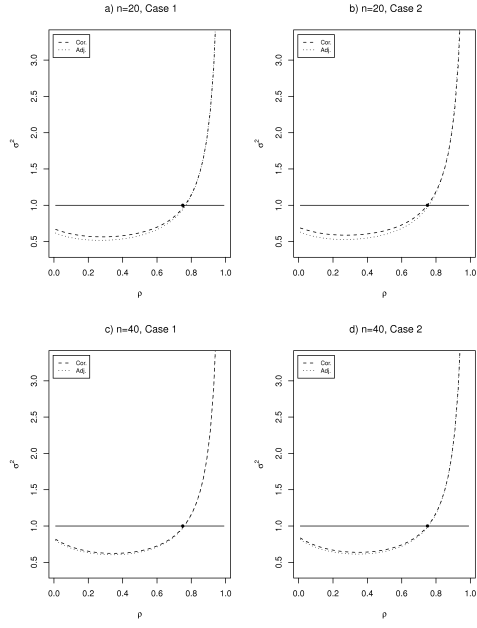
C.5 Corrected Variance-Proof-of-Concept

C.5.1 No Parsimony Correction

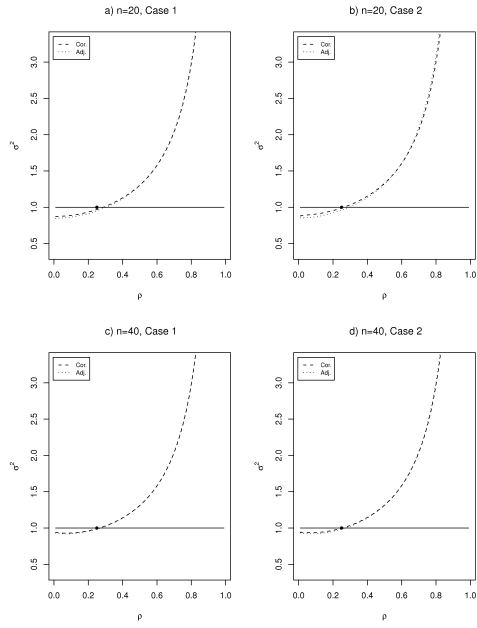
C.5.1.1 Cases 1 and 2



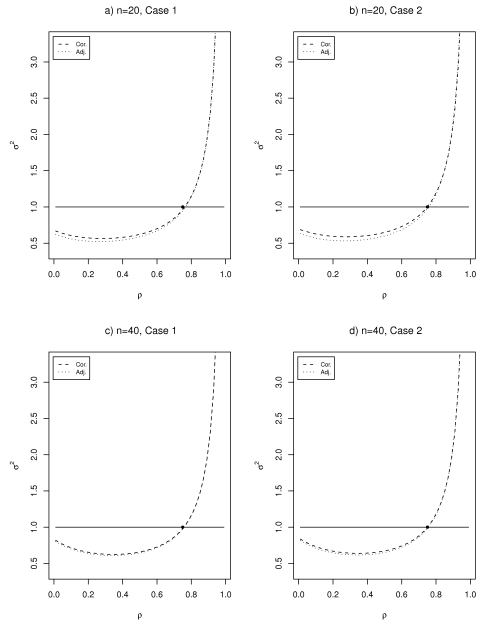
AIC : $\rho=0.25$



AIC : $\rho=0.75$

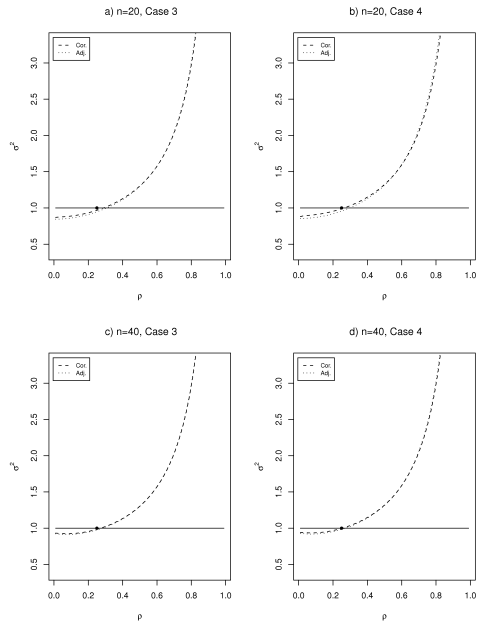


BIC : $\rho=0.25$

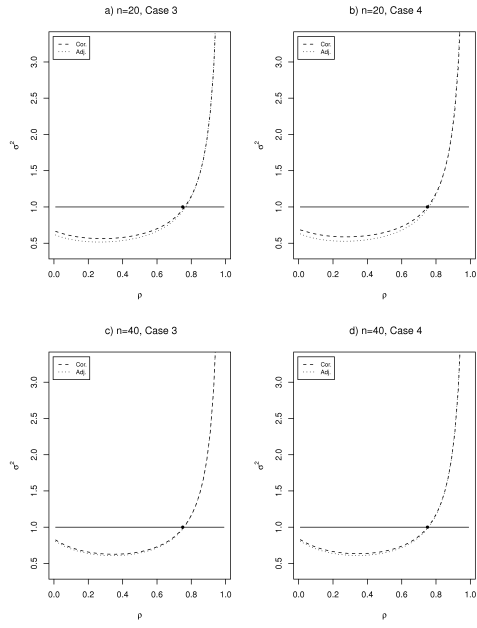


BIC : $\rho=0.75$

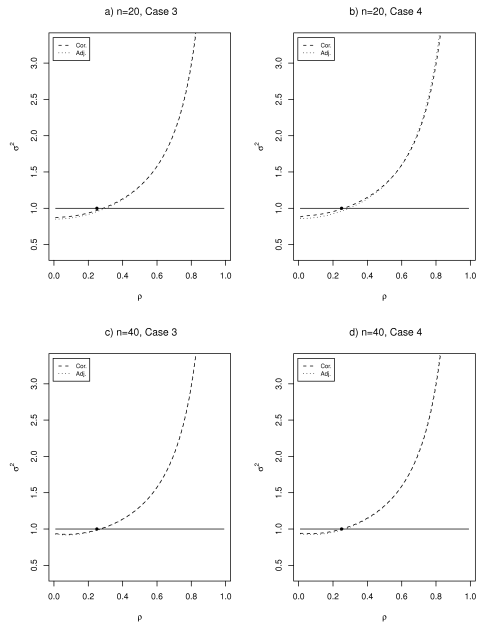
C.5.1.2 Cases 3 and 4



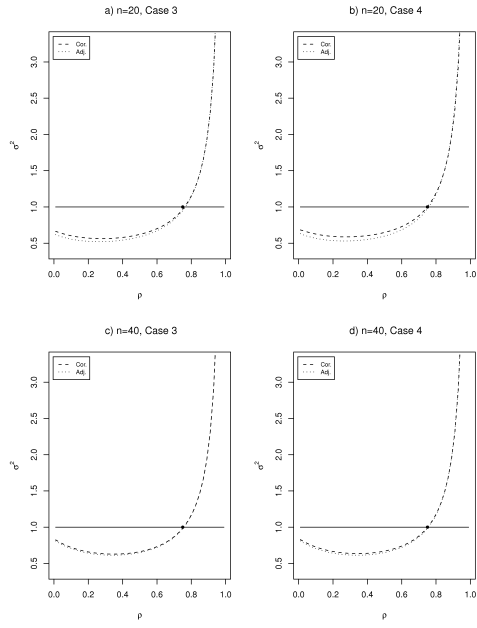
AIC : $\rho=0.25$



AIC : $\rho=0.75$



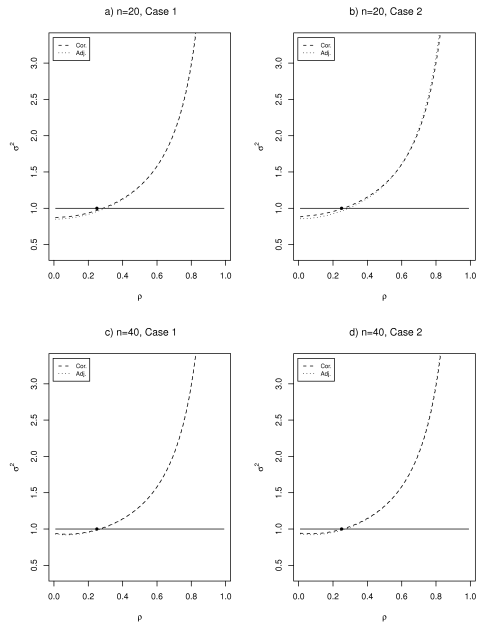
BIC : $\rho=0.25$



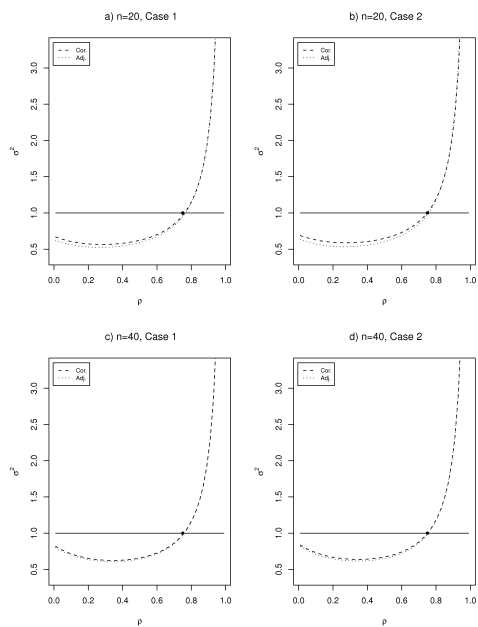
BIC : $\rho=0.75$

C.5.2 With Parsimony Correction

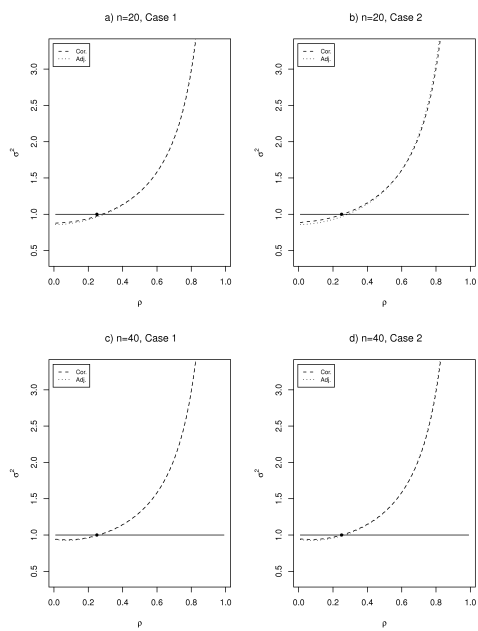
C.5.2.1 Cases 1 and 2



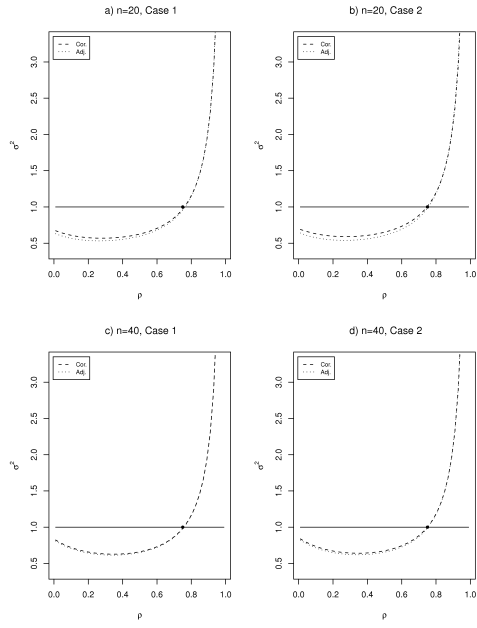
AIC : $\rho=0.25$



AIC : $\rho=0.75$

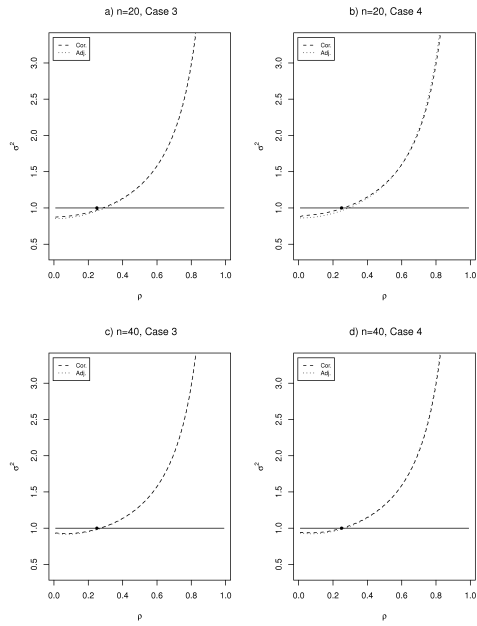


BIC : $\rho=0.25$

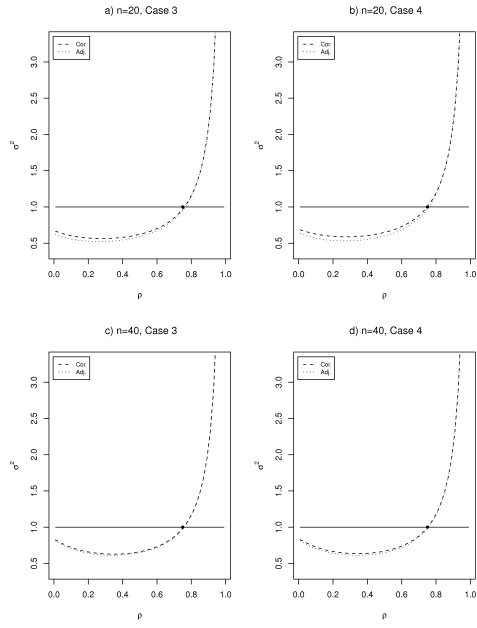


BIC : $\rho=0.75$

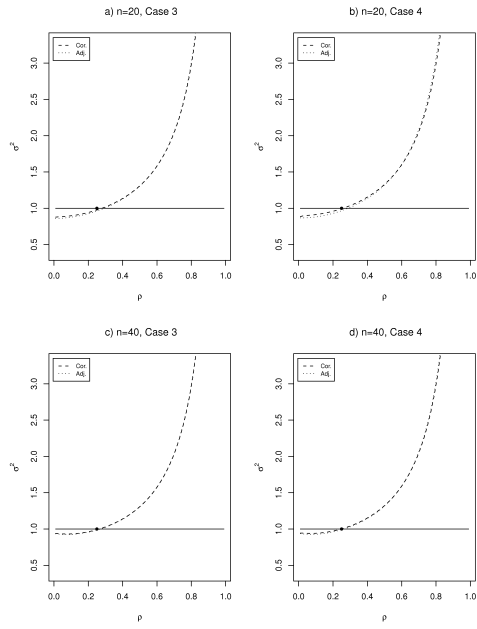
C.5.2.2 Cases 3 and 4



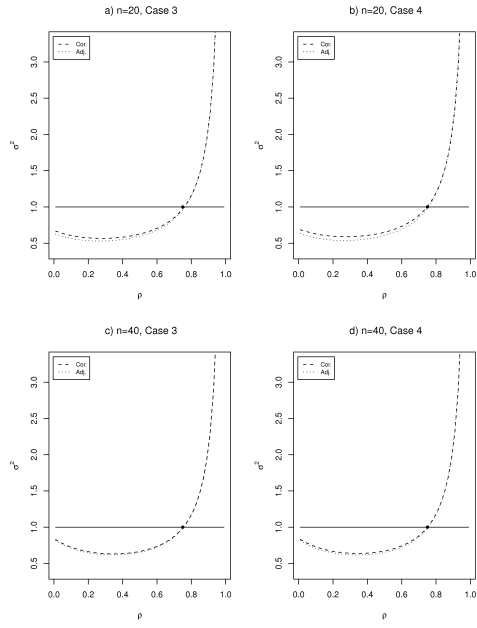
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

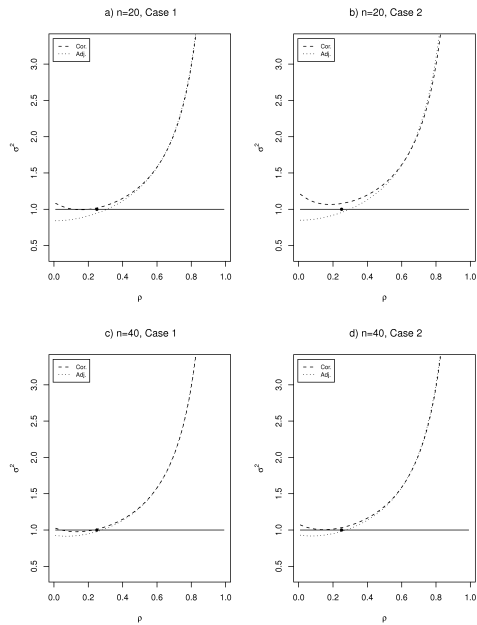


BIC : $\rho=0.75$

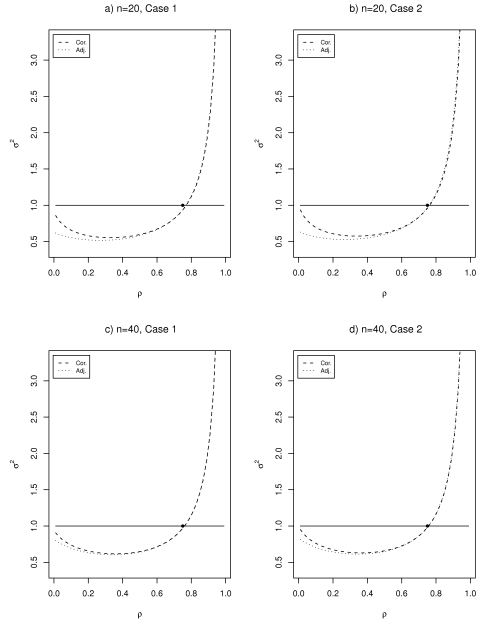
C.6 Corrected Variance-Data-Driven Method

C.6.1 No Parsimony Correction

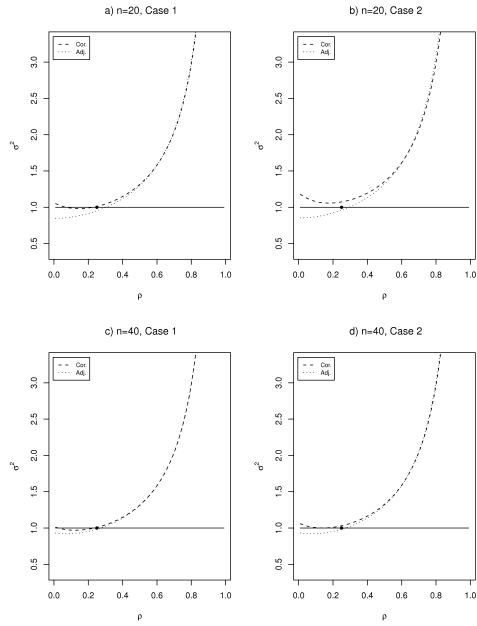
C.6.1.1 Cases 1 and 2



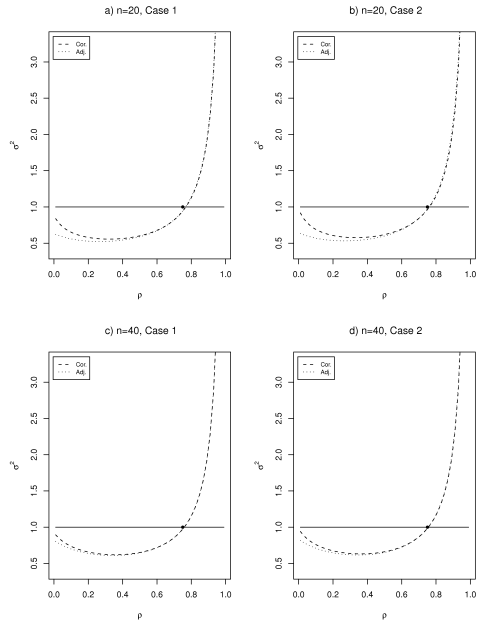
AIC : $\rho=0.25$



AIC : $\rho=0.75$

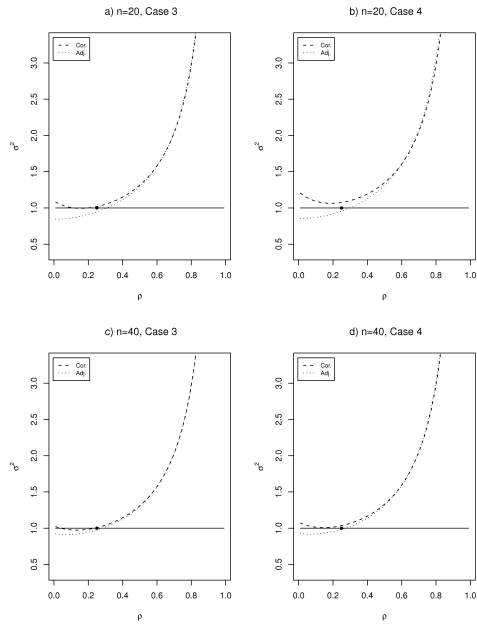


BIC : $\rho=0.25$

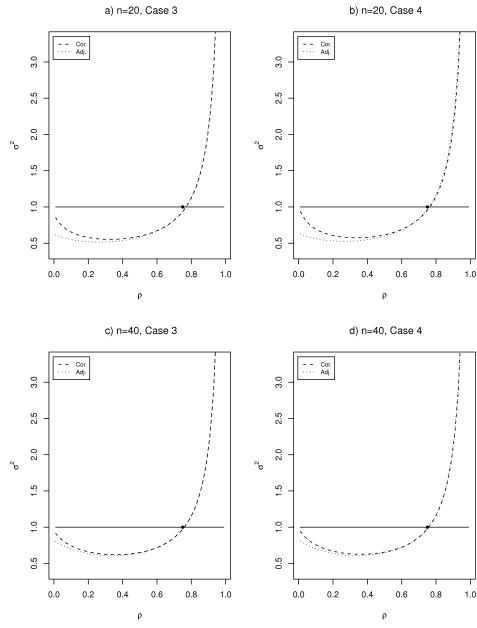


BIC : $\rho=0.75$

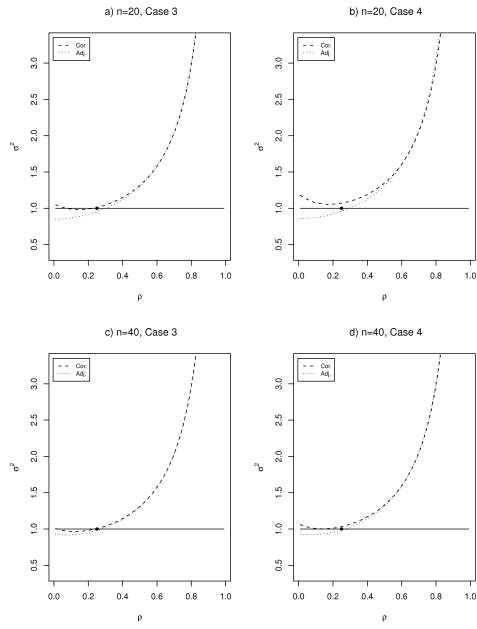
C.6.1.2 Cases 3 and 4



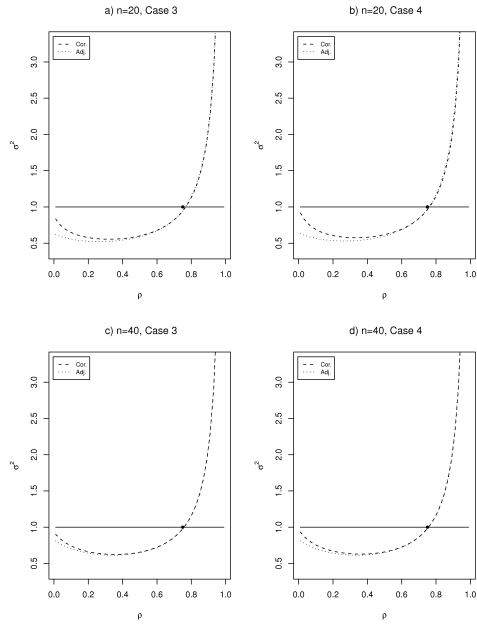
AIC : $\rho=0.25$



AIC : $\rho=0.75$



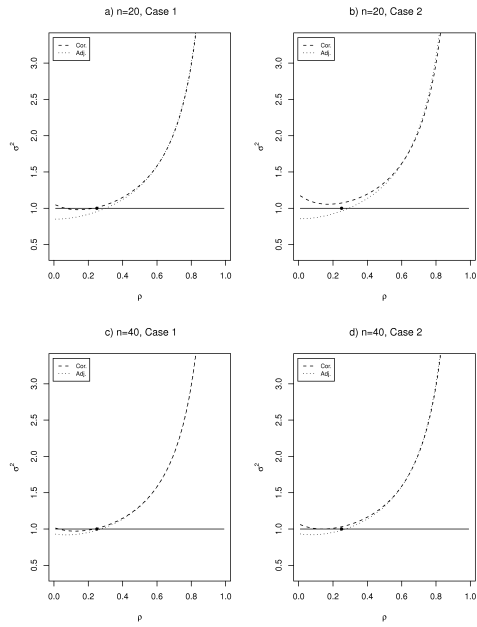
BIC : $\rho=0.25$



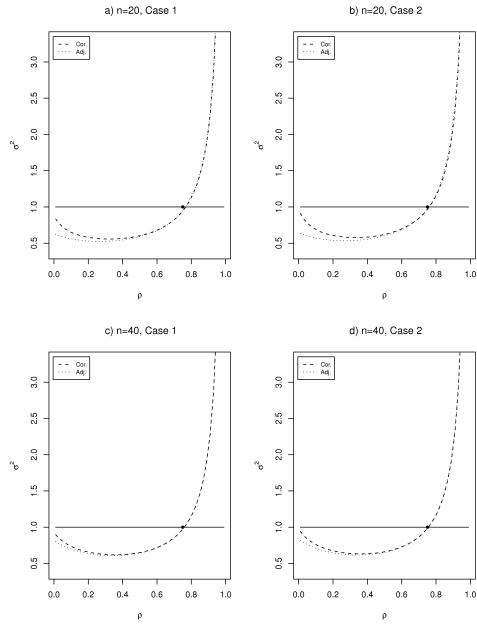
BIC : $\rho=0.75$

C.6.2 With Parsimony Correction

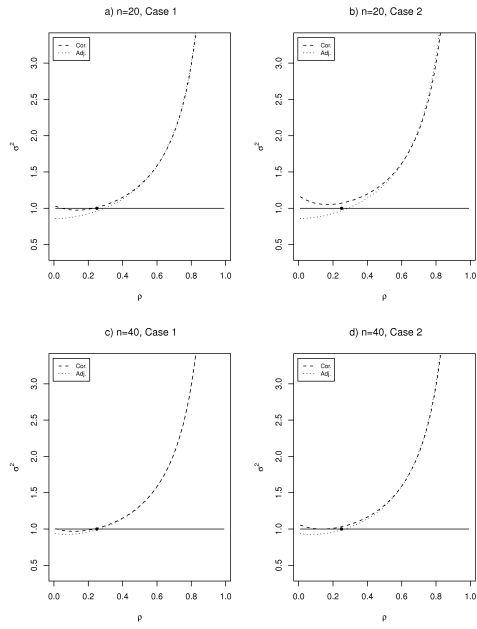
C.6.2.1 Cases 1 and 2



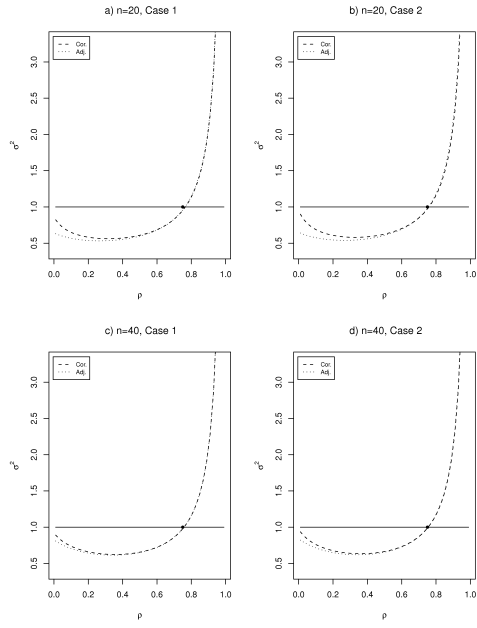
AIC : $\rho=0.25$



AIC : $\rho=0.75$

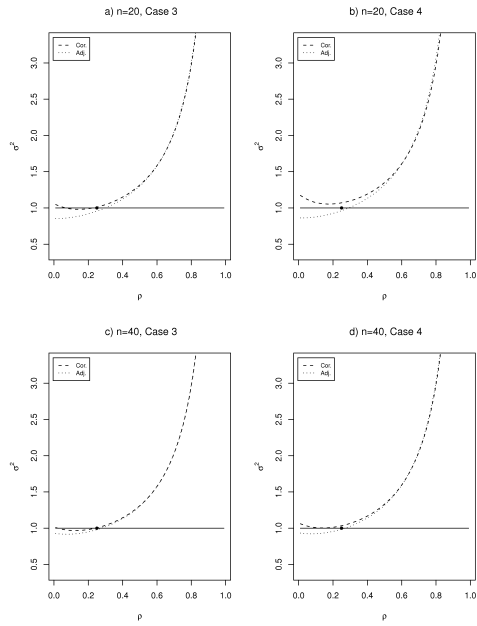


BIC : $\rho=0.25$

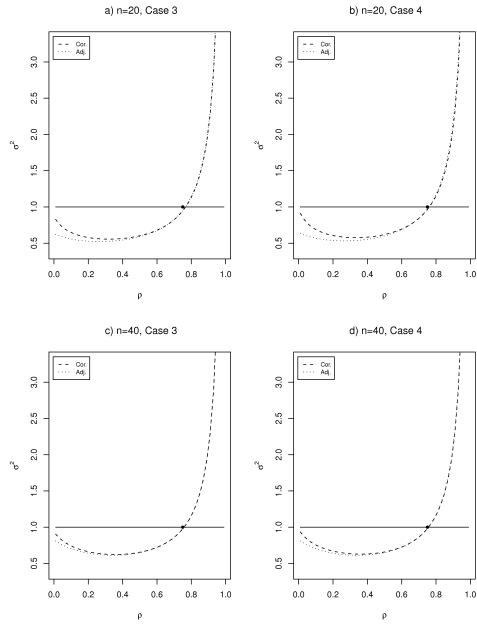


BIC : $\rho=0.75$

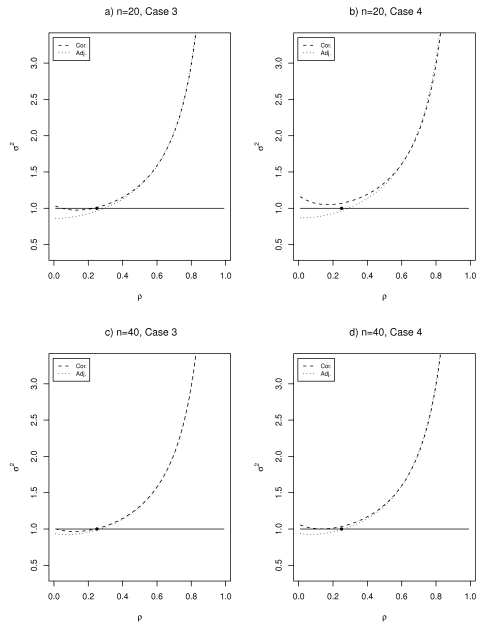
C.6.2.2 Cases 3 and 4



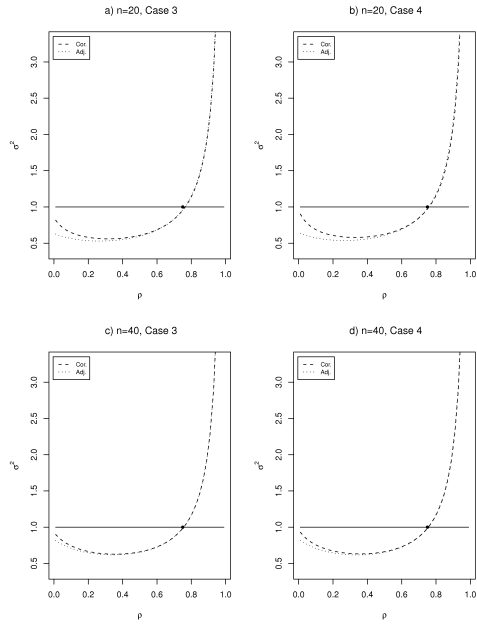
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$



BIC : $\rho=0.75$

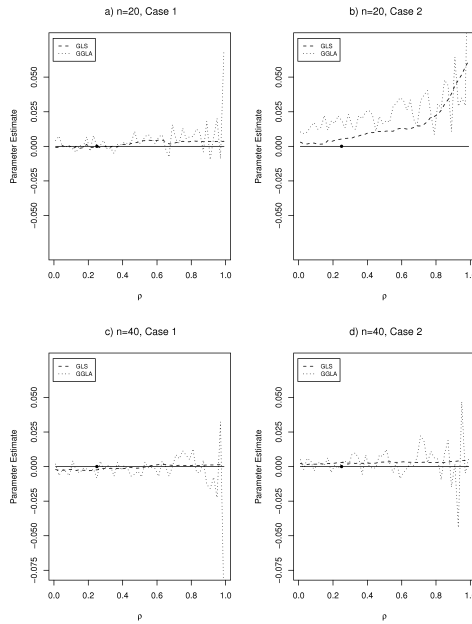
APPENDIX D

Extended GGLA Simulation Results

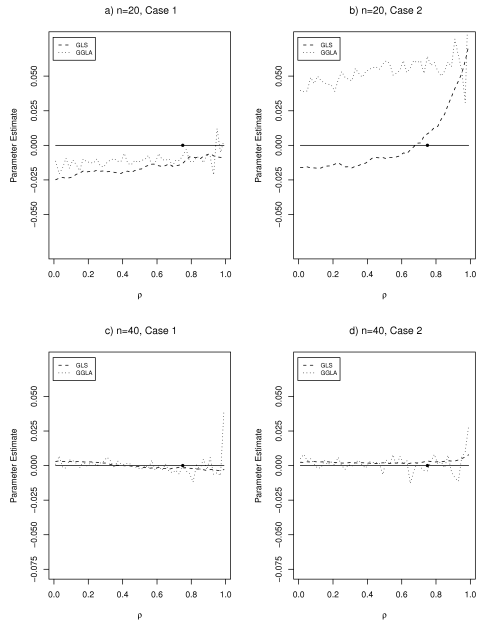
D.1 Fixed Effect Estimation–Proof-of-Concept

D.1.1 No Parsimony Correction

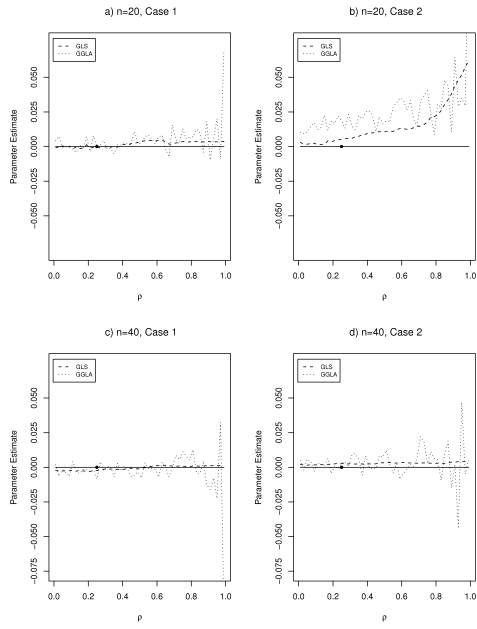
D.1.1.1 Cases 1 and 2



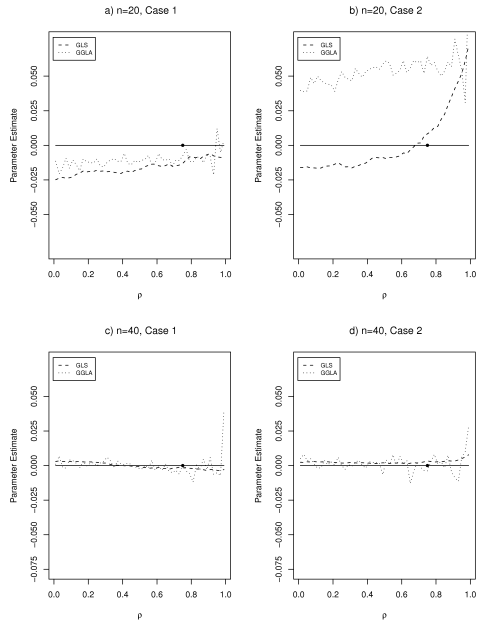
AIC : $\rho=0.25$



AIC : $\rho=0.75$

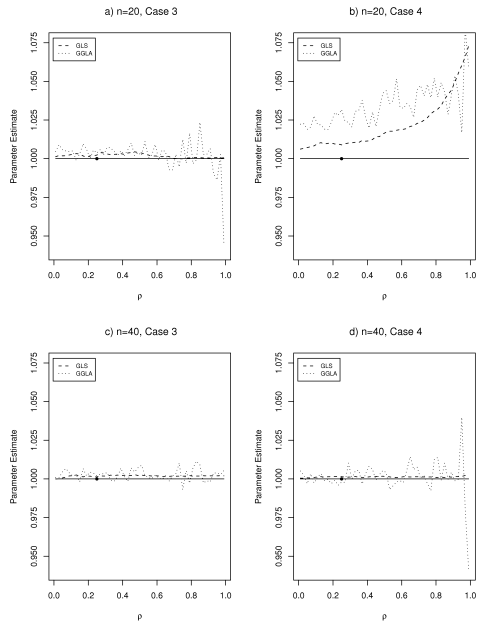


BIC : $\rho=0.25$

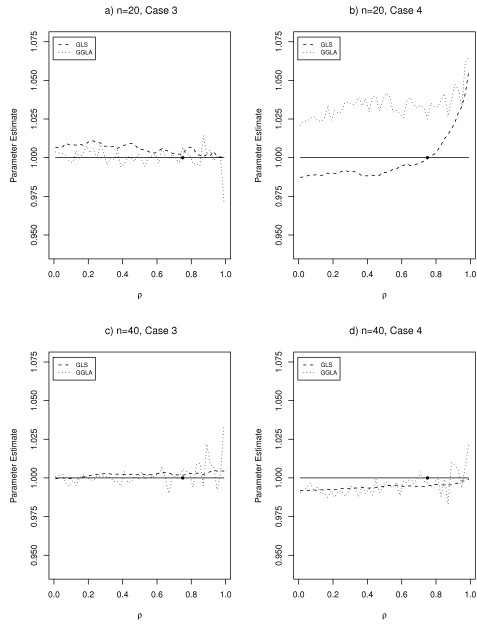


BIC : $\rho=0.75$

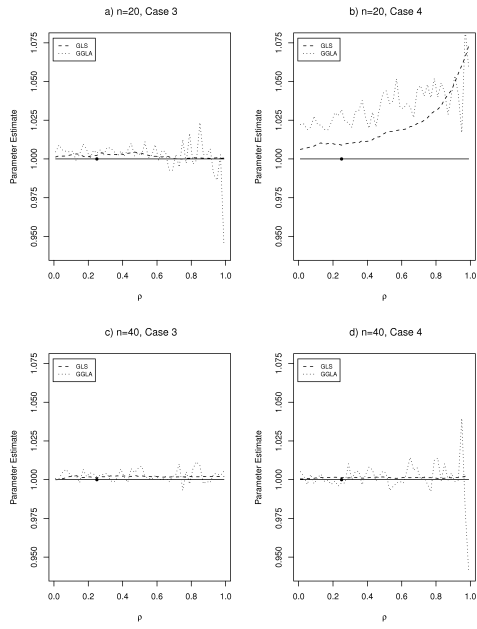
D.1.1.2 Cases 3 and 4



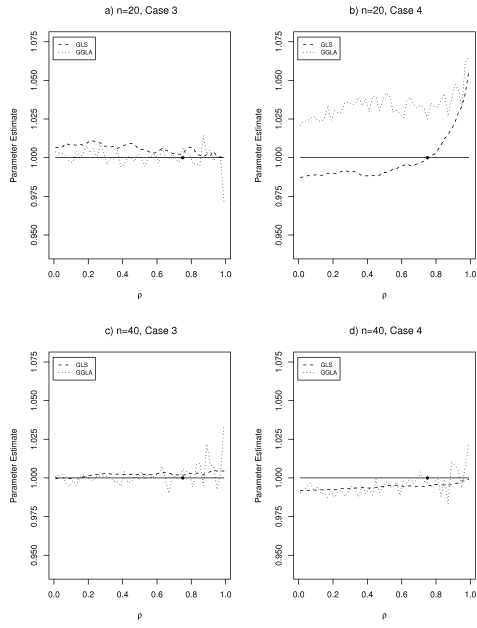
AIC : $\rho=0.25$



AIC : $\rho=0.75$



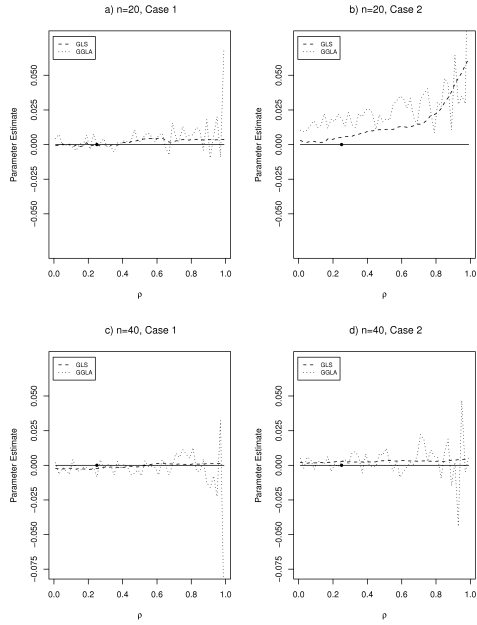
BIC : $\rho=0.25$



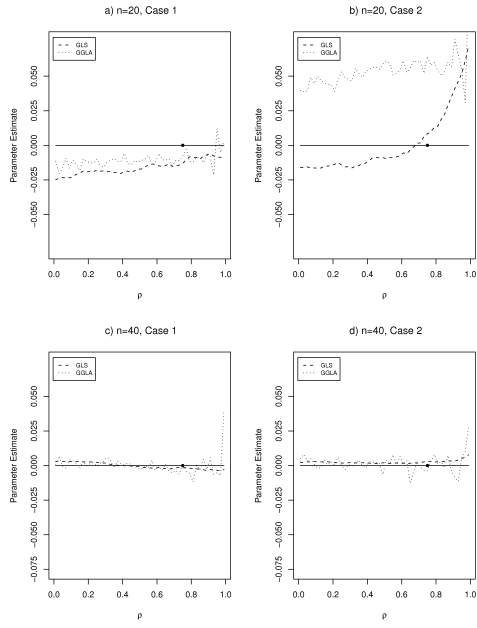
BIC : $\rho=0.75$

D.1.2 With Parsimony Correction

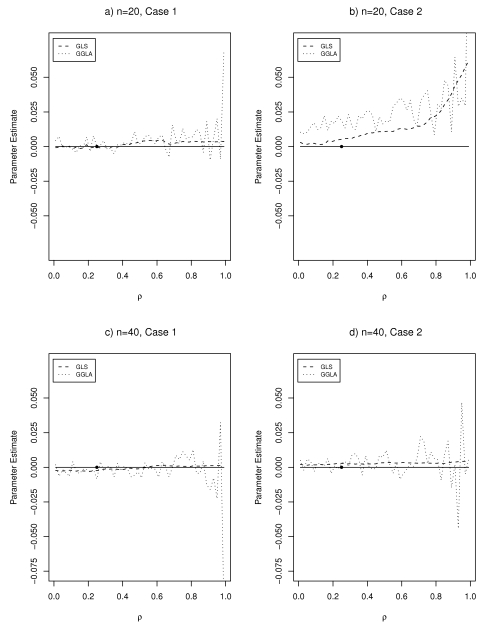
D.1.2.1 Cases 1 and 2



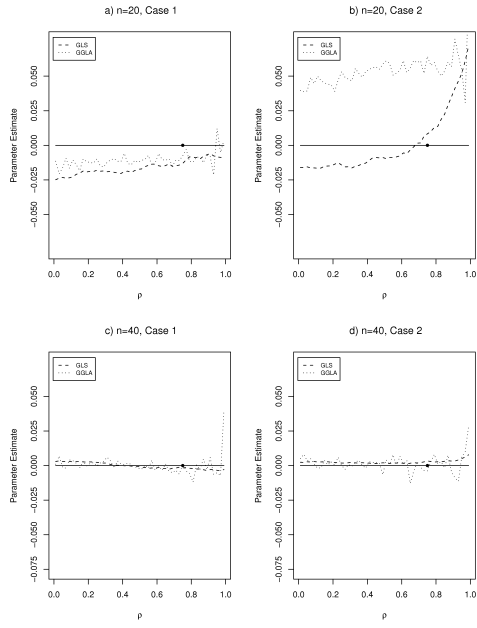
AIC : $\rho=0.25$



AIC : $\rho=0.75$

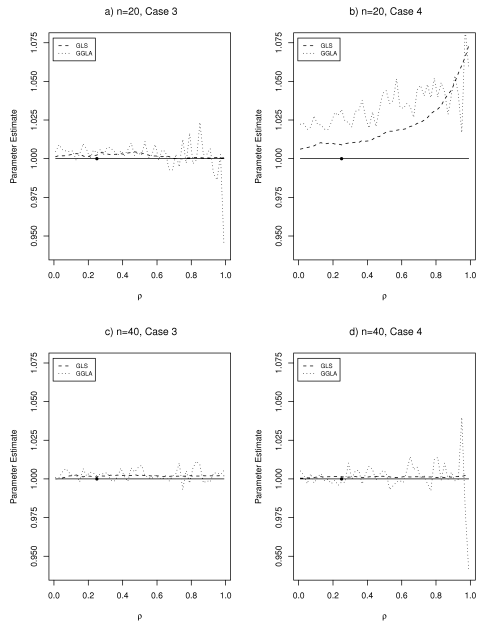


BIC : $\rho=0.25$

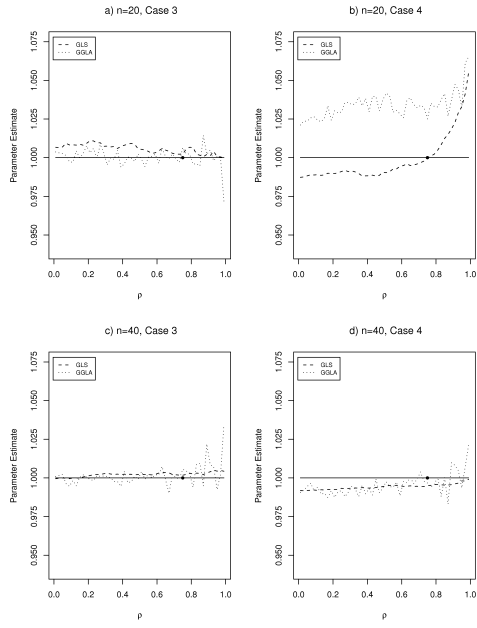


BIC : $\rho=0.75$

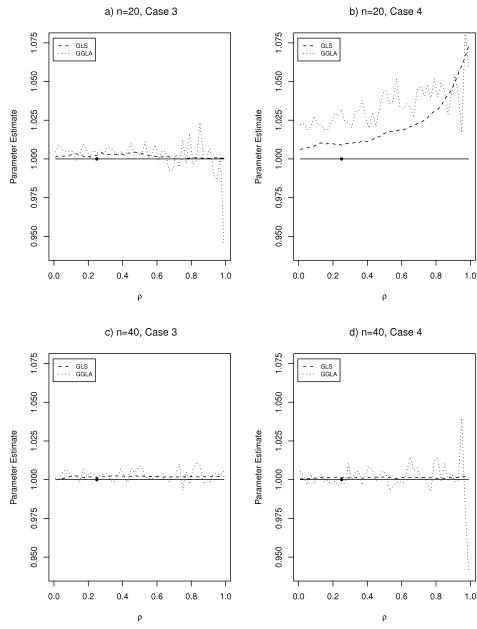
D.1.2.2 Cases 3 and 4



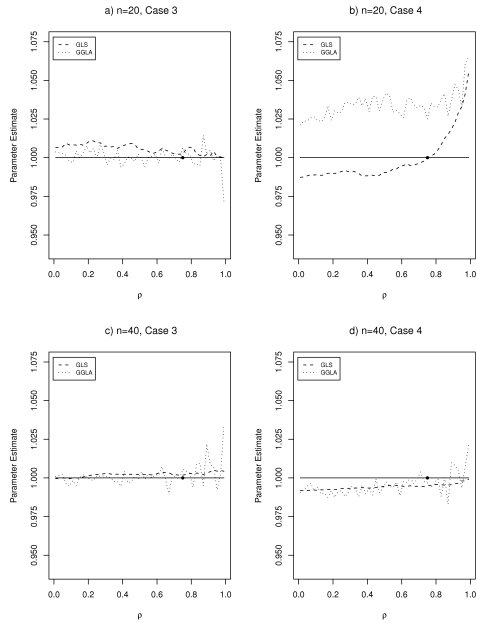
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

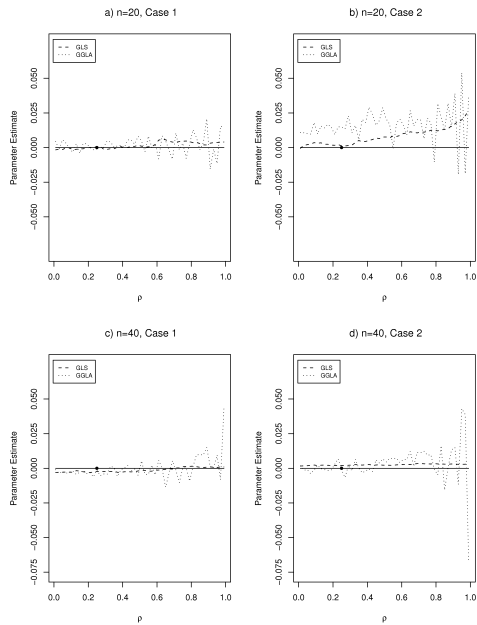


BIC : $\rho=0.75$

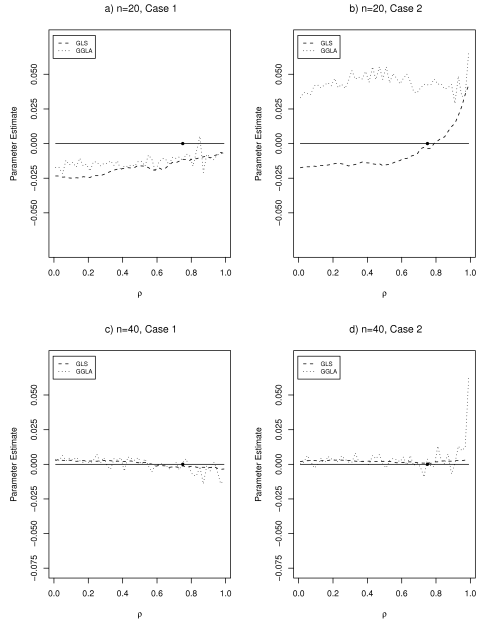
D.2 Fixed Effect Estimation-Data-Driven Method

D.2.1 No Parsimony Correction

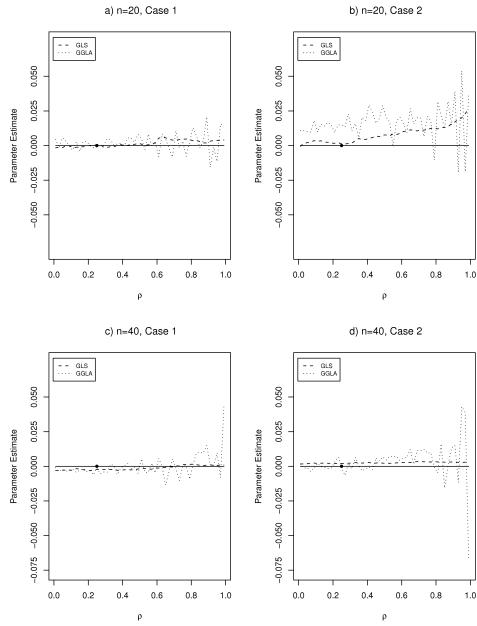
D.2.1.1 Cases 1 and 2



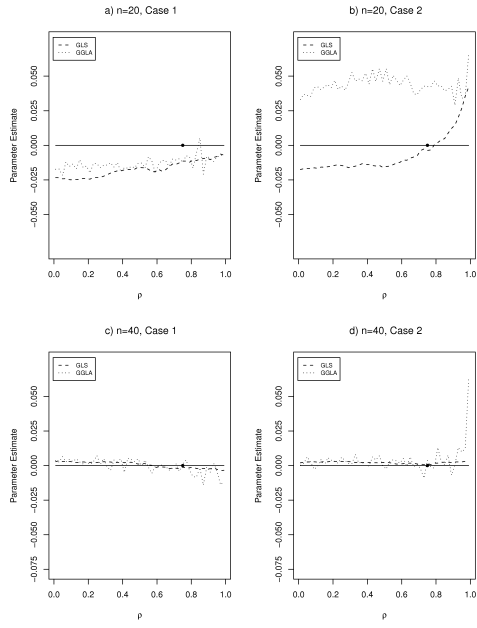
AIC : $\rho=0.25$



AIC : $\rho=0.75$

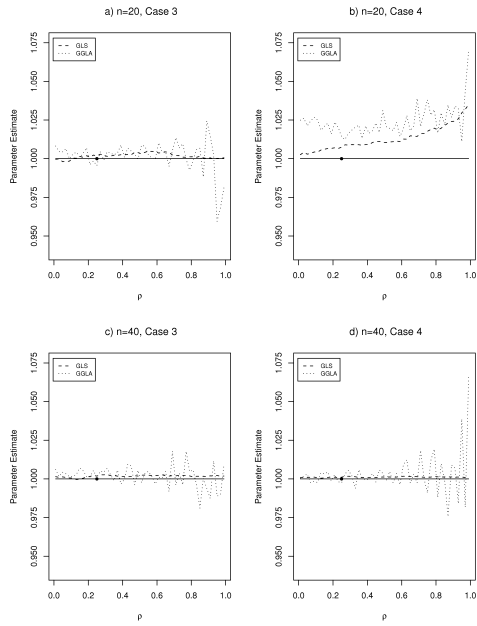


BIC : $\rho=0.25$

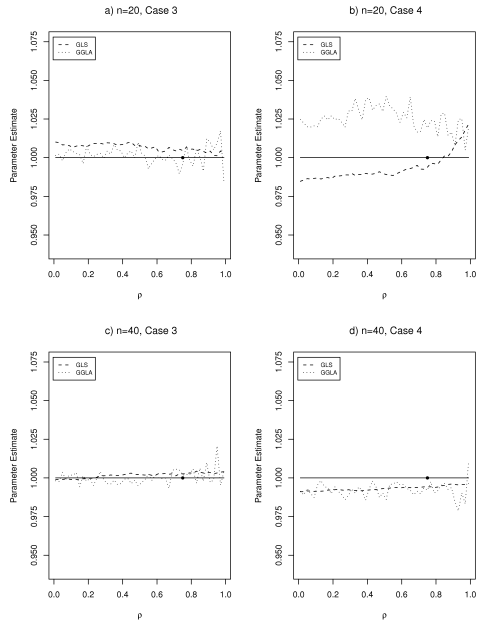


BIC : $\rho=0.75$

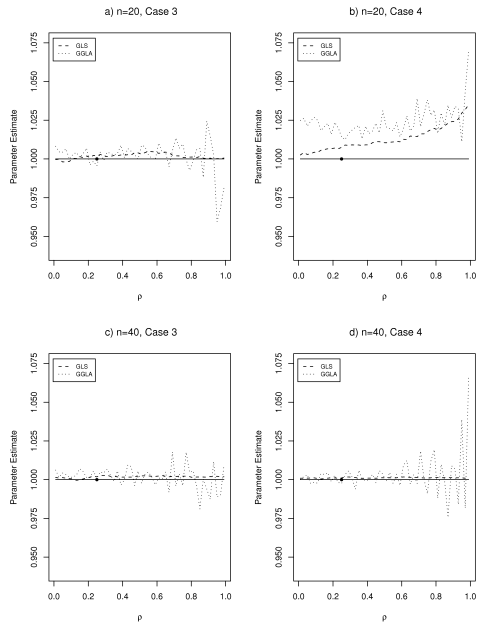
D.2.1.2 Cases 3 and 4



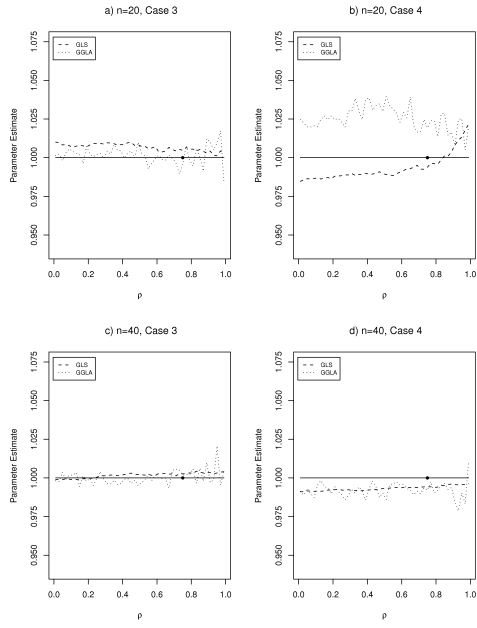
AIC : $\rho=0.25$



AIC : $\rho=0.75$



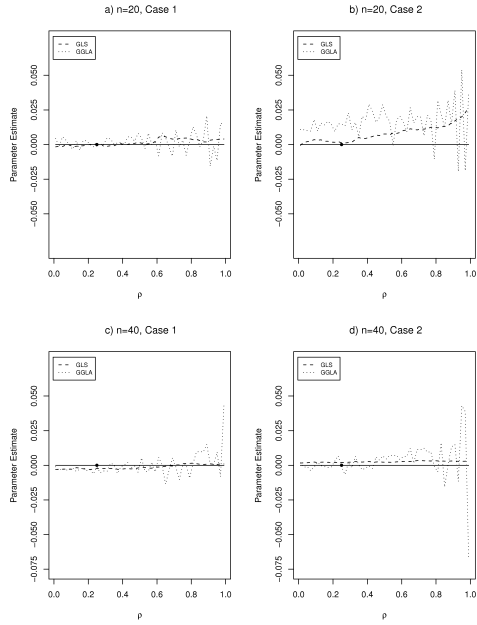
BIC : $\rho=0.25$



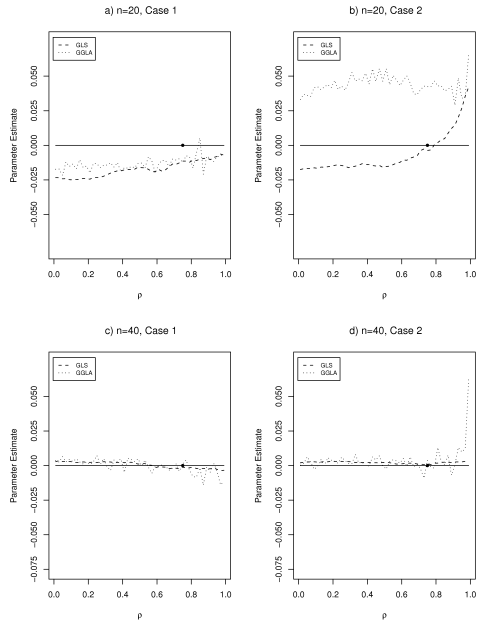
BIC : $\rho=0.75$

D.2.2 With Parsimony Correction

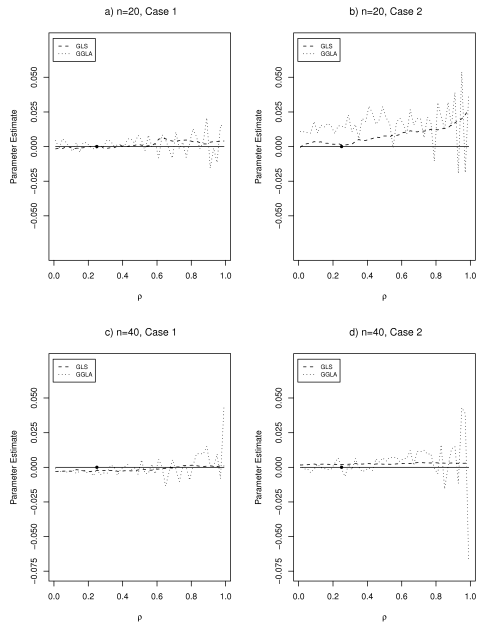
D.2.2.1 Cases 1 and 2



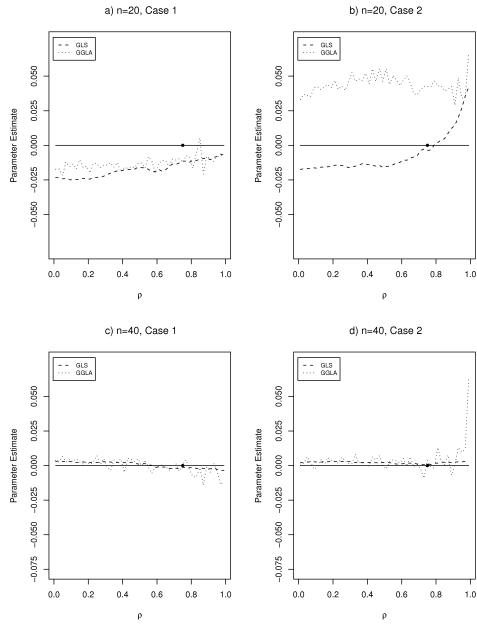
AIC : $\rho=0.25$



AIC : $\rho=0.75$

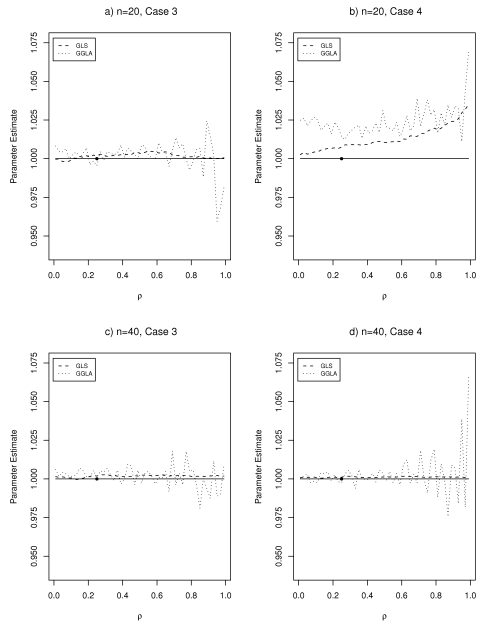


BIC : $\rho=0.25$

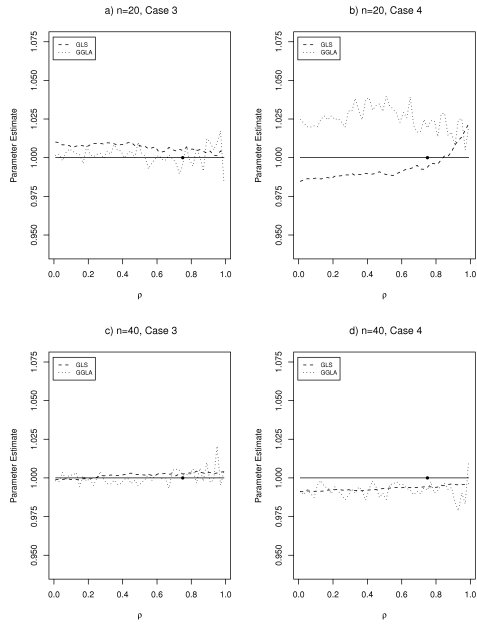


BIC : $\rho=0.75$

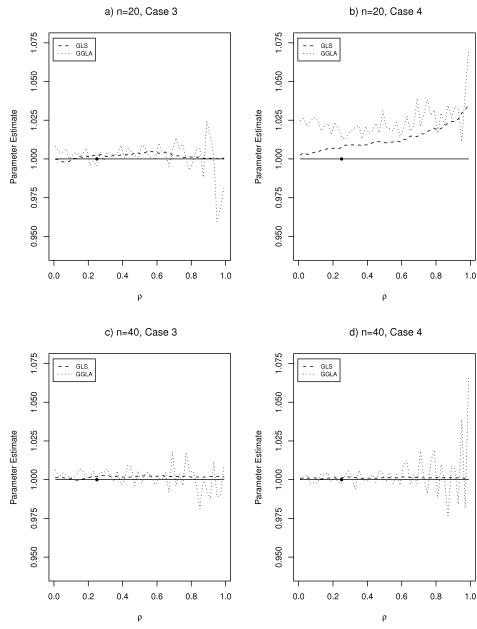
D.2.2.2 Cases 3 and 4



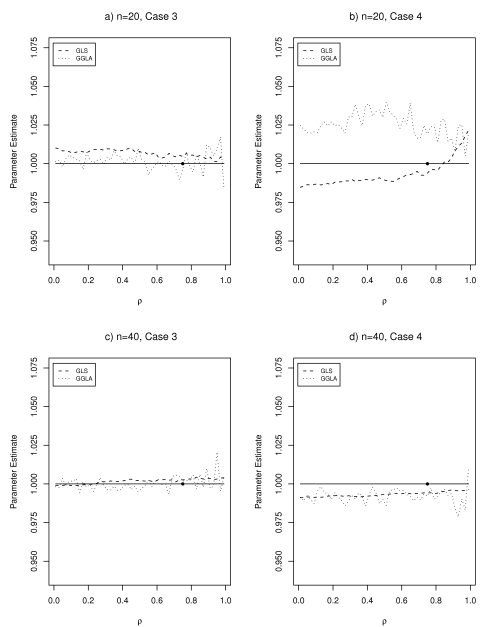
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

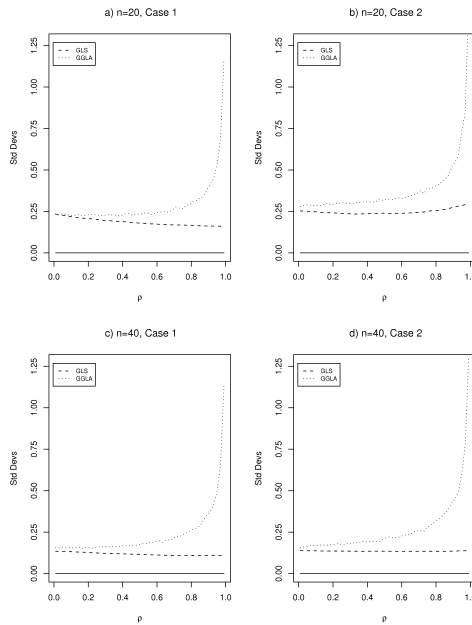


BIC : $\rho=0.75$

D.3 Simulation Standard Deviation of Fixed-Effect Estimates— Proof-of-Concept

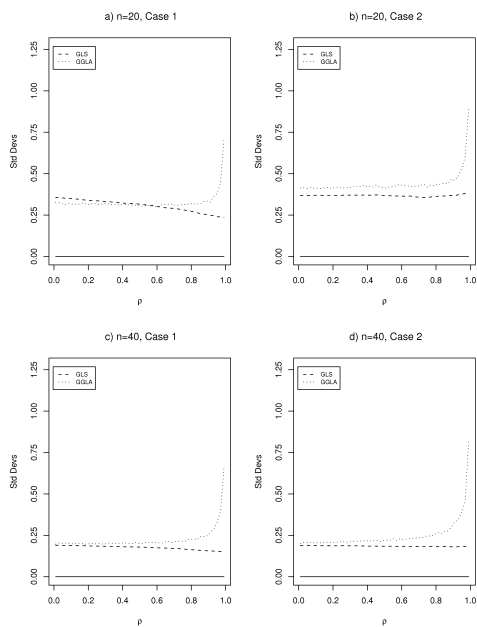
D.3.1 No Parsimony Correction

D.3.1.1 Cases 1 and 2

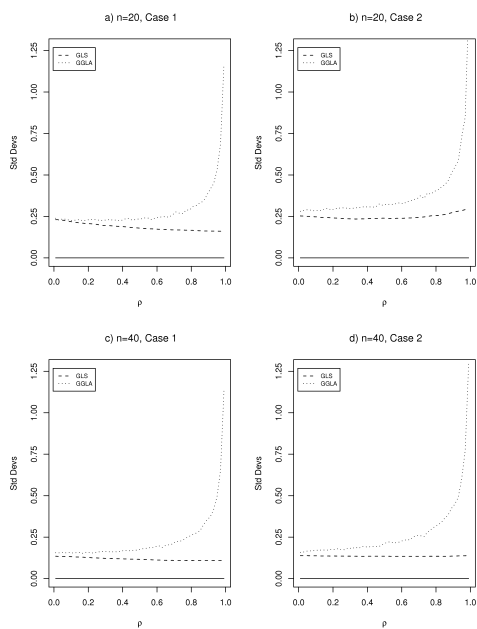


AIC : $\rho=0.25$

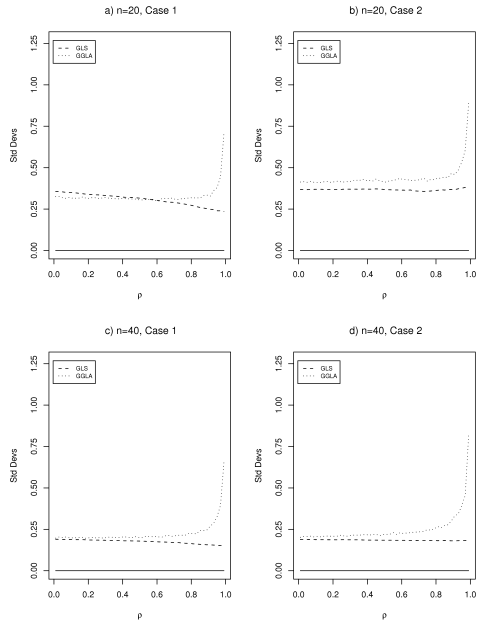
F



AIC : $\rho=0.75$

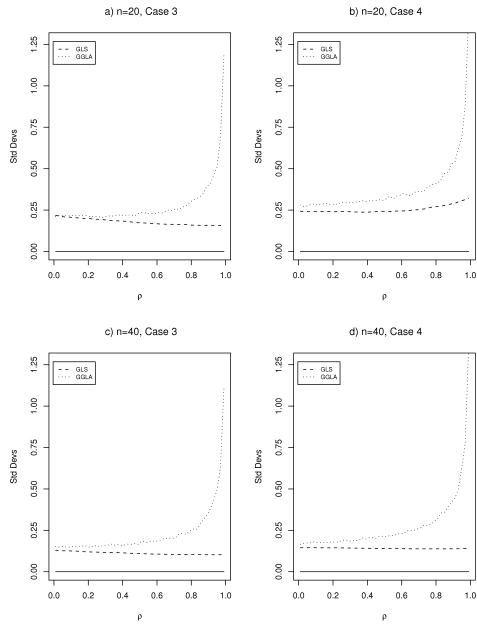


BIC : $\rho=0.25$

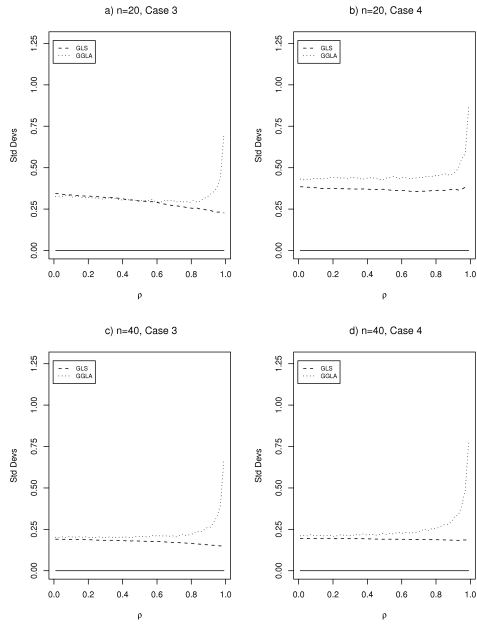


BIC : $\rho=0.75$

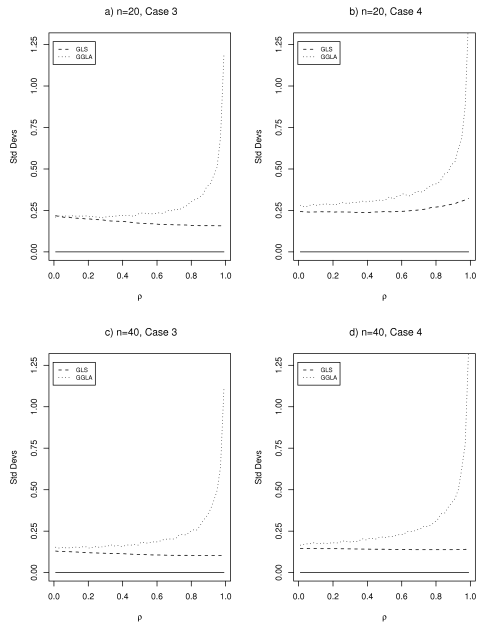
D.3.1.2 Cases 3 and 4



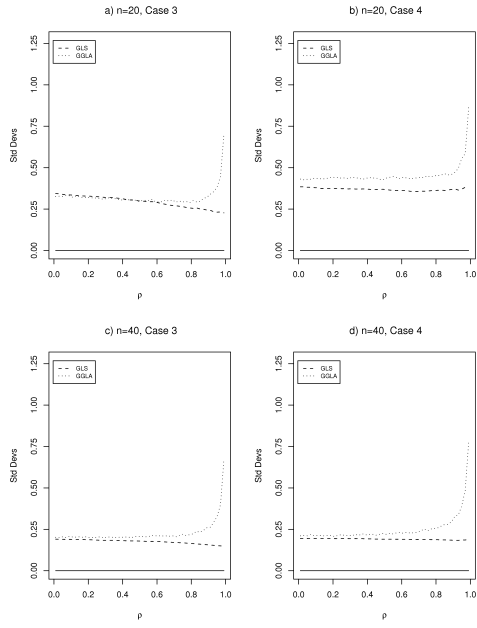
AIC : $\rho=0.25$



AIC : $\rho=0.75$



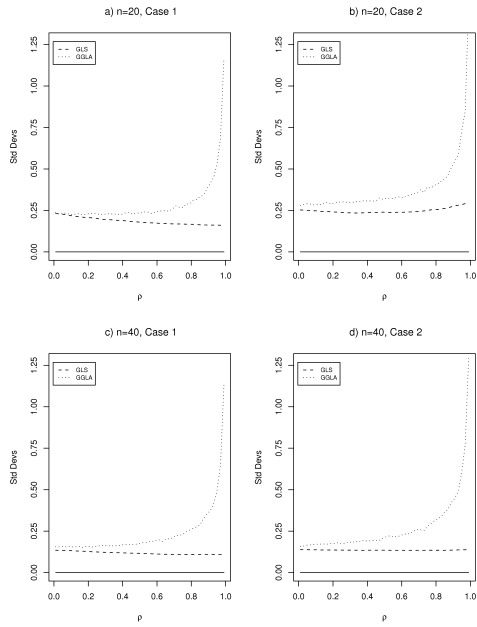
BIC : $\rho=0.25$



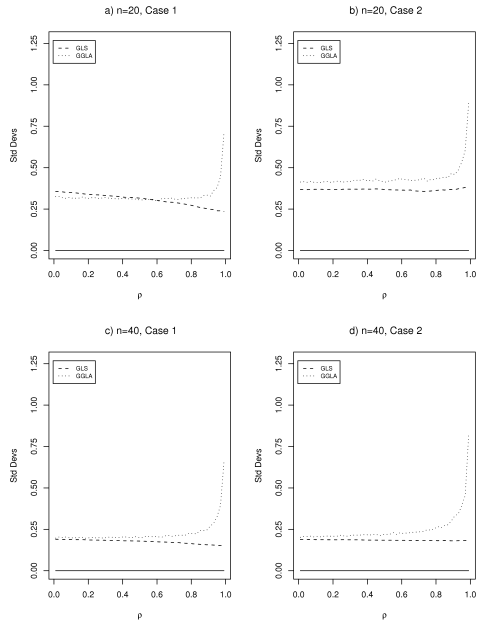
BIC : $\rho=0.75$

D.3.2 With Parsimony Correction

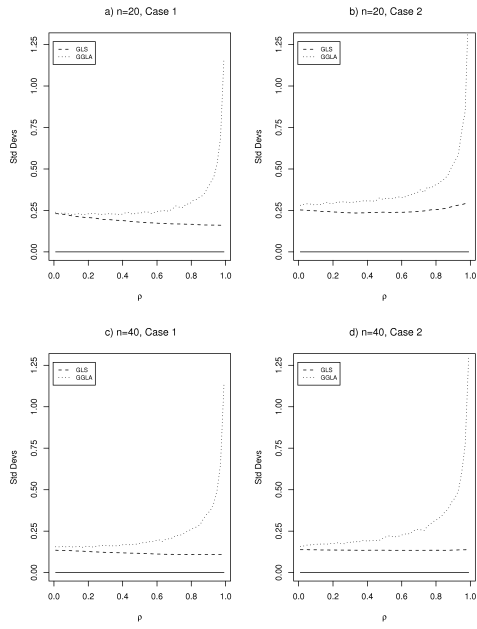
D.3.2.1 Cases 1 and 2



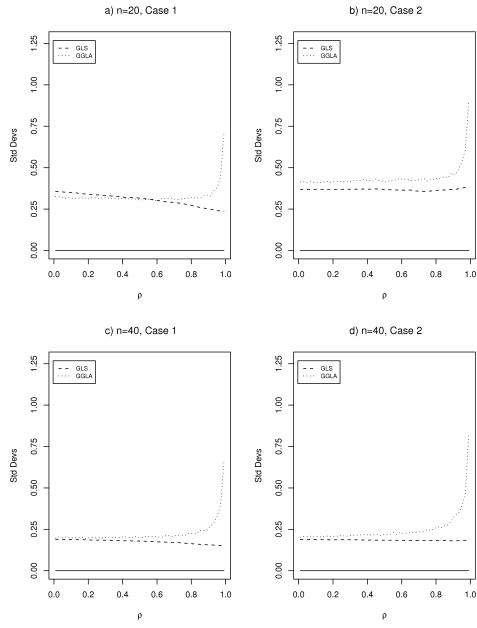
AIC : $\rho=0.25$



AIC : $\rho=0.75$

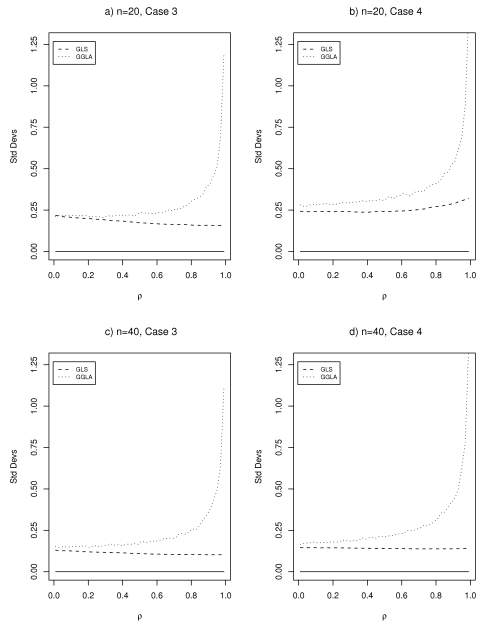


BIC : $\rho=0.25$

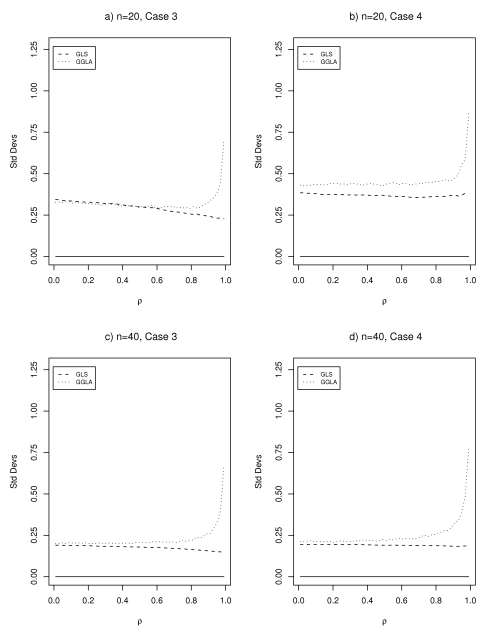


BIC : $\rho=0.75$

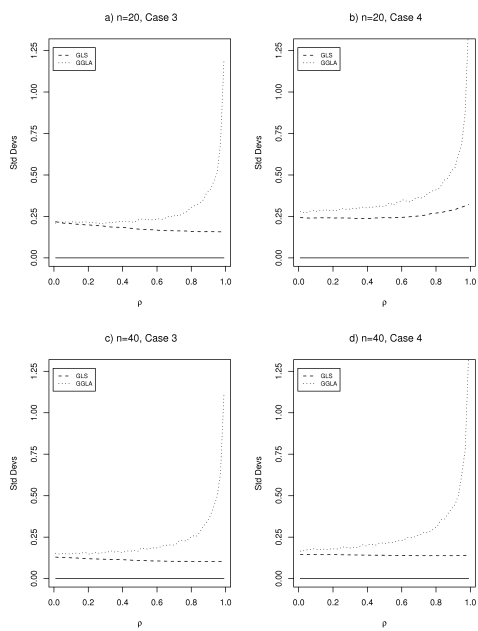
D.3.2.2 Cases 3 and 4



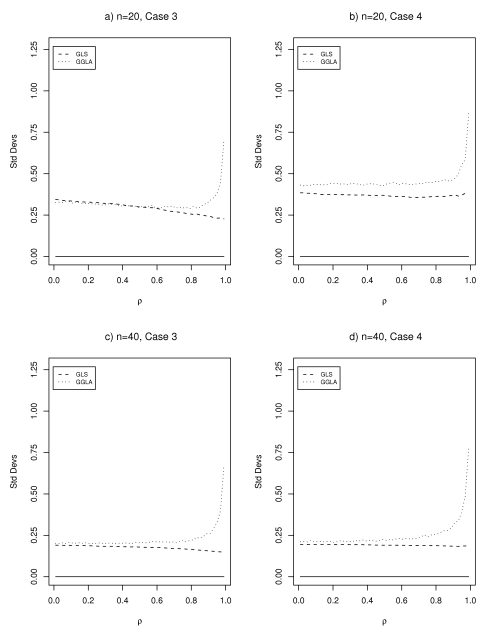
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

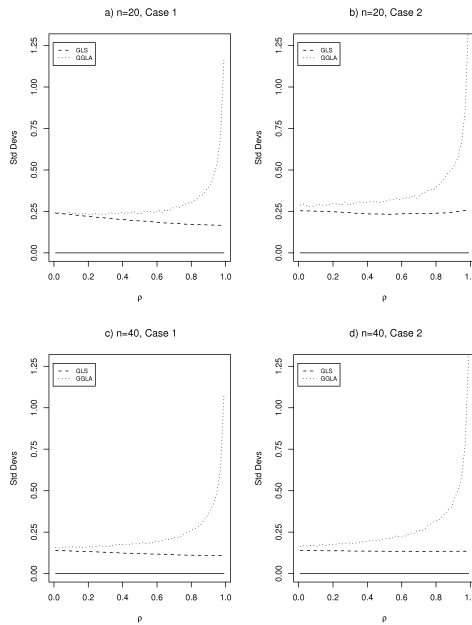


BIC : $\rho=0.75$

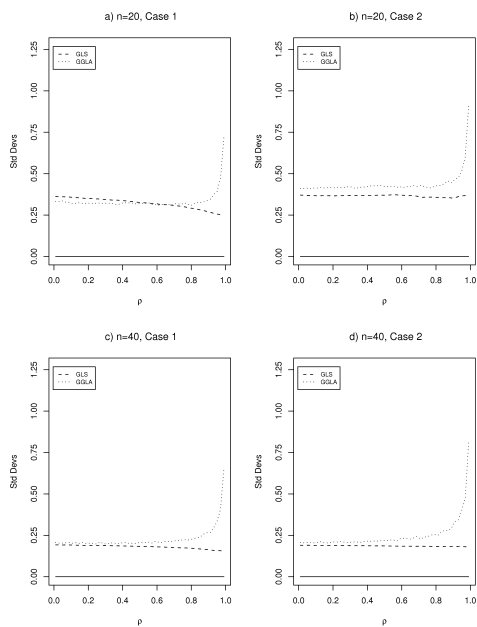
D.4 Simulation Standard Deviation of Fixed-Effect Estimates- Data-Driven Method

D.4.1 No Parsimony Correction

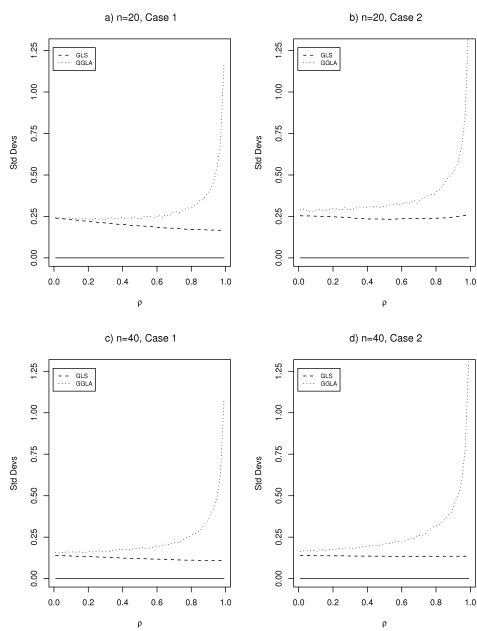
D.4.1.1 Cases 1 and 2



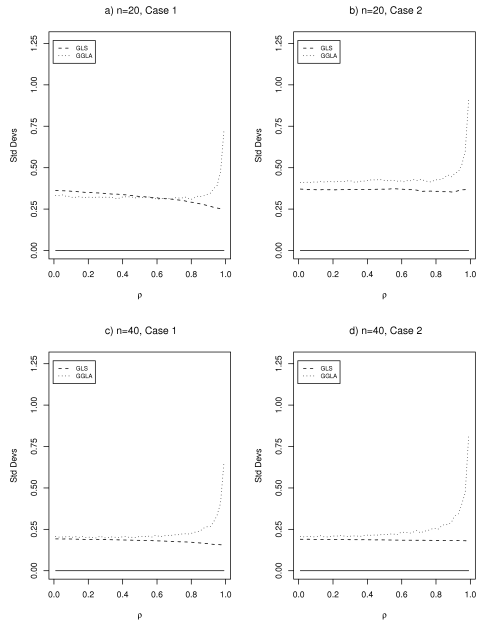
AIC : $\rho=0.25$



AIC : $\rho=0.75$

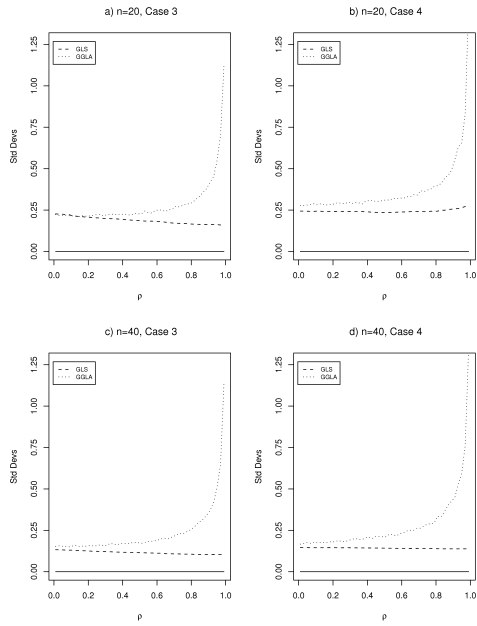


BIC : $\rho=0.25$

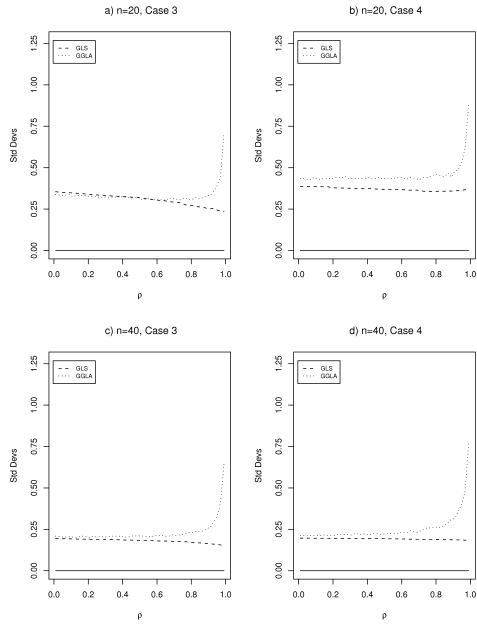


BIC : $\rho=0.75$

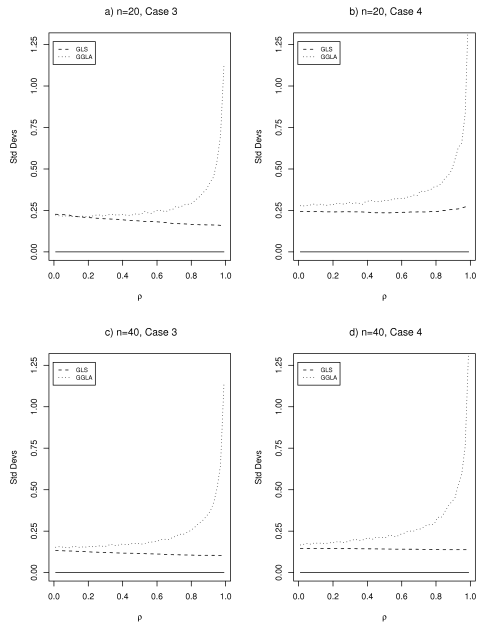
D.4.1.2 Cases 3 and 4



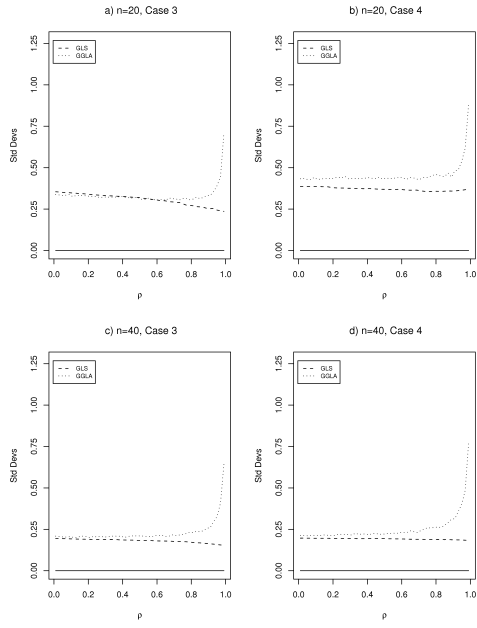
AIC : $\rho=0.25$



AIC : $\rho=0.75$



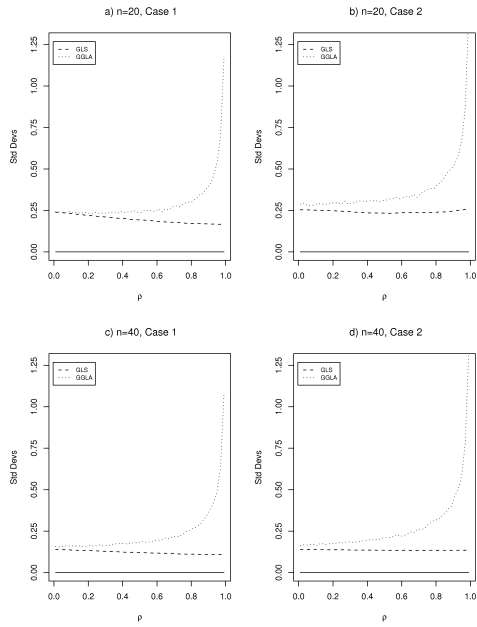
BIC : $\rho=0.25$



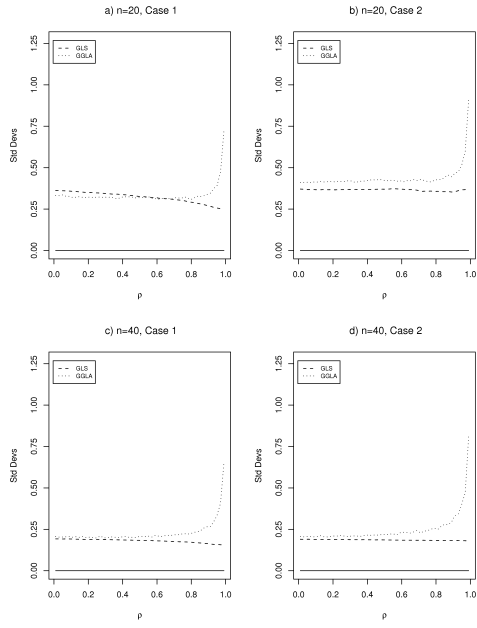
BIC : $\rho=0.75$

D.4.2 With Parsimony Correction

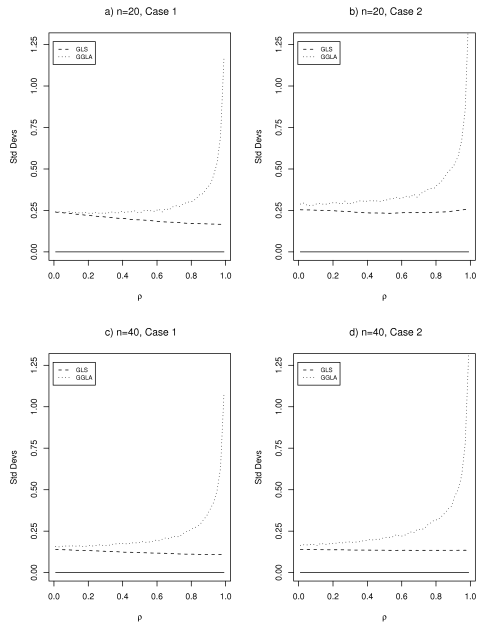
D.4.2.1 Cases 1 and 2



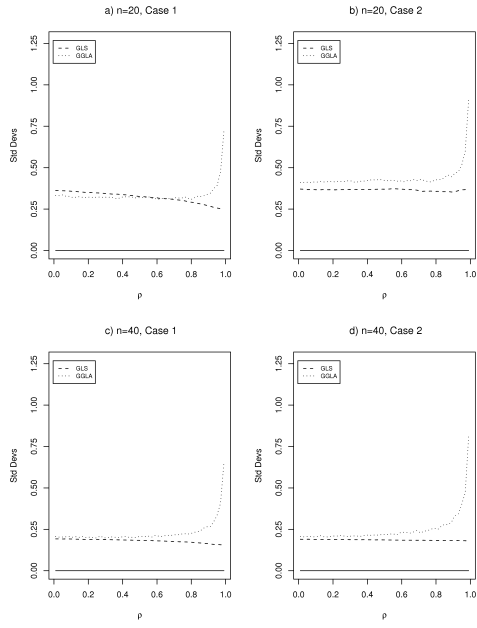
AIC : $\rho=0.25$



AIC : $\rho=0.75$

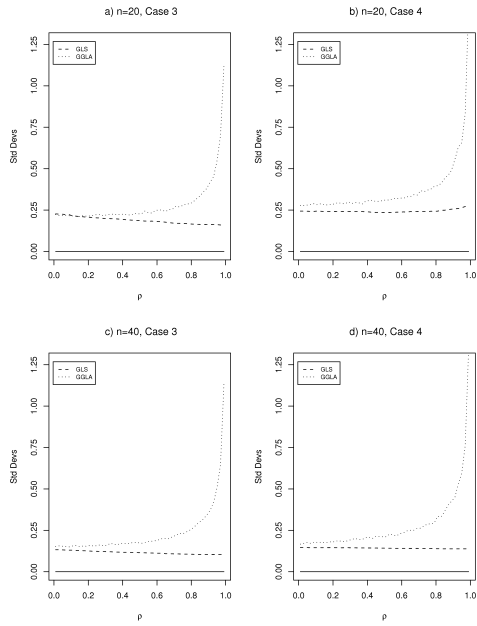


BIC : $\rho=0.25$

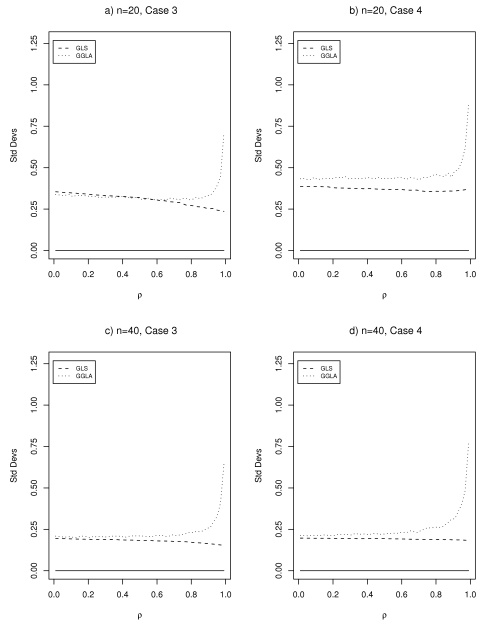


BIC : $\rho=0.75$

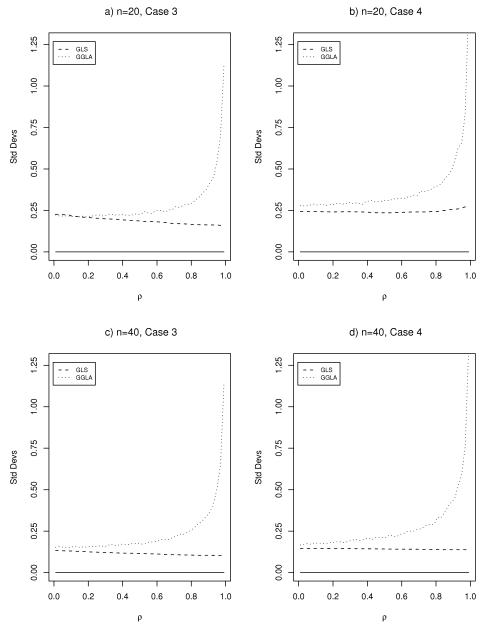
D.4.2.2 Cases 3 and 4



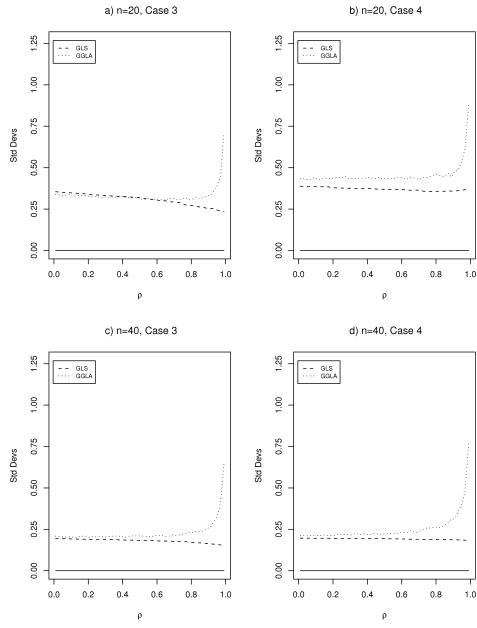
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

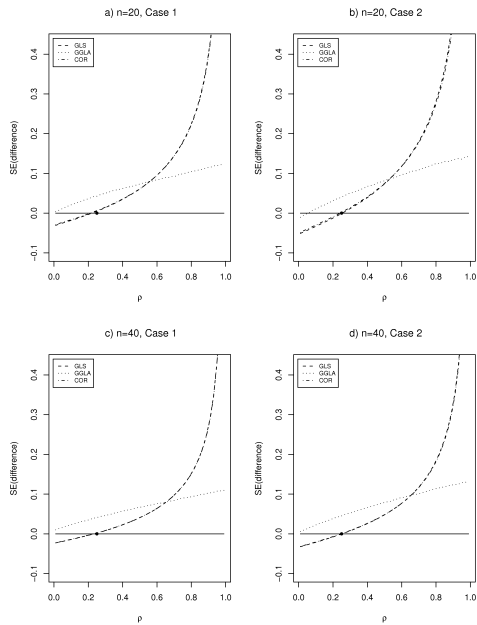


BIC : $\rho=0.75$

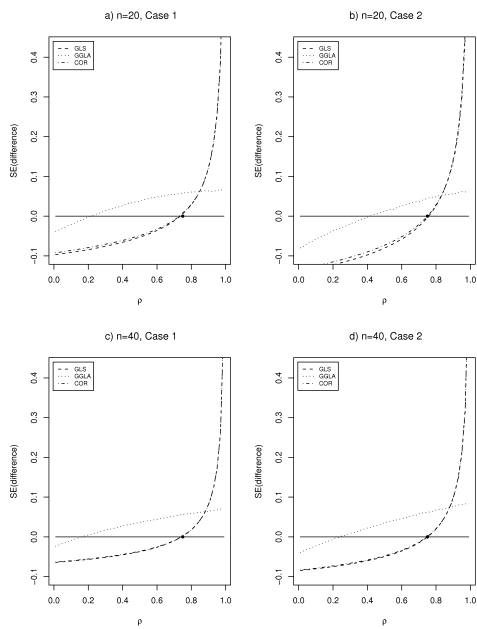
D.5 Standard Error Estimation - Proof-of-Concept

D.5.1 No Parsimony Correction

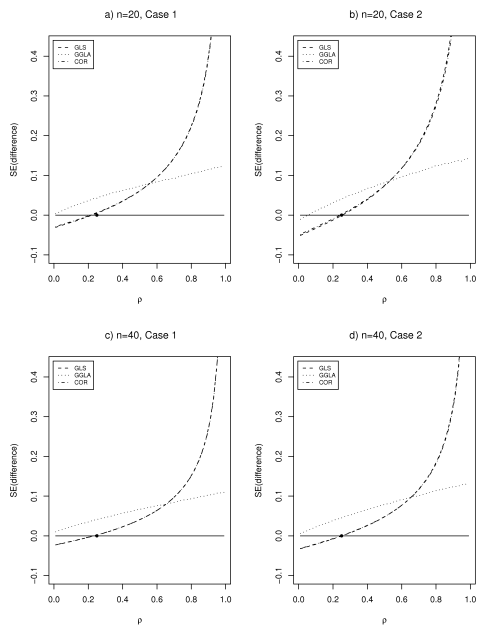
D.5.1.1 Cases 1 and 2



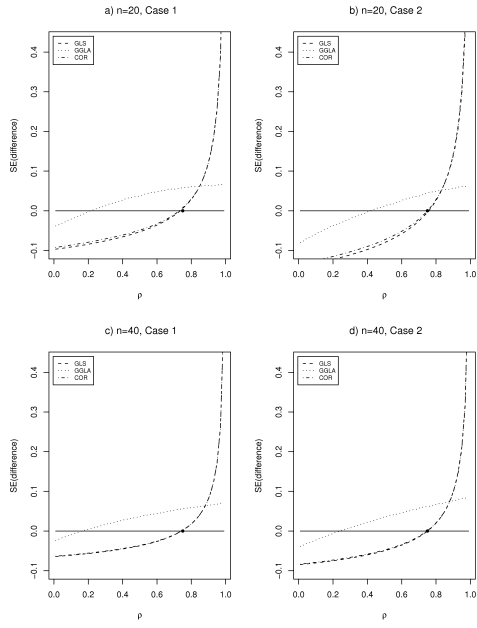
AIC : $\rho=0.25$



AIC : $\rho=0.75$

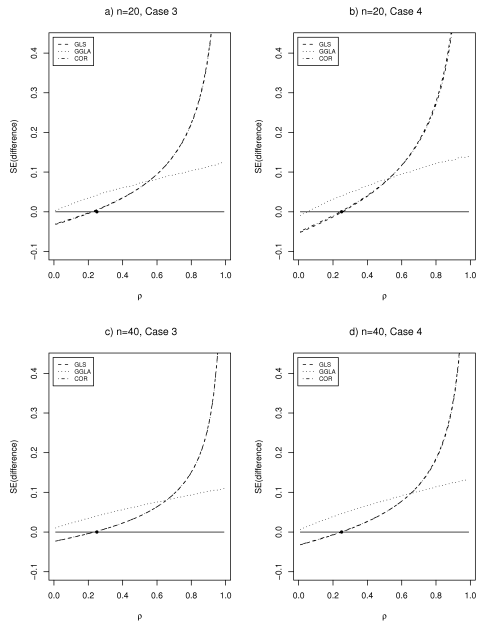


BIC : $\rho=0.25$

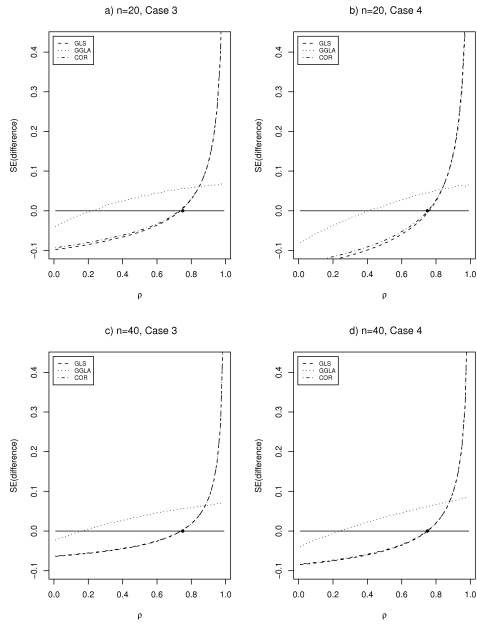


BIC : $\rho=0.75$

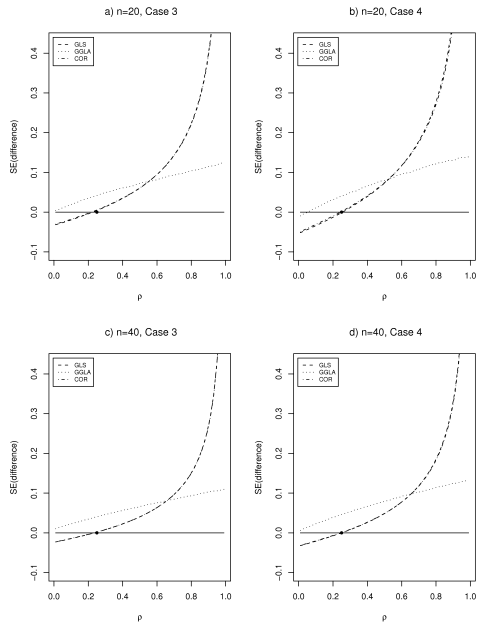
D.5.1.2 Cases 3 and 4



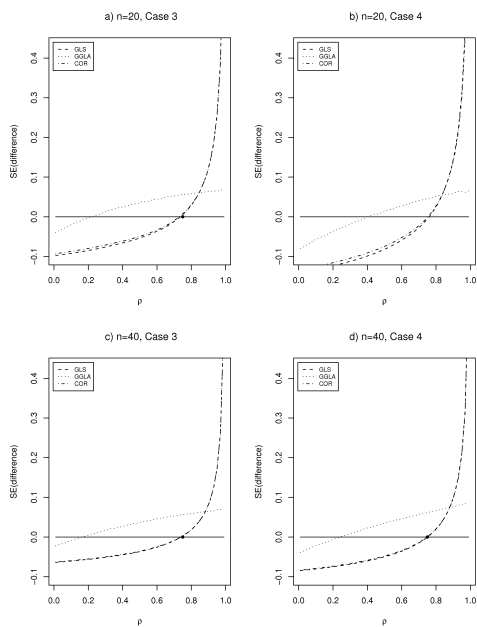
AIC : $\rho=0.25$



AIC : $\rho=0.75$



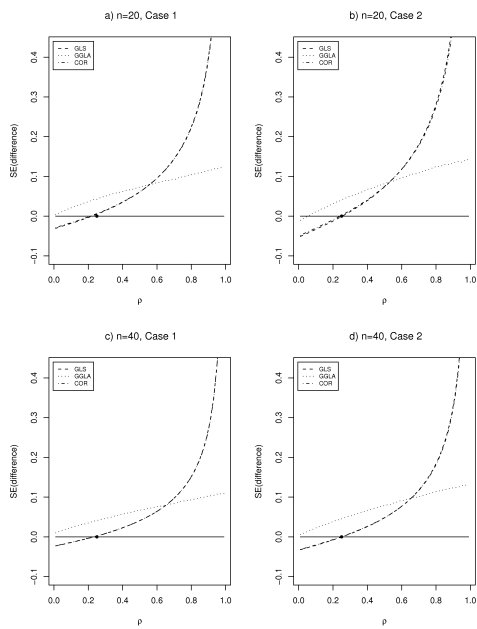
BIC : $\rho=0.25$



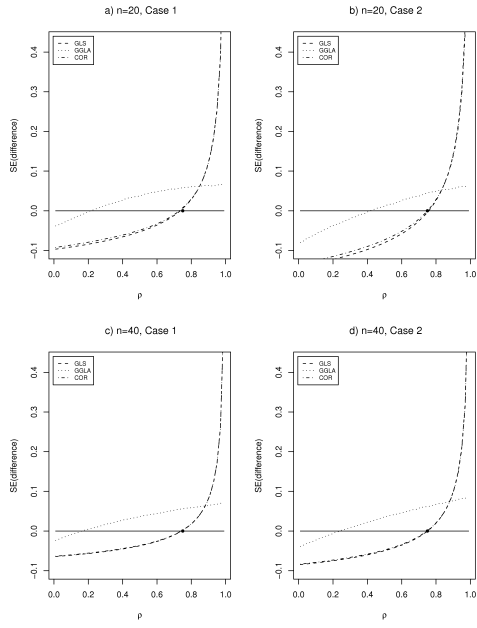
BIC : $\rho=0.75$

D.5.2 With Parsimony Correction

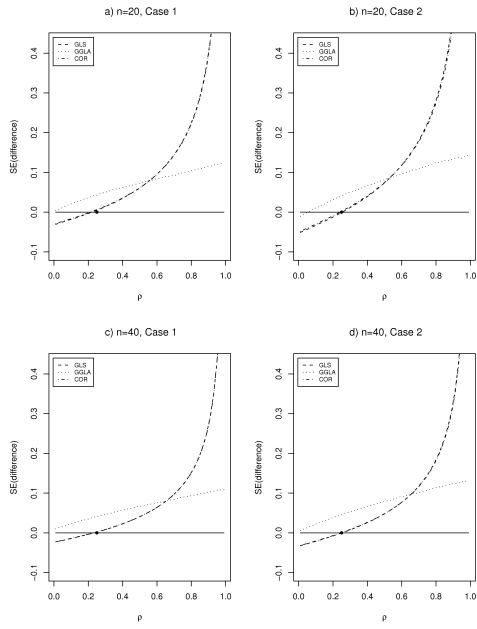
D.5.2.1 Cases 1 and 2



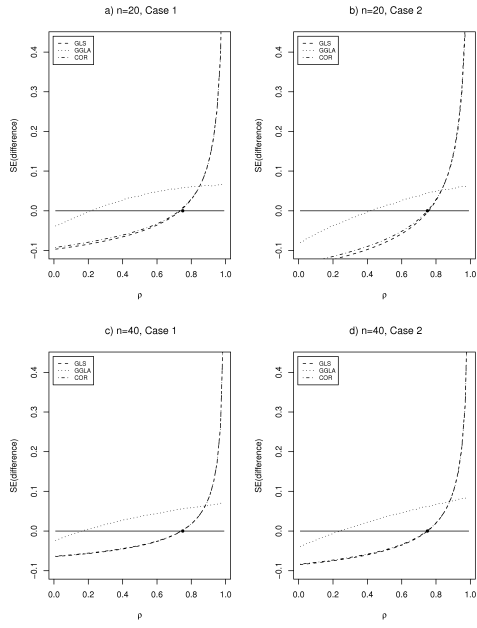
AIC : $\rho=0.25$



AIC : $\rho=0.75$

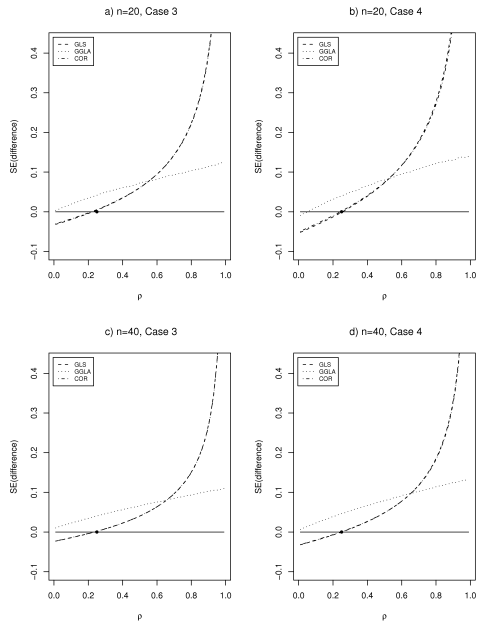


BIC : $\rho=0.25$

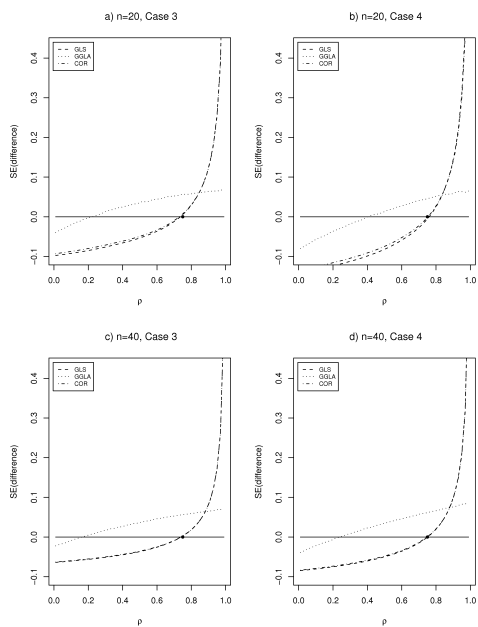


BIC : $\rho=0.75$

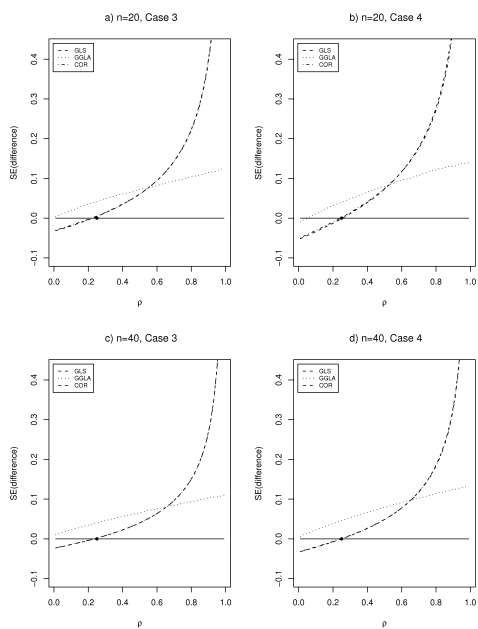
D.5.2.2 Cases 3 and 4



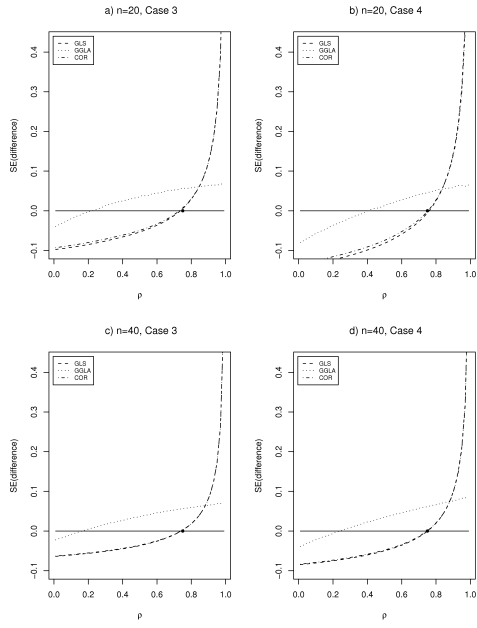
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

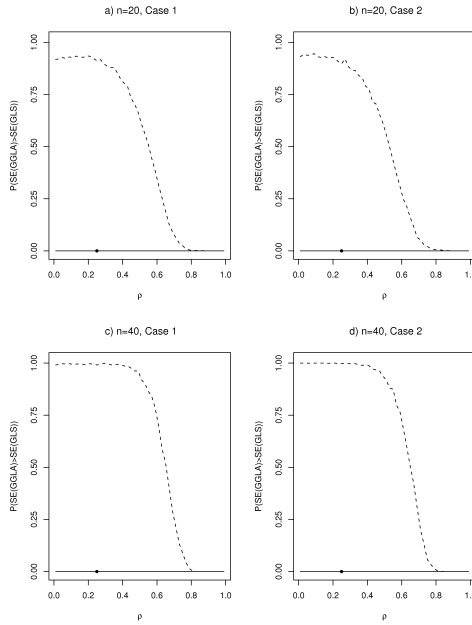


BIC : $\rho=0.75$

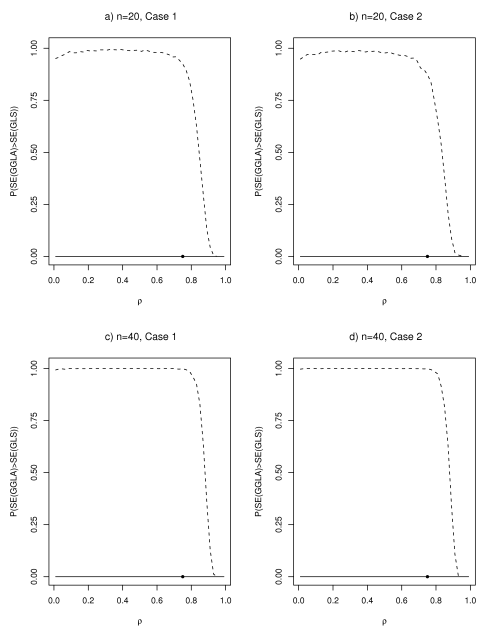
D.6 Relationship between SEs from GGLA and SEs from GLS – Proof-of-Concept

D.6.1 No parsimony correction

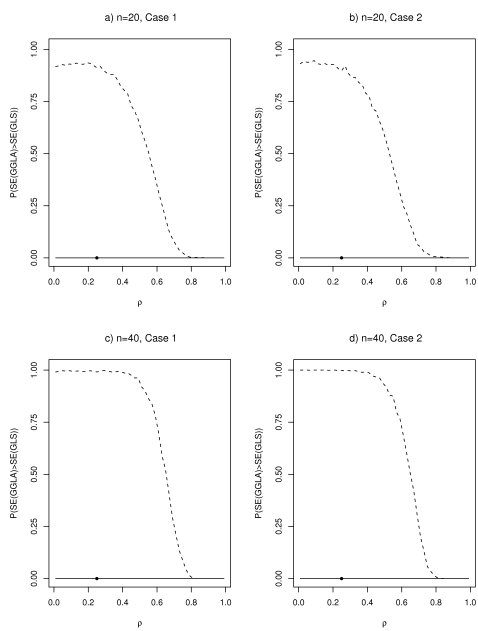
D.6.1.1 Cases 1 and 2



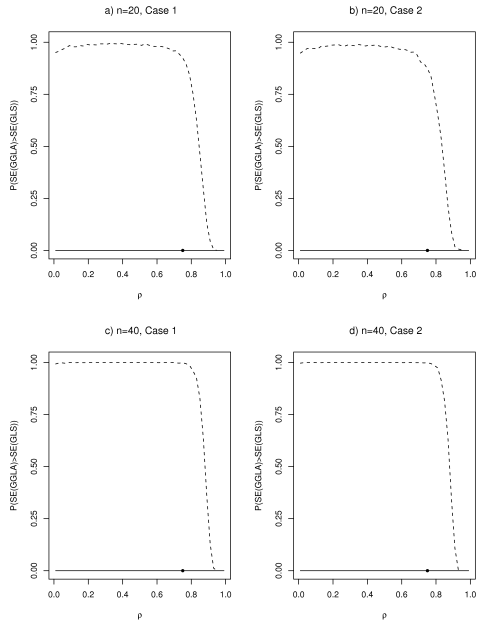
AIC : $\rho=0.25$



AIC : $\rho=0.75$

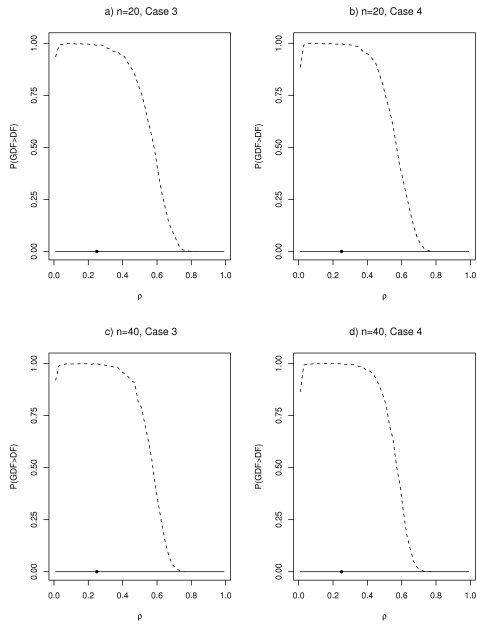


BIC : $\rho=0.25$

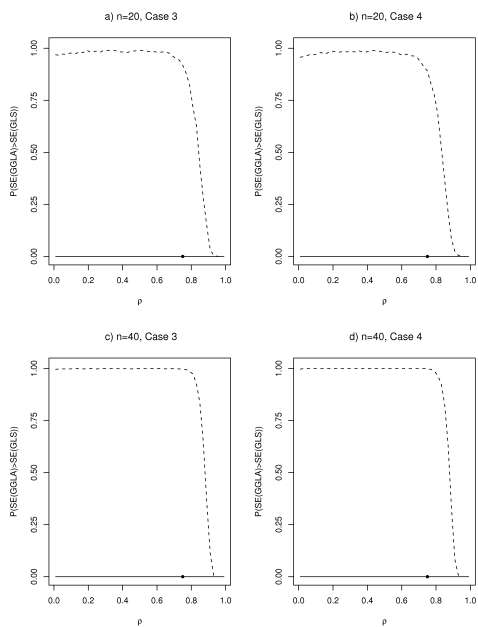


BIC : $\rho=0.75$

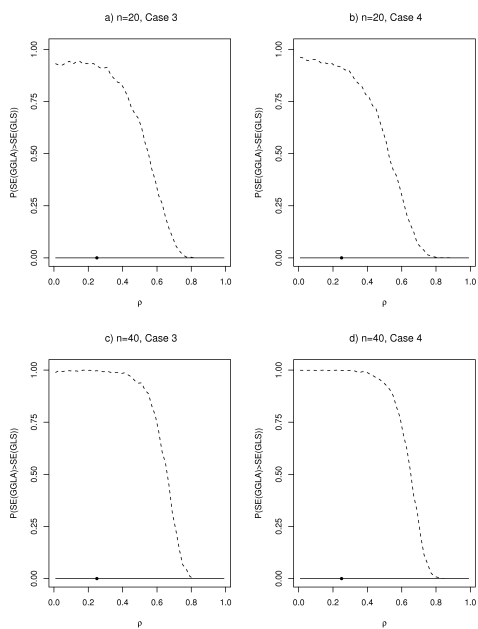
D.6.1.2 Cases 3 and 4



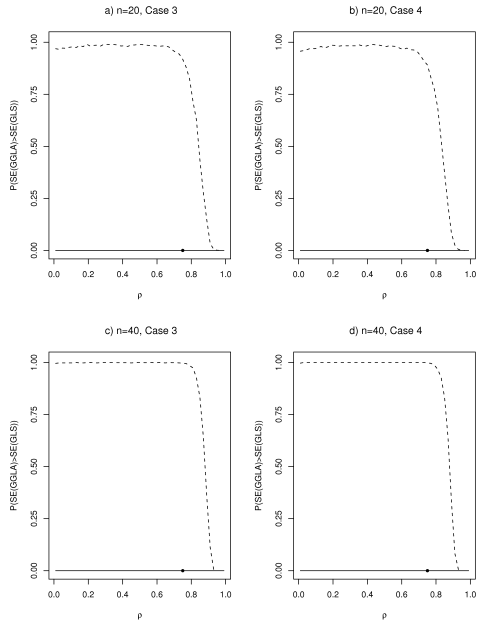
AIC : $\rho=0.25$



AIC : $\rho=0.75$



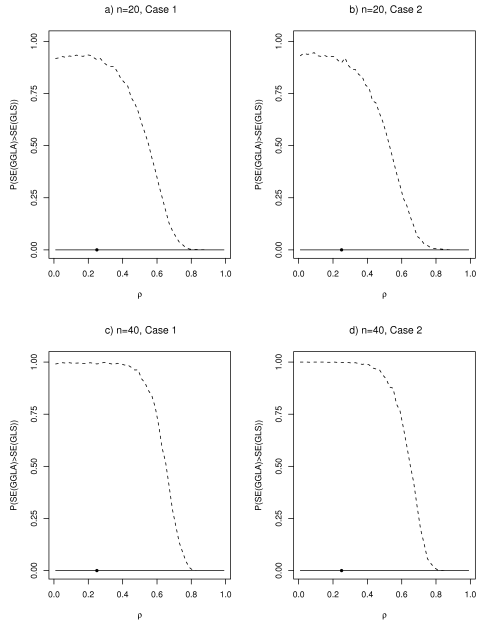
BIC : $\rho=0.25$



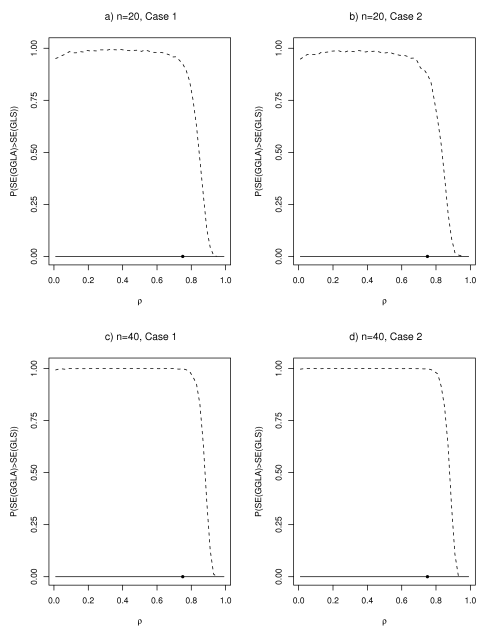
BIC : $\rho=0.75$

D.6.2 With Parsimony Correction

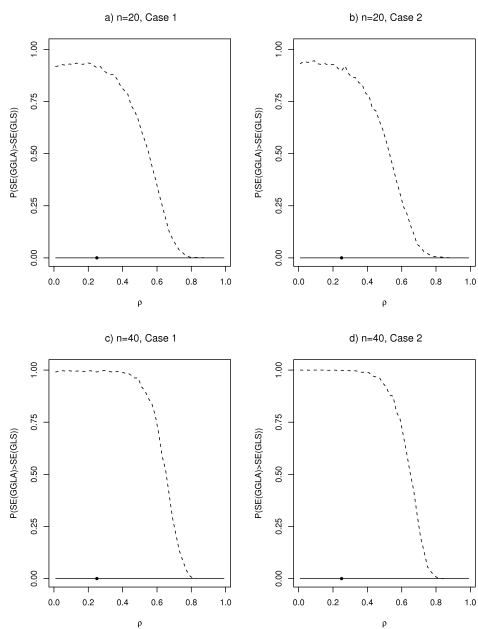
D.6.2.1 Cases 1 and 2



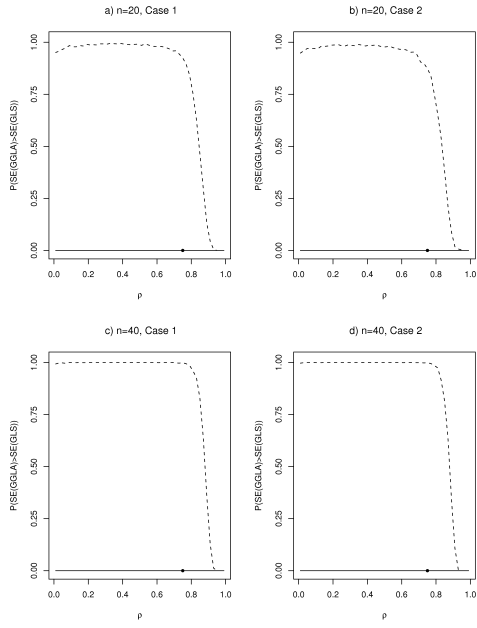
AIC : $\rho=0.25$



AIC : $\rho=0.75$

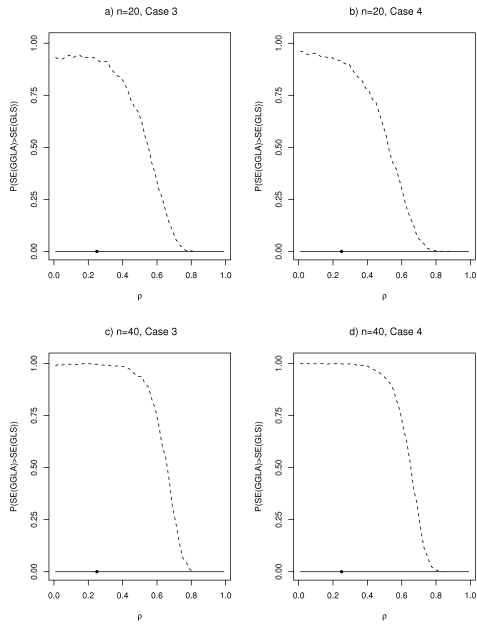


BIC : $\rho=0.25$

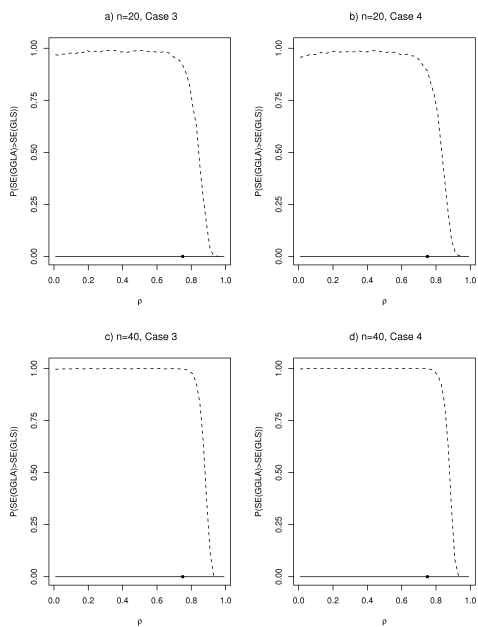


BIC : $\rho=0.75$

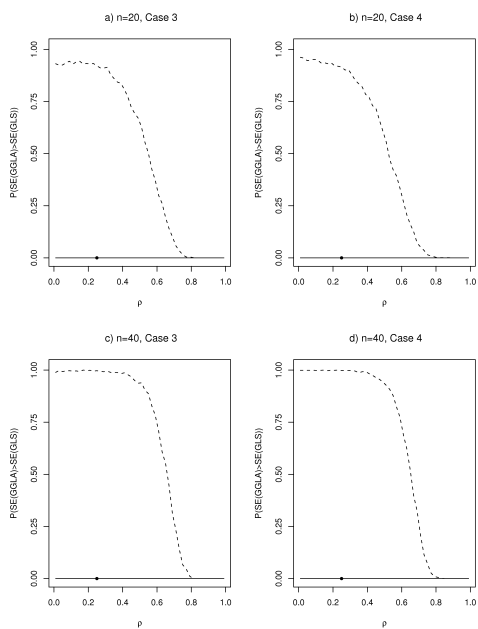
D.6.2.2 Cases 3 and 4



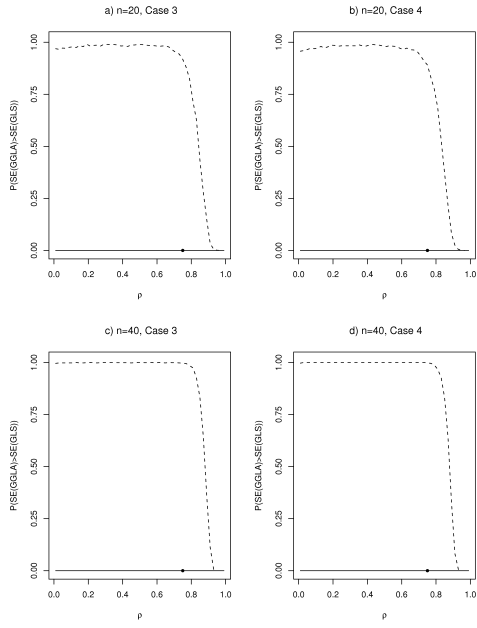
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

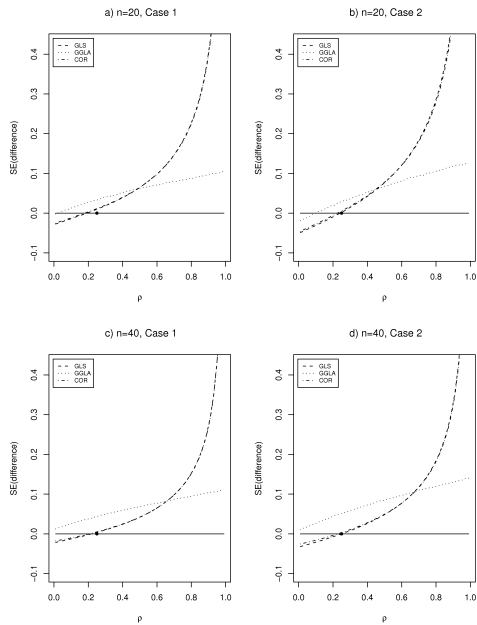


BIC : $\rho=0.75$

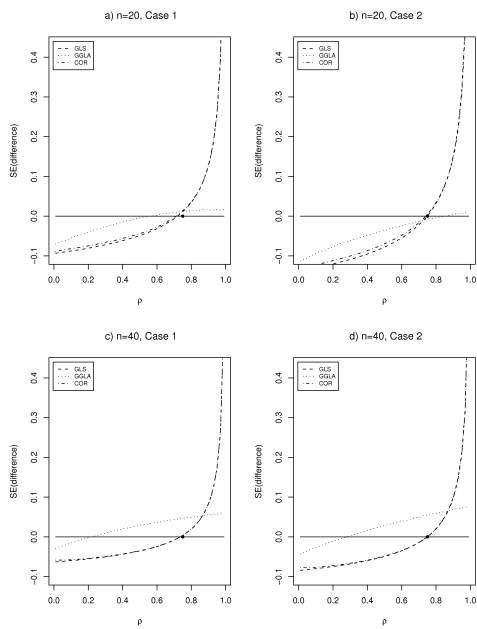
D.7 Standard Error Estimation - Data-Driven Method

D.7.1 No Parsimony Correction

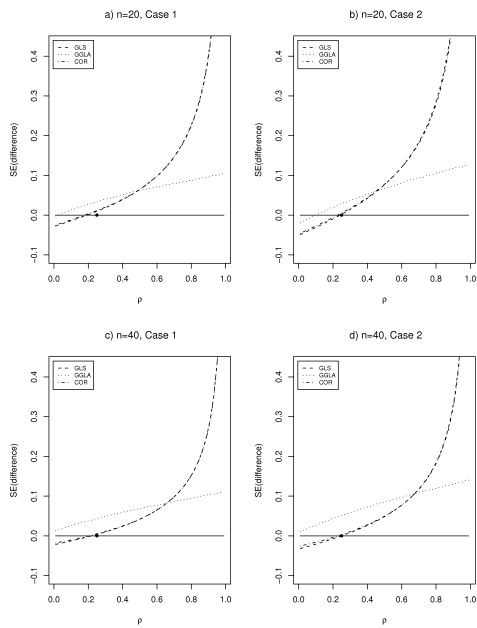
D.7.1.1 Cases 1 and 2



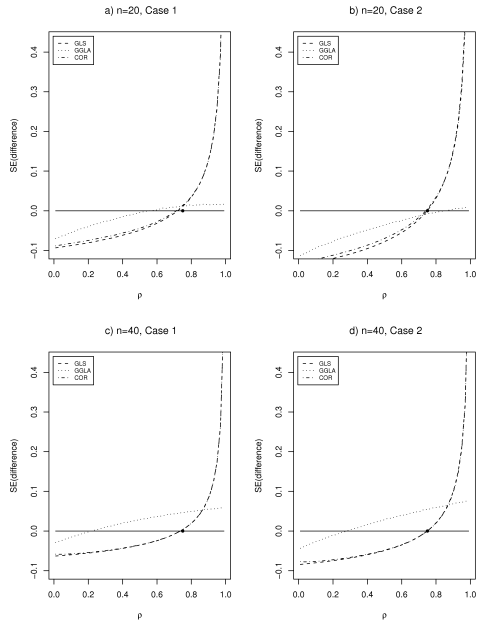
AIC : $\rho=0.25$



AIC : $\rho=0.75$

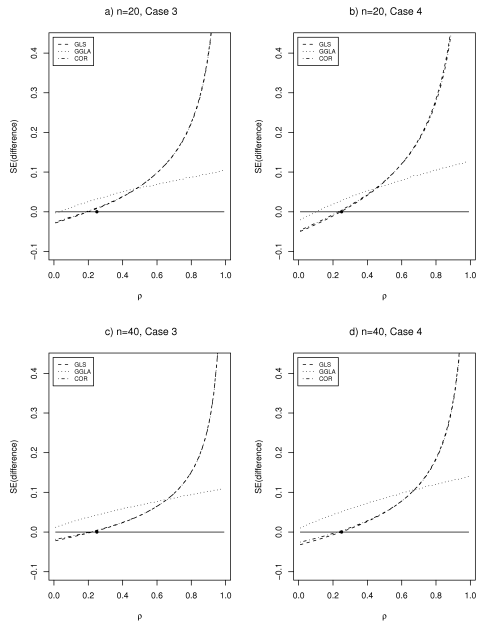


BIC : $\rho=0.25$

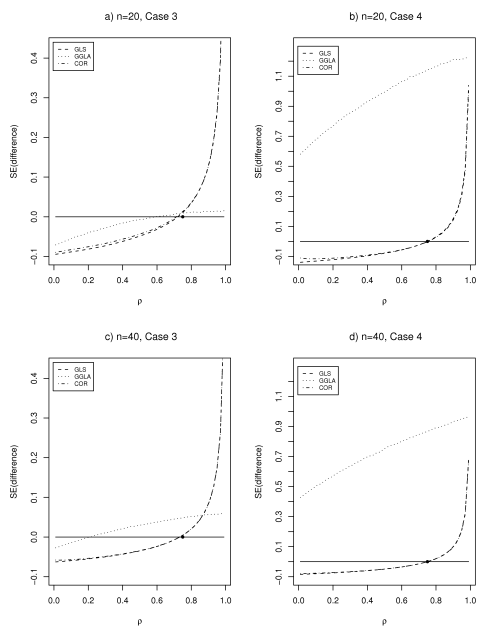


BIC : $\rho=0.75$

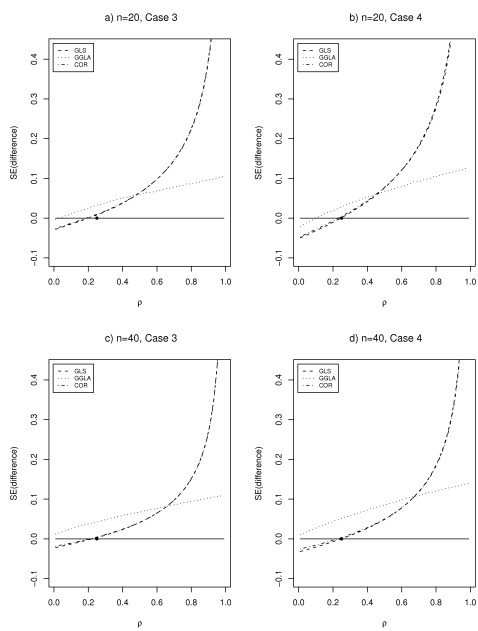
D.7.1.2 Cases 3 and 4



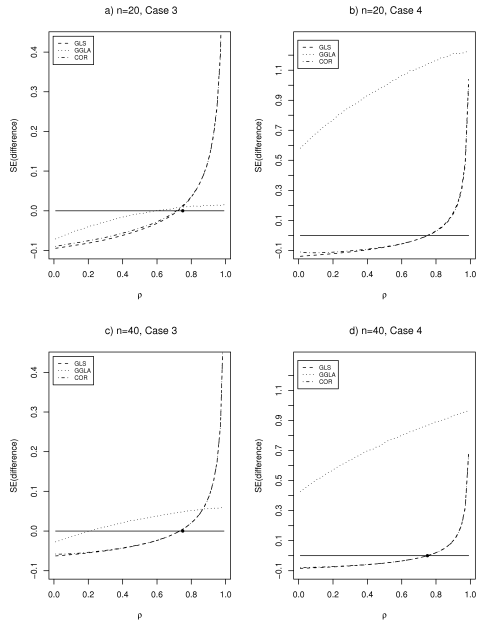
AIC : $\rho=0.25$



AIC : $\rho=0.75$



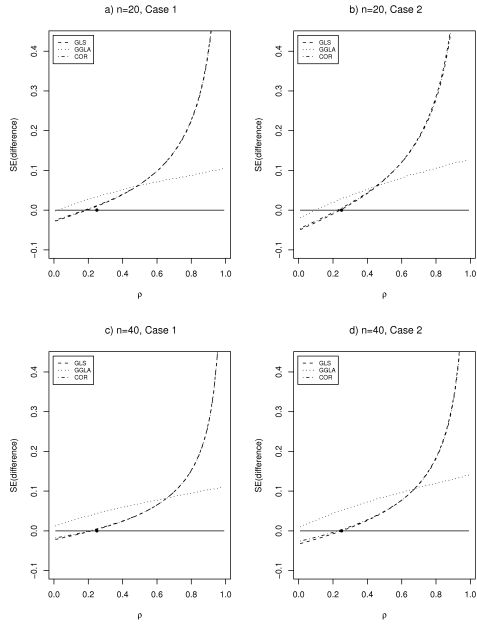
BIC : $\rho=0.25$



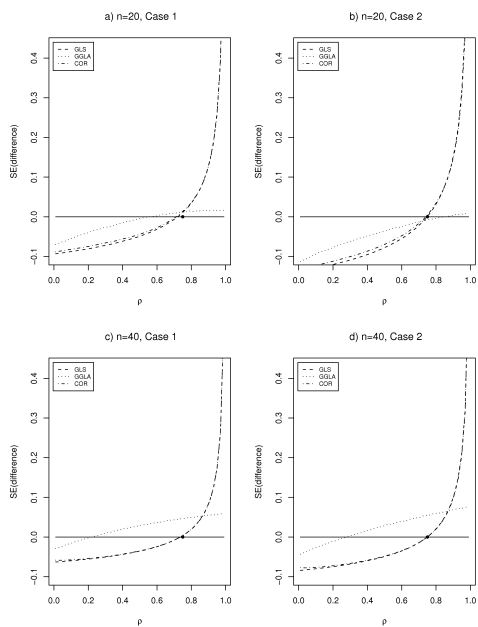
BIC : $\rho=0.75$

D.7.2 With Parsimony Correction

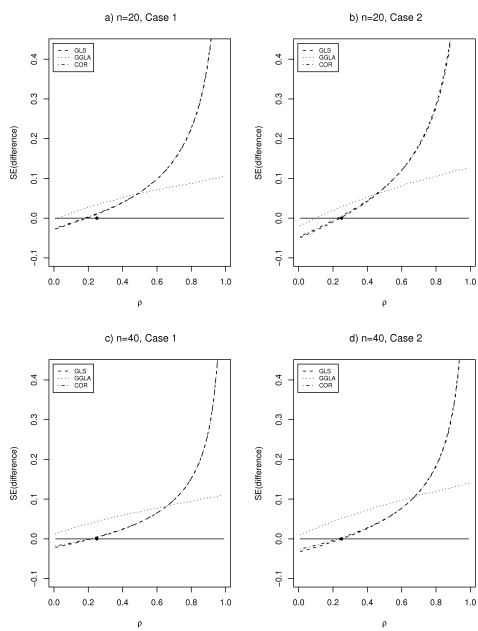
D.7.2.1 Cases 1 and 2



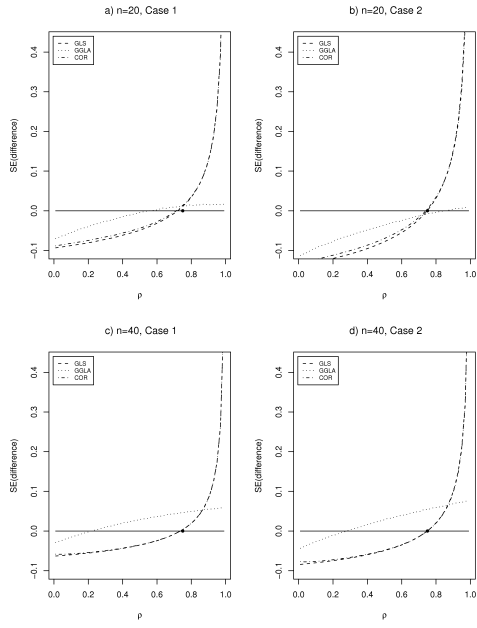
AIC : $\rho=0.25$



AIC : $\rho=0.75$

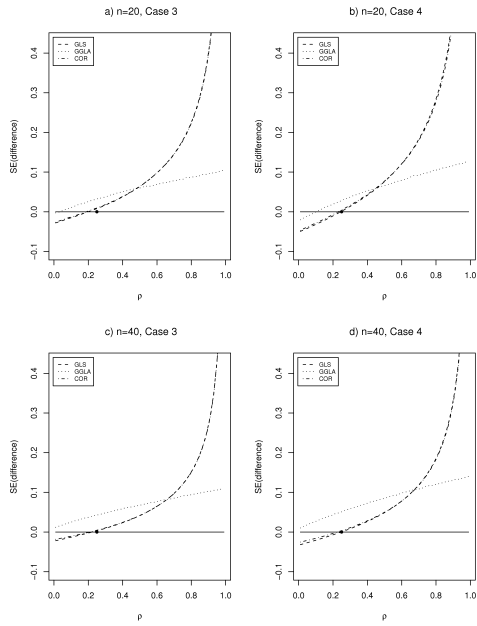


BIC : $\rho=0.25$

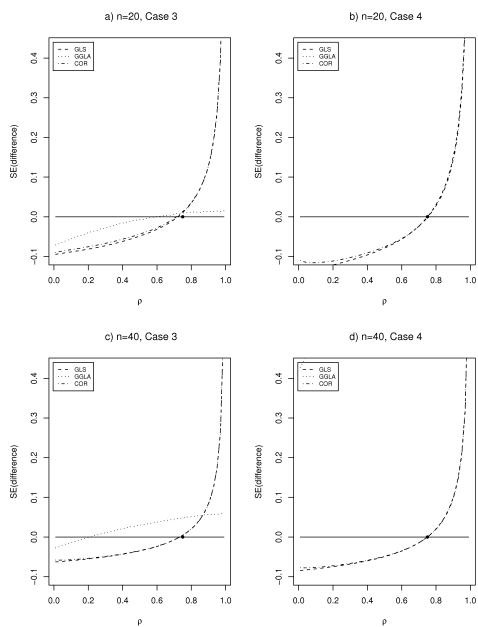


BIC : $\rho=0.75$

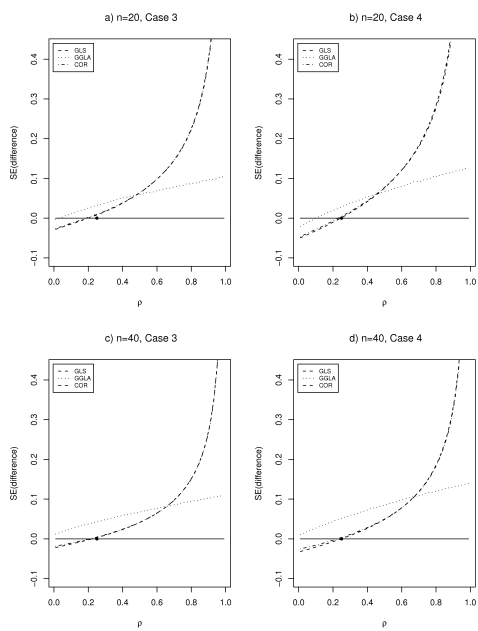
D.7.2.2 Cases 3 and 4



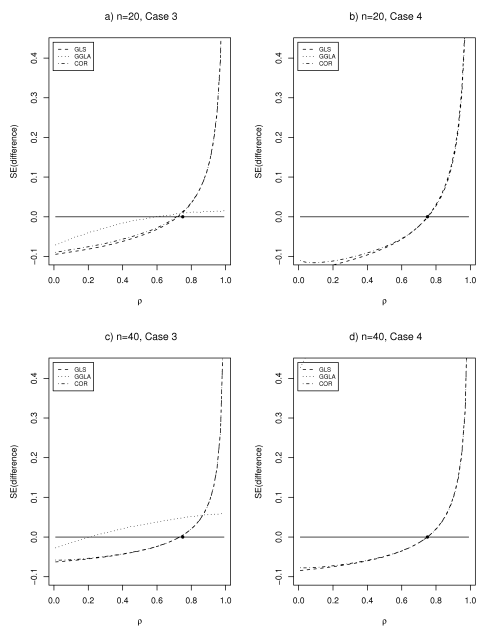
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

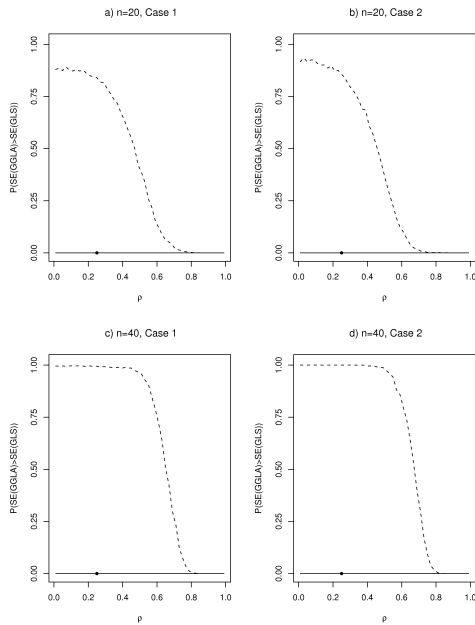


BIC : $\rho=0.75$

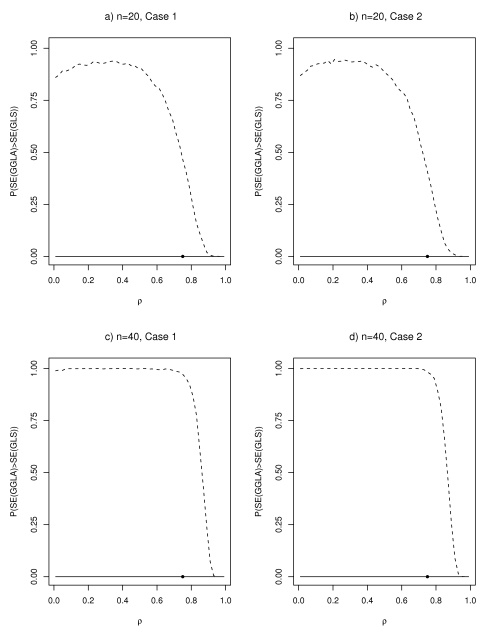
D.8 Relationship between SEs from GGLA and SEs from GLS – Data-Driven

D.8.1 No parsimony correction

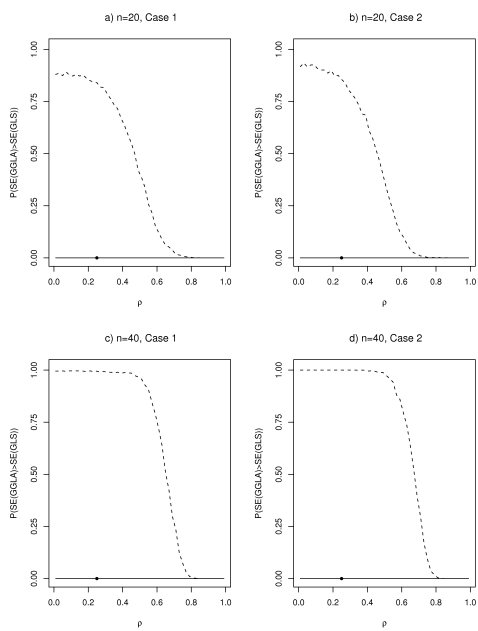
D.8.1.1 Cases 1 and 2



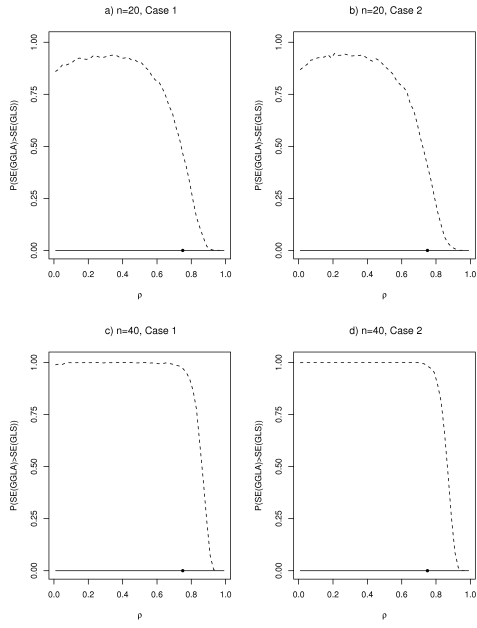
AIC : $\rho=0.25$



AIC : $\rho=0.75$

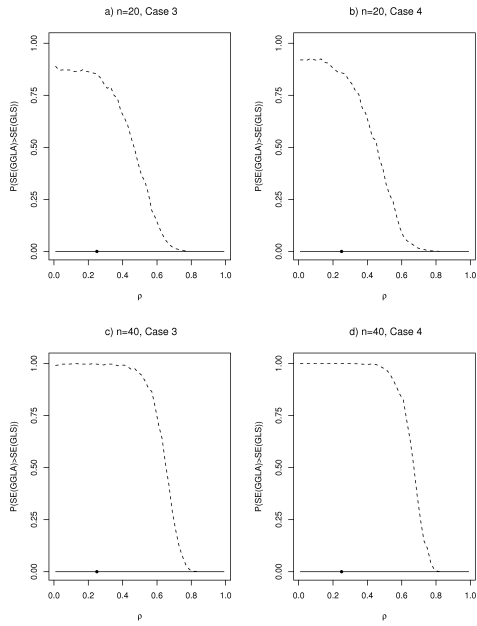


BIC : $\rho=0.25$

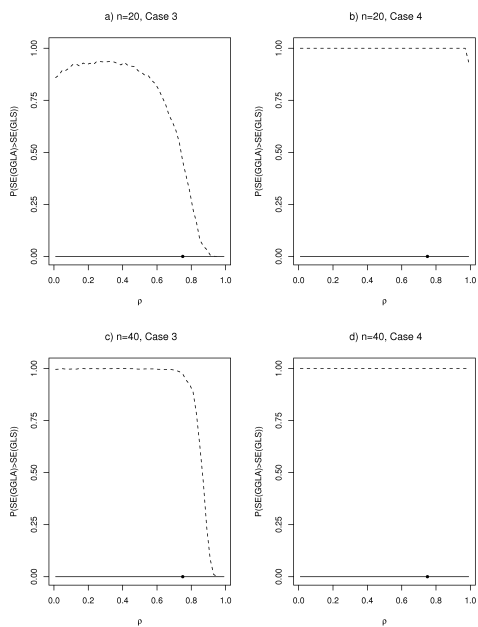


BIC : $\rho=0.75$

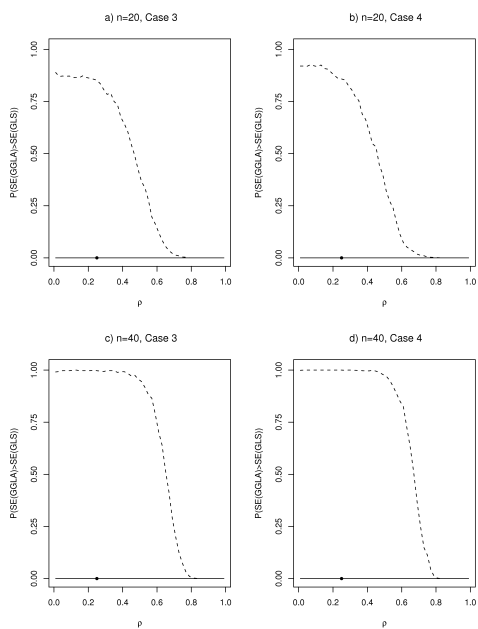
D.8.1.2 Cases 3 and 4



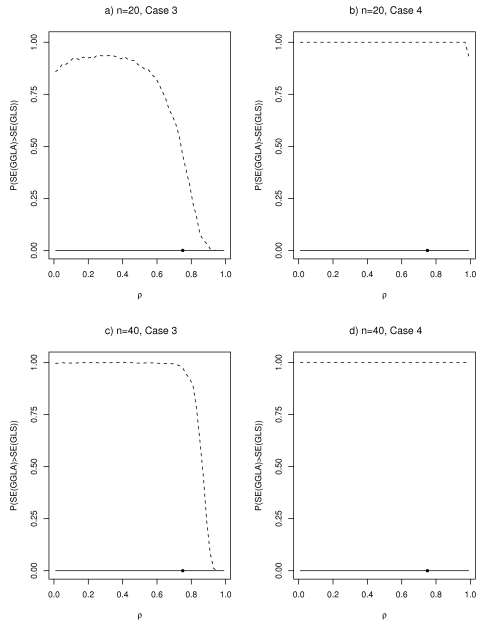
AIC : $\rho=0.25$



AIC : $\rho=0.75$



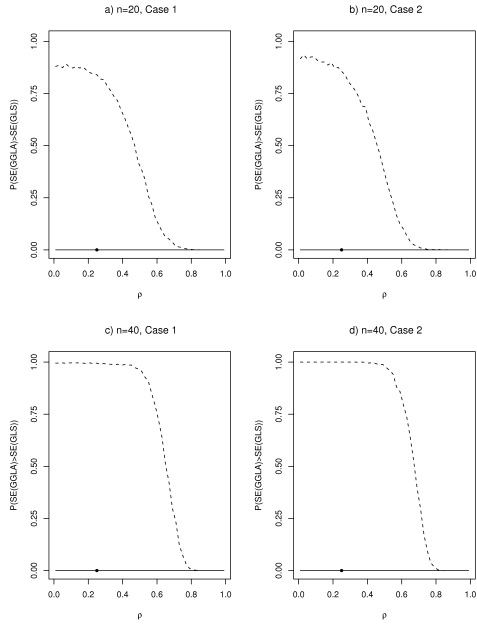
BIC : $\rho=0.25$



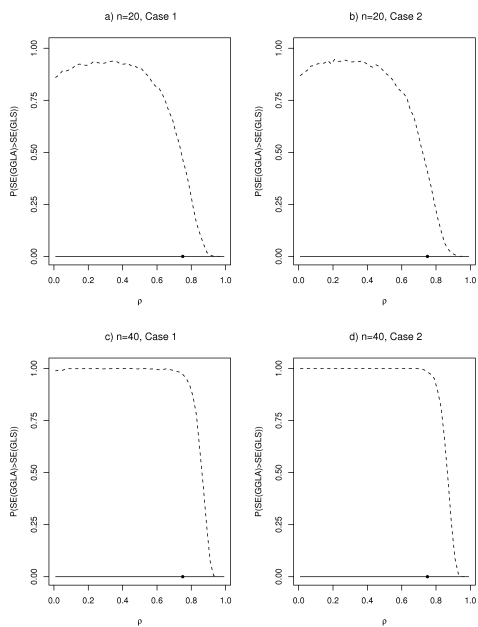
BIC : $\rho=0.75$

D.8.2 With Parsimony Correction

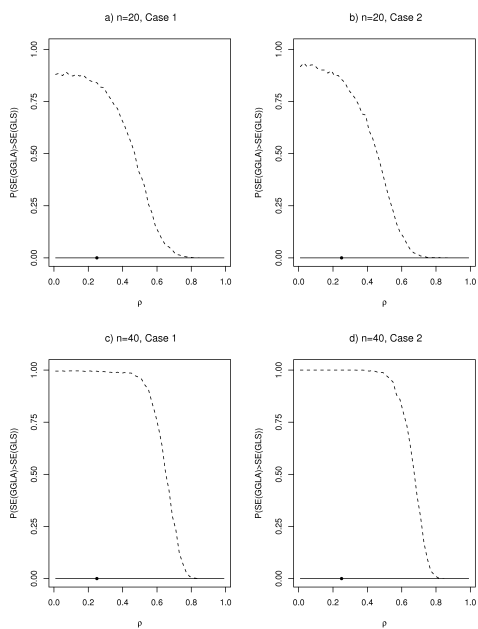
D.8.2.1 Cases 1 and 2



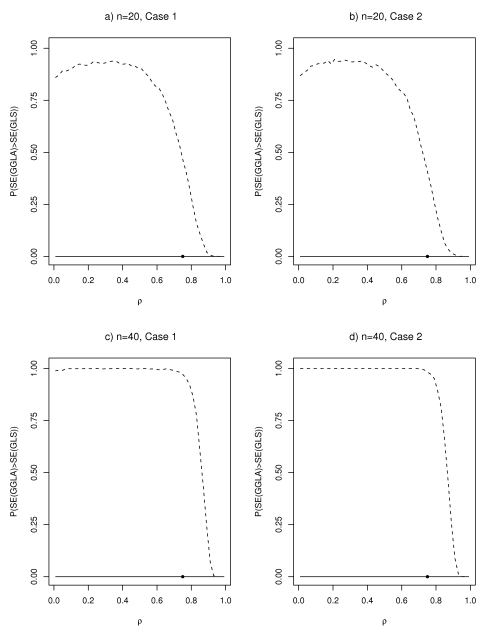
AIC : $\rho=0.25$



AIC : $\rho=0.75$

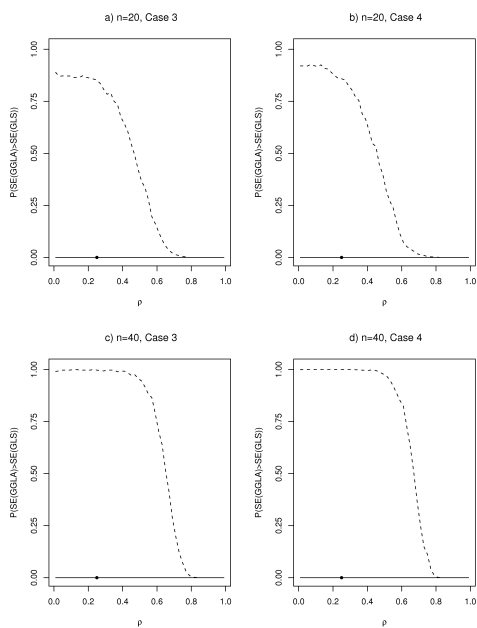


BIC : $\rho=0.25$

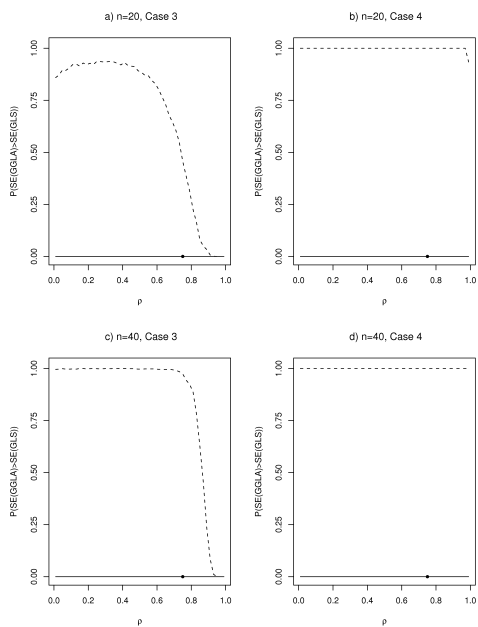


BIC : $\rho=0.75$

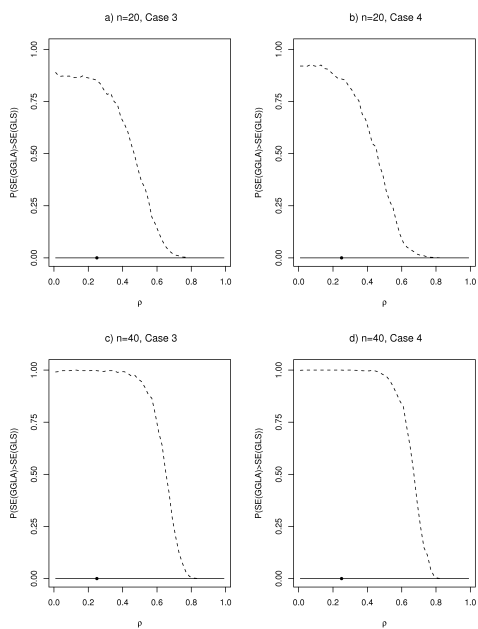
D.8.2.2 Cases 3 and 4



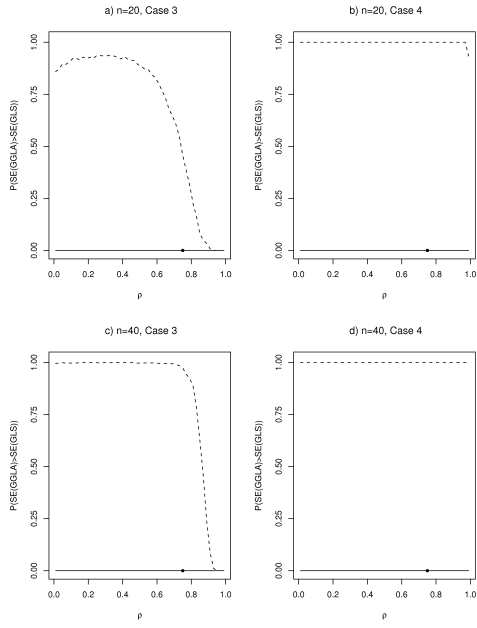
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

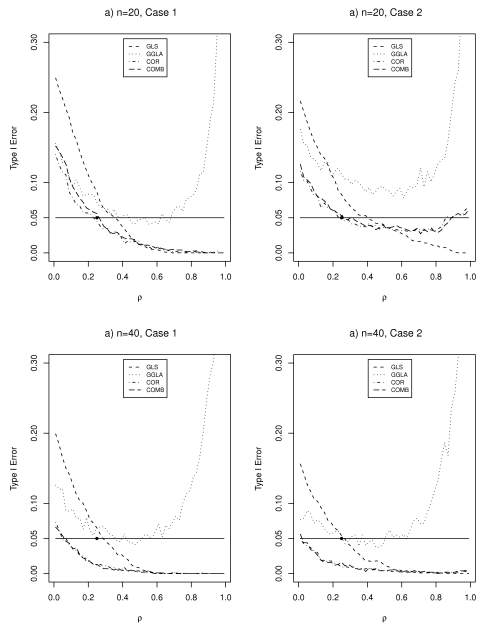


BIC : $\rho=0.75$

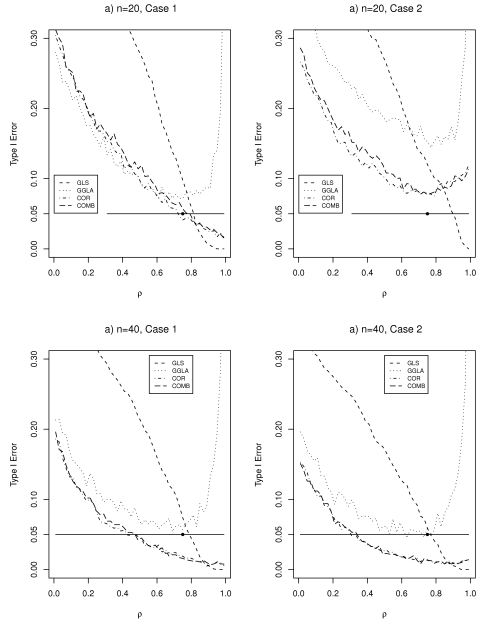
D.9 Type-I Error and Coverage Probability - Proof-of-Concept

D.9.1 No Parsimony Correction

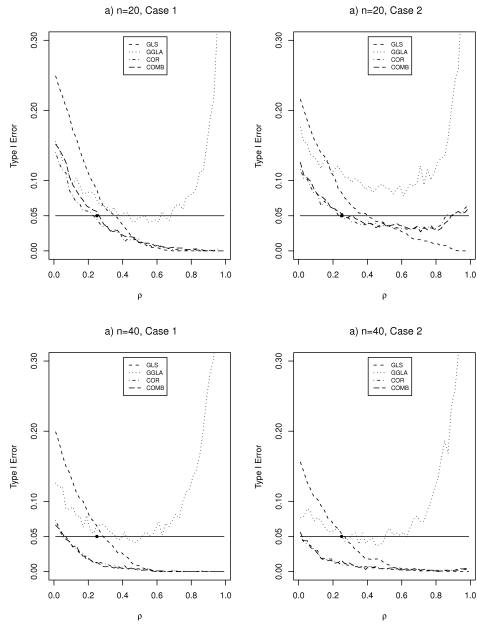
D.9.1.1 Cases 1 and 2



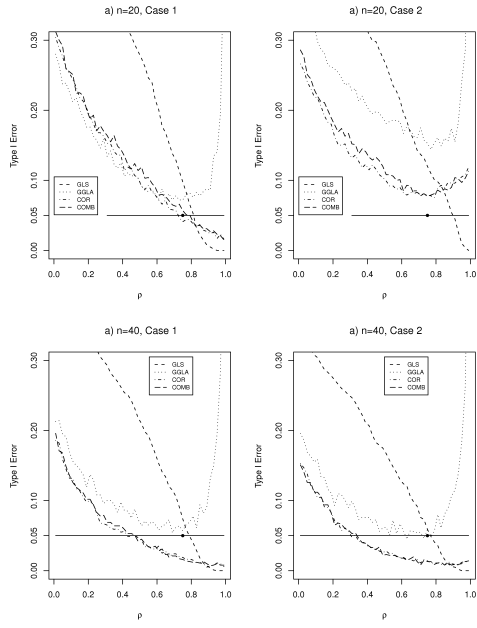
AIC : $\rho=0.25$



AIC : $\rho=0.75$

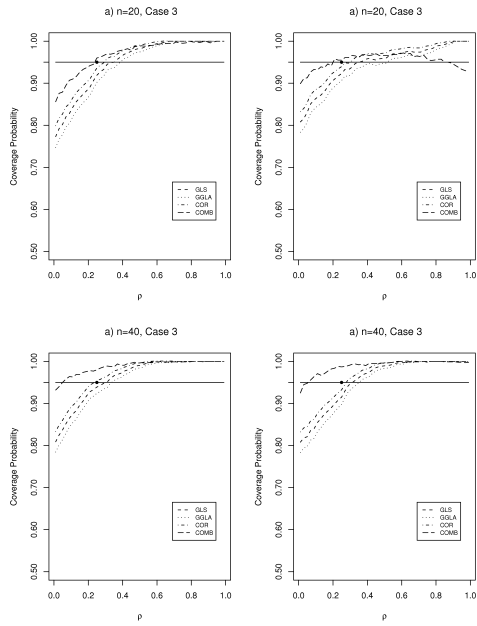


BIC : $\rho=0.25$

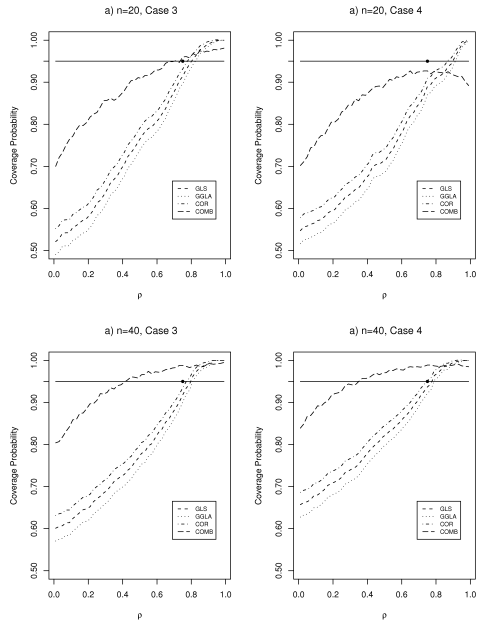


BIC : $\rho=0.75$

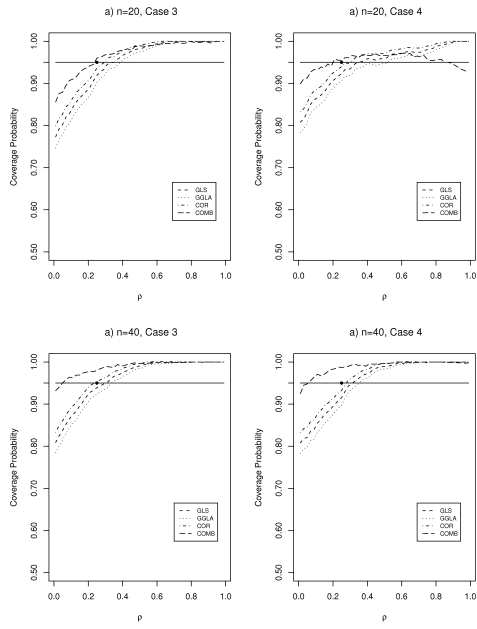
D.9.1.2 Cases 3 and 4



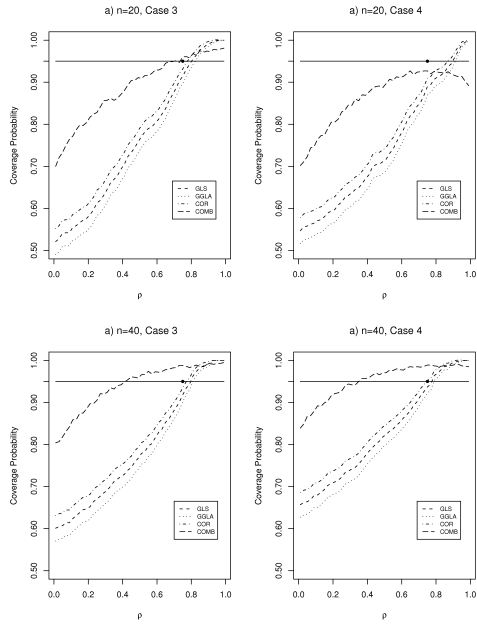
AIC : $\rho=0.25$



AIC : $\rho=0.75$



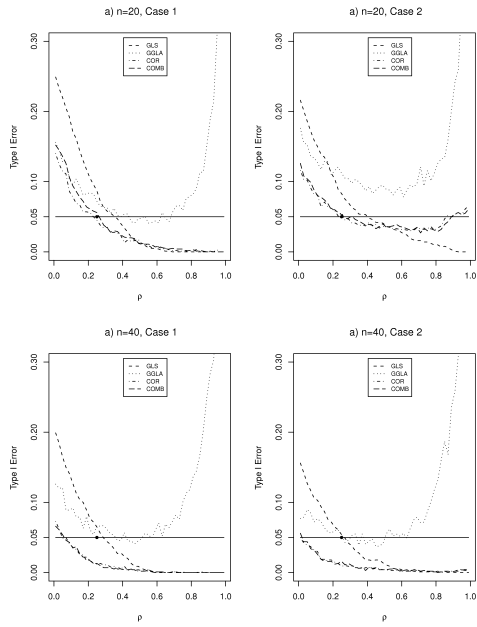
BIC : $\rho=0.25$



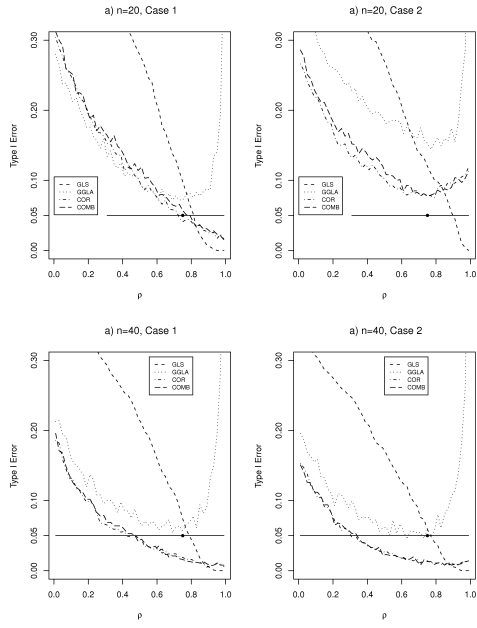
BIC : $\rho=0.75$

D.9.2 With Parsimony Correction

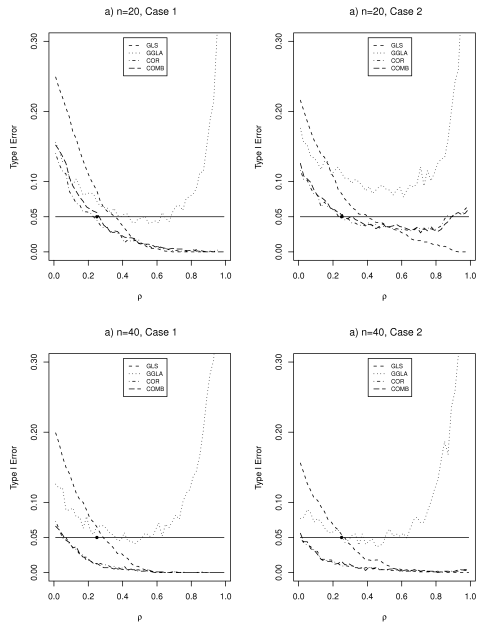
D.9.2.1 Cases 1 and 2



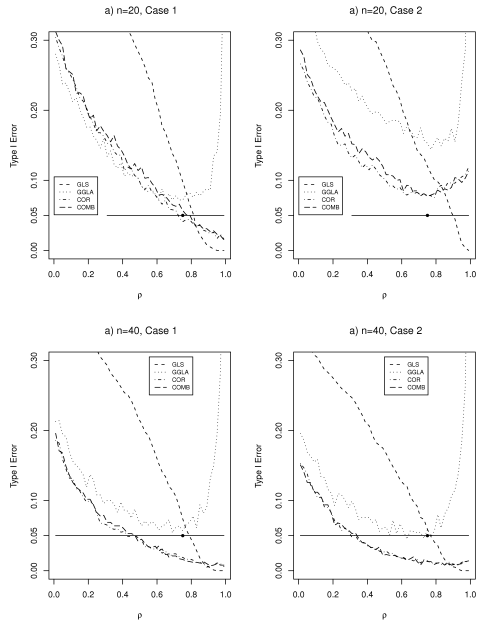
AIC : $\rho=0.25$



AIC : $\rho=0.75$

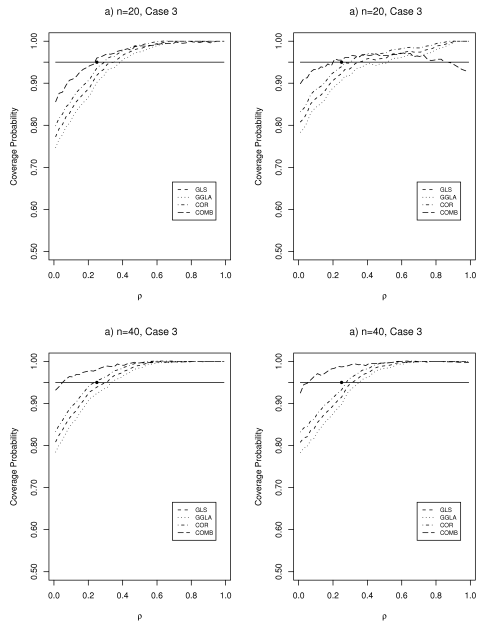


BIC : $\rho=0.25$

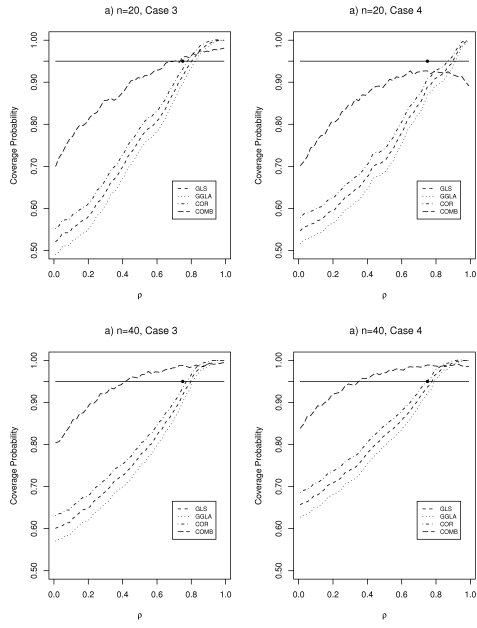


BIC : $\rho=0.75$

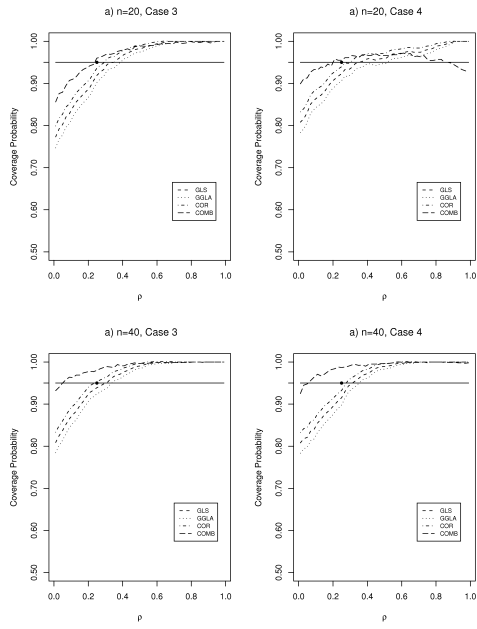
D.9.2.2 Cases 3 and 4



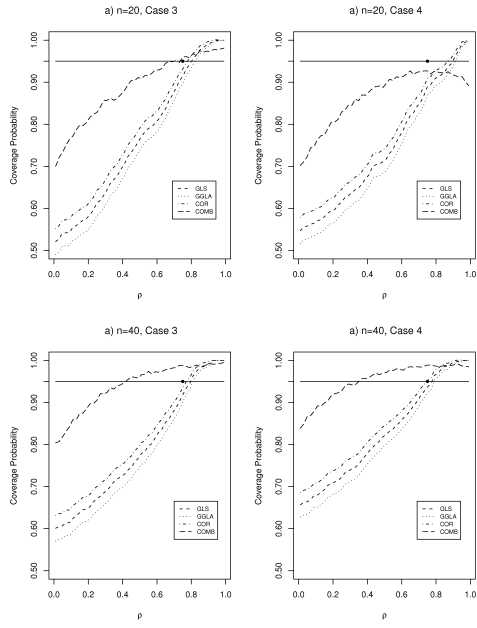
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$

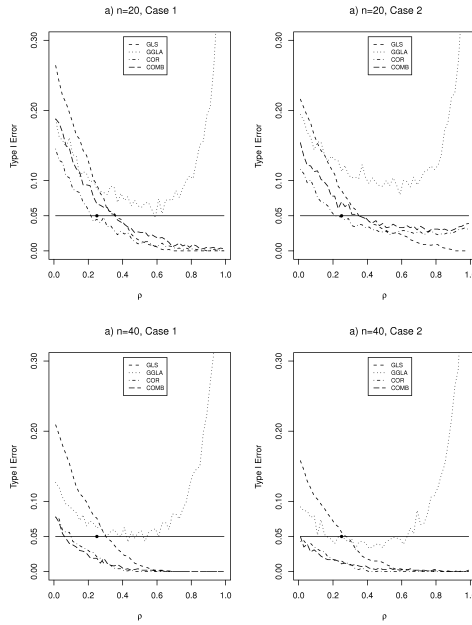


BIC : $\rho=0.75$

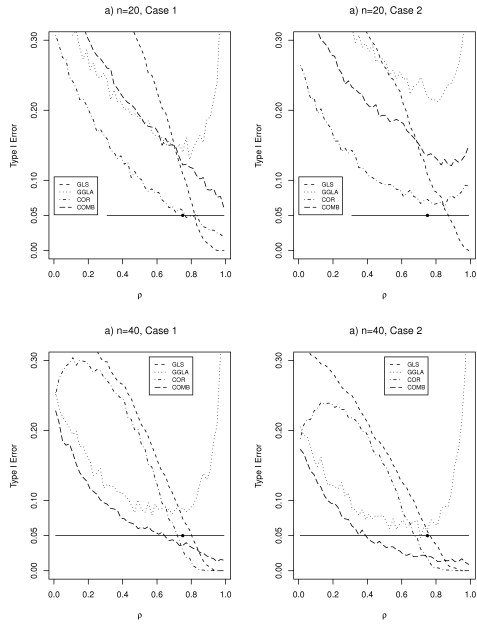
D.10 Type-I Error and Coverage Probability - Data-Driven Method

D.10.1 No Parsimony Correction

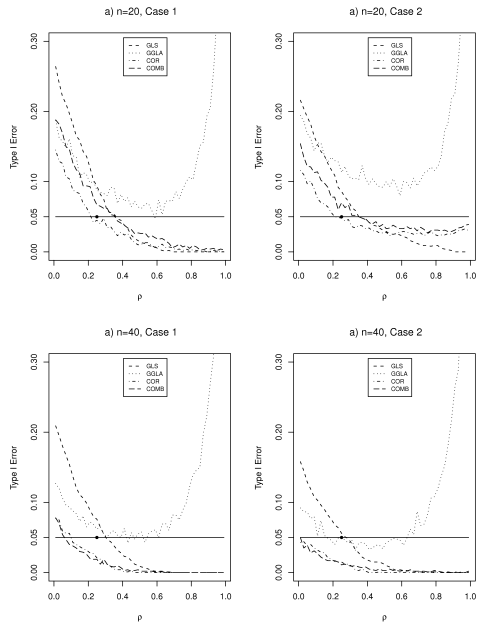
D.10.1.1 Cases 1 and 2



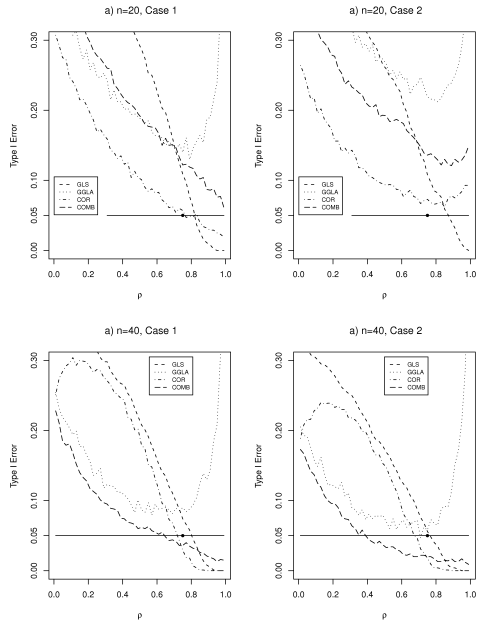
AIC : $\rho=0.25$



AIC : $\rho=0.75$

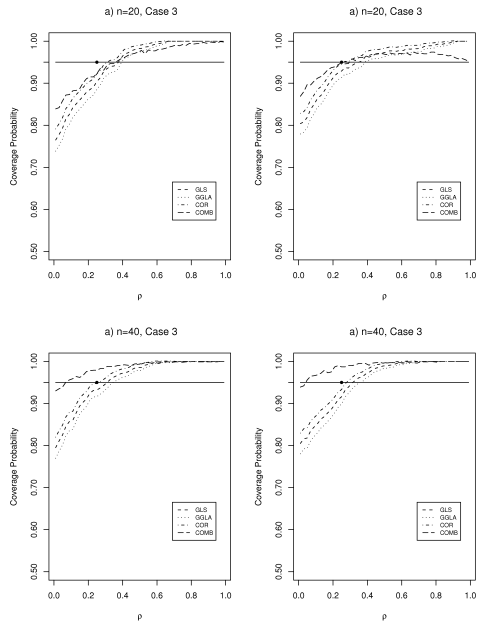


BIC : $\rho=0.25$

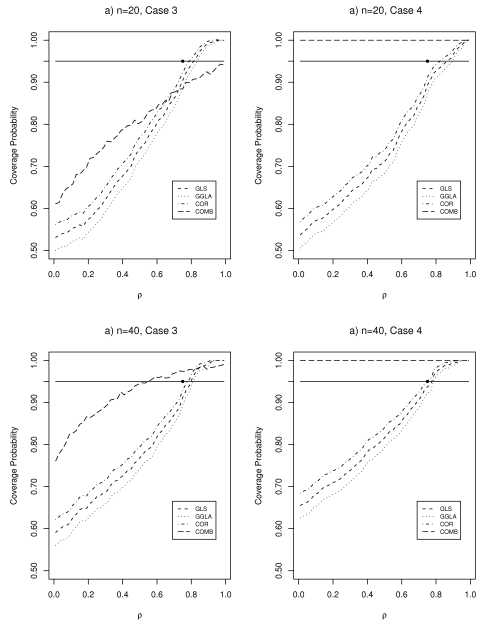


BIC : $\rho=0.75$

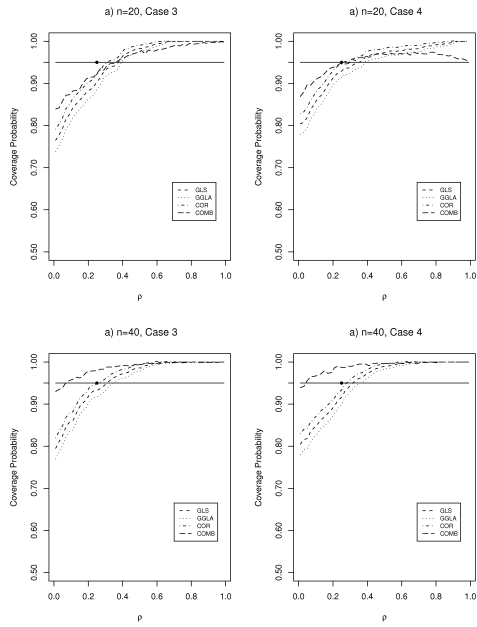
D.10.1.2 Cases 3 and 4



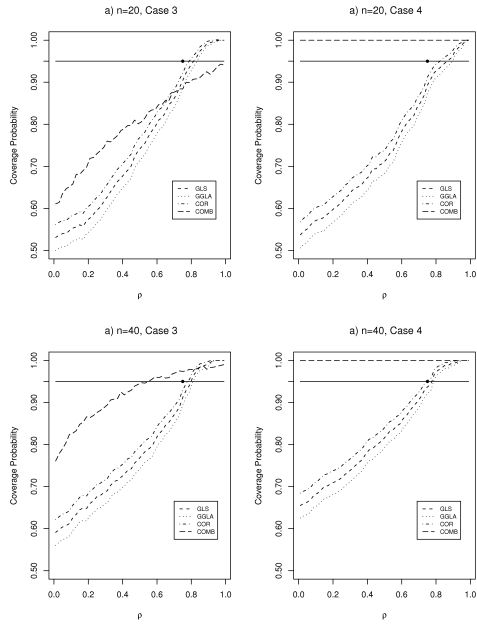
AIC : $\rho=0.25$



AIC : $\rho=0.75$



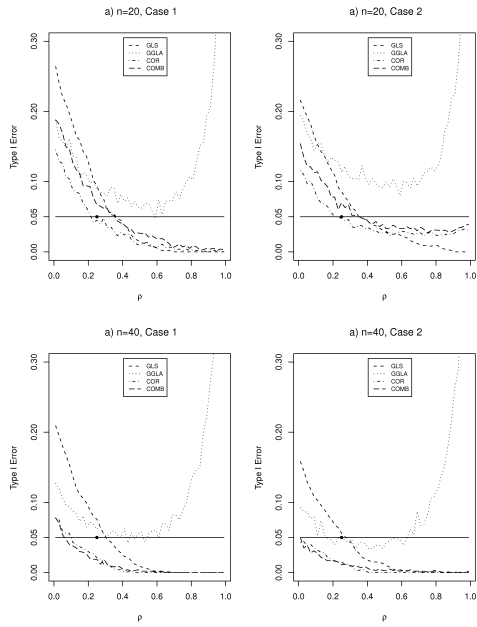
BIC : $\rho=0.25$



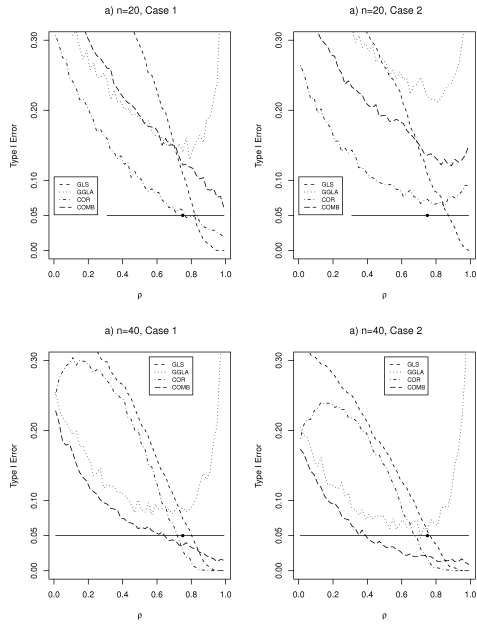
BIC : $\rho=0.75$

D.10.2 With Parsimony Correction

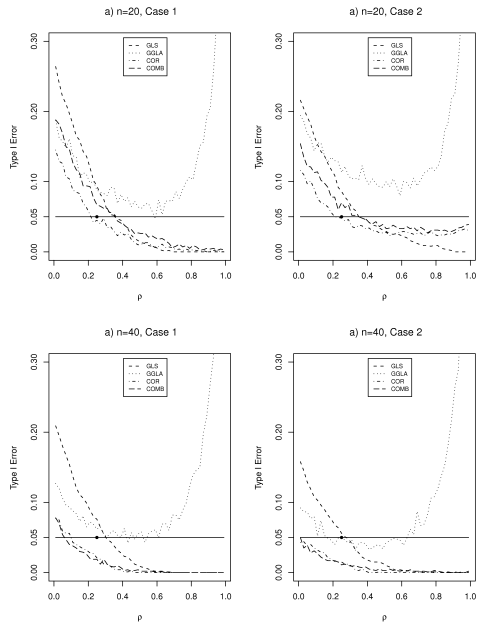
D.10.2.1 Cases 1 and 2



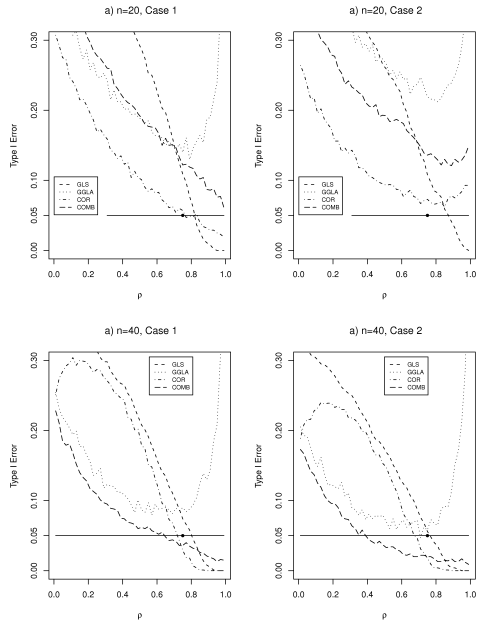
AIC : $\rho=0.25$



AIC : $\rho=0.75$

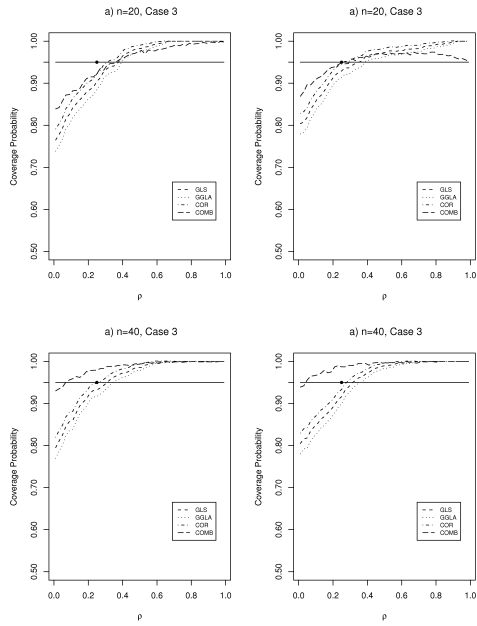


BIC : $\rho=0.25$

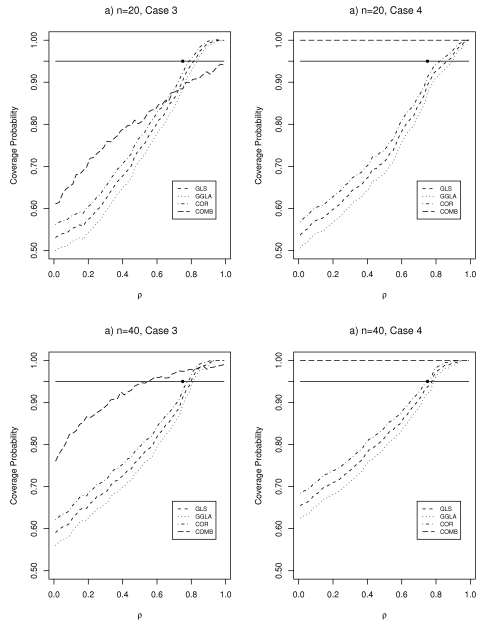


BIC : $\rho=0.75$

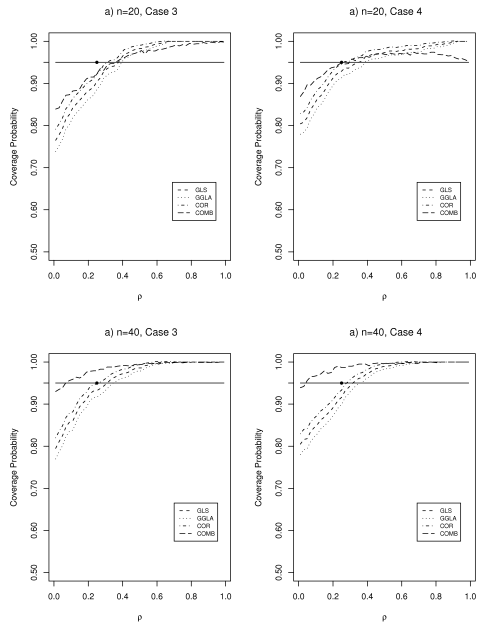
D.10.2.2 Cases 3 and 4



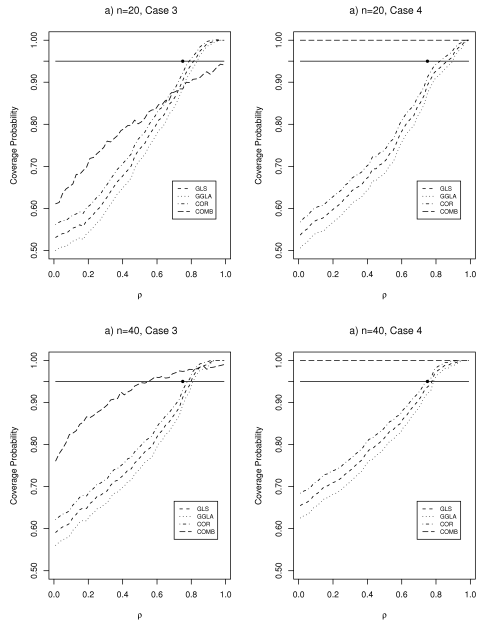
AIC : $\rho=0.25$



AIC : $\rho=0.75$



BIC : $\rho=0.25$



BIC : $\rho=0.75$

BIBLIOGRAPHY

BIBLIOGRAPHY

- Arditi, R. (1989), Avoiding fallacious significance tests in stepwise regression - a monte-carlo method applied to a meteorological theory for the canadian lynx cycle, *International journal of biometeorology*, 33(1), 24–26.
- Austin, P. C. (2008), Using the bootstrap to improve estimation and confidence intervals for regression coefficients selected using backwards variable elimination, *Statistics in medicine*, 27(17), 3286–3300.
- Berk, R., L. Brown, and L. Zhao (2010), Statistical inference after model selection, *Journal of Quantitative Criminology*, 26(2), 217–236.
- Burnham, K. P., and D. R. Anderson (2002), *Model Selection and Multimodel Inference*.
- Cheng, J., L. J. Edwards, M. M. Maldonado-Molina, K. A. Komro, and K. E. Muller (2010), Real longitudinal data analysis for real people: Building a good enough mixed model, *Statistics in medicine*, 29(4), 504–520.
- Danilov, D., and J. R. Magnus (2004), On the harm that ignoring pretesting can cause, *Journal of Econometrics*, 122(1), 27–46.
- Demetrescu, M., U. Hassler, and V. Kuzin (2011), Pitfalls of post-model-selection testing: experimental quantification, *Empirical Economics*, 40(2), 359–372.
- Dimova, R. B., M. Markatou, and A. H. Talal (2011), Information methods for model selection in linear mixed effects models with application to hcv data., *Computational statistics data analysis*, 55(9), 2677.
- Efron, B. (1986), How biased is the apparent error rate of a prediction rule, *Journal of the American Statistical Association*, 81(394).
- Efron, B. (2004), The estimation of prediction error: Covariance penalties and cross-validation, *Journal of the American Statistical Association*, 99(467), 619–632.
- Finos, L., C. Brombin, and L. Salmaso (2010), Adjusting stepwise p-values in generalized linear models, *Communications in Statistics-Theory and Methods*, 39(10), 1832–1846.

- Foster, S. D., A. P. Verbyla, and W. S. Pitchford (2009), Estimation, prediction and inference for the lasso random effects model, *Australian & New Zealand Journal of Statistics*, 51(1), 43–61.
- Girard, D. A. (1989), A fast monte-carlo cross-validation procedure for large least-squares problems with noisy data, *Numerische Mathematik*, 56(1).
- Giri, K., and P. Kabaila (2008), The coverage probability of confidence intervals in 2(r) factorial experiments after preliminary hypothesis testing, *Australian & New Zealand Journal of Statistics*, 50(1), 69–79.
- Gurka, M. J., L. J. Edwards, and K. E. Muller (2011), Avoiding bias in mixed model inference for fixed effects, *Statistics in medicine*, 30(22), 2696–2707.
- Hastie, T., R. Tibshirani, and J. Friedman (2001), *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York Inc., New York, NY, USA.
- Hurvich, C., and C. Tsai (1990), The impact of model selection on inference in linear-regression, *American Statistician*, 44(3), 214–217.
- Ibrahim, J. G., H. Zhu, R. I. Garcia, and R. Guo (2011), Fixed and random effects selection in mixed effects models, *Biometrics*, 67(2), 495–503.
- Ioannidis, J. P. A. (2008), Why most discovered true associations are inflated., *Epidemiology*, 19(5), 640–648.
- Joghataei, M. T., A. M. Arab, and H. Khaksar (2004), The effect of cervical traction combined with conventional therapy on grip strength on patients with cervical radiculopathy, *Clinical rehabilitation*, 18(8).
- Kabaila, P. (1995), The effect of model selection on confidence-regions and prediction regions, *Econometric Theory*, 11(3), 537–549.
- Kabaila, P. (1998), Valid confidence intervals in regression after variable selection, *Econometric Theory*, 14(4), 463–482.
- Kabaila, P., and H. Leeb (2006), On the large-sample minimal coverage probability of confidence intervals after model selection, *Journal of the American Statistical Association*, 101(474), 619–629.
- Kutner, M., C. Nachtsheim, and J. Neter (2004), *Applied linear regression models*, Irwin/McGraw-Hill series in operations and decision sciences, McGraw-Hill/Irwin.
- Leeb, H. (2005), The distribution of a linear predictor after model selection: conditional finite-sample distributions and asymptotic approximations, *Journal of Statistical Planning and Inference*, 134(1), 64–89.

- Leeb, H. (2006), The distribution of a linear predictor after model selection: Unconditional finite-sample distributions and asymptotic approximations, *Lecture Notes-Monograph Series*, 49(, Optimality: The Second Erich L. Lehmann Symposium), 291–311.
- Leeb, H. (2009), Conditional predictive inference post model selection, *Annals of Statistics*, 37(5B), 2838–2876.
- Leeb, H., and B. M. Poetscher (2008), Can one estimate the unconditional distribution of post-model-selection estimators?, *Econometric Theory*, 24(2), 338–376.
- Leeb, H., and B. Potscher (2005), Model selection and inference: Facts and fiction, *Econometric Theory*, 21(1), 21–59.
- Leeb, H., and B. M. Potscher (2003), The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations, *Econometric Theory*, 19(1), 100–142.
- Littell, R. C., G. A. Milliken, W. W. Stroup, R. D. Wolfinger, and O. Schabenberger (2006), *SAS for mixed models*, Cary : SAS institute, iD: 988447.
- Mundry, R., and C. L. Nunn (2009), Stepwise model fitting and statistical inference: Turning noise into signal pollution, *American Naturalist*, 173(1), 119–123.
- Myers, R. H., and J. S. Milton (1998), *A First Course in the Theory of Linear Statistical Models*, 2 ed., Primis Custom Publishing.
- Poetscher, B. M., and H. Leeb (2009), On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding, *Journal of Multivariate Analysis*, 100(9), 2065–2082.
- Potscher, B. M. (1991), Effects of model selection on inference, *Econometric Theory*, 7(2), 163–185.
- Potscher, B. M. (1995), Comment on the effect of model selection on confidence-regions and prediction regions by kabaila,p, *Econometric Theory*, 11(3), 550–559.
- Potscher, B. M., and A. J. Novak (1998), The distribution of estimators after model selection: Large and small sample results, *Journal of Statistical Computation and Simulation*, 60(1), 19–56.
- Salt, D. W., S. Ajmani, R. Crichton, and D. J. Livingstone (2007), An improved approximation to the estimation of the critical f values in best subset regression, *Journal of Chemical Information and Modeling*, 47(1), 143–149.
- Sen, P. K. (1979), Asymptotic properties of maximum likelihood estimators based on conditional specification, *Annals of Statistics*, 7(5), 1019–1033.
- Shao, J. (1997), An asymptotic theory for linear model selection, *Statistica Sinica*, 7(2), 221–242.

- Shen, X., and H.-C. Huang (2006), Optimal model assessment, selection, and combination, *Journal of the American Statistical Association*, 101(474), pp. 554–568.
- Shen, X., H. Huang, and J. Ye (2004a), Inference after model selection, *Journal of the American Statistical Association*, 99(467), 751–762.
- Shen, X. T., H. C. Huang, and J. Ye (2004b), Adaptive model selection and assessment for exponential family distributions, *Technometrics*, 46(3), pT: J; TC: 15; UT: WOS:000222767500004.
- Shibata, R. (1984), Approximate efficiency of a selection procedure for the number of regression variables, *Biometrika*, 71(1), 43–49.
- Vaida, F., and S. Blanchard (2005), Conditional akaike information for mixed-effects models, *Biometrika*, 92(2), 351–370.
- Vallejo, G., M. Fernandez, P. L. Rojas, and E. T. Herrero (2010), Selecting the best unbalanced repeated measures model, *Behavior research methods*, 43(1), 18–36.
- Verbeke, G., and G. Molenberghs (2000), *Linear mixed models for longitudinal data*, New York : Springer, New York.
- Wang, R., and S. W. Lagakos (2009), Inference after variable selection using restricted permutation methods, *Canadian Journal of Statistics-Revue Canadienne De Statistique*, 37(4), 625–644.
- Weisberg, S. (2005), *Applied linear regression*, 3 ed., 310 pp., John Wiley & Sons, Hoboken, N.J.
- Ye, J. M. (1998), On measuring and correcting the effects of data mining and model selection, *Journal of the American Statistical Association*, 93(441), 120–131.
- Young, I. A., L. A. Michener, and J. A. Cleland (2010), Manual therapy, exercise, and traction for patients with cervical radiculopathy ... (vol 89, pg 628, 2009), *Physical Therapy*, 90(5).
- Zhang, B., X. Shen, and S. L. Mumford (2012), Generalized degrees of freedom and adaptive model selection in linear mixed-effects models, *Computational Statistics & Data Analysis*, 56(3), 574–586.
- Zhang, P. (1992), Inference after variable selection in linear regression models, *Biometrika*, 79(4), 741–746.