Bootstrapping linear models, and its applications.

Author:
Tsang, Lester Hing Fung

Publication Date:
2011

DOI:
https://doi.org/10.26190/unsworks/23766

Bootstrapping linear models and its applications

Lester Hing Fung Tsang
B. A., M. Stat.

School of Mathematics and Statistics

The University of New South Wales, Australia

Supervisors: Sally Galbraith
David Warton

Submitted for the degree of Master of Science (research), July 13, 2011

Contents

Acknowledgements

Thanks to my supervisors Sally Galbraith and David Warton for their time, guidance and support. Thanks also to the School Computing Helpdesk for their assistance in running simulations. Finally, I would like to thank my parents for their support while I undertook this project.

Abstract

The bootstrap is a computationally intensive data analysis technique. It is particularly useful for analysing small datasets, and for estimating the sampling distribution of a statistic when it is intractable. We focus on bootstrap hypothesis testing of linear models. In this context, at present, various versions of the bootstrap are available, and it is not entirely clear from the literature which method is optimal for each situation.

The existing literature on bootstrapping linear models was reviewed, and three "rules" were found in the literature. We confirmed these via simulation. We also identified two outstanding issues. Firstly, which variance estimator should be used when constructing a bootstrap test statistic? Secondly, if resampling residuals, should this be done using the model that was fitted under the null hypothesis ("null model") or under the alternative hypothesis ("full model")? To our knowledge, these two questions have not been previously addressed. We provided theoretical results to answer these questions, and subsequently confirmed these via simulation. Our simulations were designed to evaluate both the size and (size-adjusted) power characteristics of the proposed bootstrap schemes.

We proposed the use of a sandwich variance estimator for case and score resampling, rather than the naive statistic that is commonly used in practice. Via simulation, we showed that bootstrap test statistics using the sandwich estimator tend to have superior Type I error for case and score resampling, but there was still an issue of which estimator (naive or sandwich) to use for the observed test statistic (t). Best results were achieved when using t-naive for score resampling and t-sandwich for case resampling. One possible explanation for this result is that score

resampling conditions on X whereas case resampling does not, and instead treats X as random.

We also studied full versus null model residual resampling. We showed that null model resampling has better Type I error in theory, having an asymptotic correlation of one with a "true bootstrap" procedure, analogous to a result derived in the permutation testing case by Anderson & Robinson (2001). However in practice, this superiority holds only for non-pivotal statistics: for pivotal statistics, both null and full model resampling had accurate Type I error, a discrepancy which we were able to explain theoretically.

Chapter 1

Introduction

1.1 Overview of the bootstrap technique for data analysis

The bootstrap is a computationally intensive statistical technique that depends on modern computing power. It is a generally applicable technique that is particularly useful for making inferences about a statistic that has an unknown sampling distribution. A particular case where this arises is in analysing small datasets, when distributional assumptions are not (or may not be) satisfied. The method is widely used and several introductory texts have been written on the topic (for example, Davison & Hinkley, 1997; Manly, 1997; Chernick, 2008). The basic idea is to treat the empirical distribution from the data as the true distribution, and then to resample from this distribution to estimate the sampling distribution of any statistic of interest. There are many different variations on the bootstrap method, which resample different quantities in different ways. For example, in case resampling (Davison & Hinkley, 1997), the observations from the original dataset are resampled, whereas in residual resampling (Davison & Hinkley, 1997), the objects being resampled are residuals obtained from some model fitted to the original dataset. Different methods make different assumptions and have different properties, a topic which will be explained in this thesis. In some sense, case resampling is less restrictive than residual resampling, because case resampling does not assume a particular distribution for

the residuals, while residual resampling assumes that the residuals are exchangeable and thus come from the same distribution.

Many papers have been written on applying the bootstrap to practical problems, and on the behaviour of the bootstrap when applied to linear models, generalized linear models (GLM's) and even more general classes of models. Some key references for this thesis are as follows. Davison & Hinkley (1997) describe the theoretical basis and the practical application of the bootstrap, including discussing situations in which the bootstrap may be applied. Hall & Wilson (1991) present some guidelines to follow in constructing test statistics and using the bootstrap to evaluate significance. Wu (1986) compares the bootstrap with the jackknife and other resampling methods in regression analysis, and proposes "score resampling", a method for handling heteroscedasticity. Friedl & Stadlober (1997) describe resampling methods in the GLM context that may be used to analyse environmental datasets. Both Freedman (1981), and Moulton & Zeger (1991), investigate the theoretical, and in particular the asymptotic, behaviour of the bootstrap, when applied to linear regression models and GLM's respectively. Another paper of interest is Anderson & Robinson (2001), who provide theoretical properties of different permutation testing methods for linear models, as well as presenting results of simulations. This thesis focuses on *bootstrap* methods for *linear models*.

1.2 Definition of linear models

We consider the general linear model

$$Y = X\beta + \epsilon$$

where Y is the vector of responses, $Y = [Y_1, \ldots, Y_N]^T$, X is the design matrix with $1 \times p$ row vector X_i containing explanatory variables for the ith case, $i = 1, \ldots, N$, β is a $p \times 1$ vector of unknown parameters, and ϵ is a vector of residual errors, $\epsilon = [\epsilon_1, \ldots, \epsilon_N]^T$, which are assumed to be mutually independent and to have mean zero. The distribution of the ϵ_i is unknown, and they may have constant variance (homoscedastic) or non-constant variance (heteroscedastic).

The theory of linear models has been well developed, and we use it as a starting point for an examination of three different bootstrap resampling strategies (residual, case and score resampling, to be defined in Section 1.3). The theoretical properties of these bootstrap techniques will be compared to results obtained by simulation.

This thesis focuses on hypothesis testing for linear models, and specifically on "partial tests" of the form $H_0 : \beta_k = 0$ for some k. In this setting, there is generally no exact permutation test available, contrary to $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ (Anderson & Robinson, 2001), and surprisingly, there are many unanswered questions in how to construct such tests.

The aims of the thesis are:

- To present a review of the current literature concerning bootstrapping linear models (Chapter 2).
- To confirm the validity of three "known" rules for bootstrap hypothesis testing already in the literature (Chapter 3).
- To identify which variance estimator should be used with which resampling method in the construction of pivotal statistics (Chapter 4).
- To compare the performance of residuals from the "full" (that is, under H_1) and "null" (that is, under H_0) models, in residual resampling (Chapter 5).

1.3 Linear model bootstrap methods

In this section, we review the definitions and some properties of the different resampling methods. We follow the terminology of Davison & Hinkley (1997).

1.3.1 Case resampling

Case resampling is the most intuitive of the bootstrap methods. It involves resampling cases from the original dataset, and then fitting linear models to these bootstrap resamples.

We define case resampling mathematically as follows. Let i^* be a vector of length N with elements obtained by randomly resampling with replacement from $\{1, 2, \ldots, N\}$. Let i^*_j be the jth element of i^*. Then $Y_* = (Y_{i^*_1}, Y_{i^*_2}, \ldots, Y_{i^*_N})^T$ and the jth row of X_* is $(X_{i^*_1}, X_{i^*_2}, \ldots, X_{i^*_N})$.

The case resample estimate of $\hat{\beta}$ is obtained as:

$$\hat{\beta}_* = (X_*^T X_*)^{-1} X_*^T Y_*.$$

It has been shown by Davison & Hinkley (1997) that in general, case resampling is less efficient than residual resampling, but that it is more robust against variance heteroscedasticity and model misspecification.

This is often thought of as the "correlation model" (Freedman, 1981). Note that by definition the design matrix X is not held fixed, and so inference based upon it is not conditional on the design points.

1.3.2 Residual resampling

Residual resampling involves resampling the residuals from the fitted model:

$$Y_* = X\hat{\beta} + r^*$$

where each element of r^* has probability $1/N$ of taking each value in $r_1, \ldots, r_N$.

In the context of hypothesis testing, the linear model from which the residuals are obtained may be the full (alternative) model, which will be referred to as full model residual resampling, or may be the null model (i.e. the model estimated under the null hypothesis), which will be referred to as null model residual resampling. One issue this thesis considers is the choice between full and null model residual resampling.

Another issue that arises in residual resampling is whether raw or modified residuals should be used. If the raw (or unmodified) residuals are denoted by r_i, then the modified residuals m_i are defined as:

$$m_i = r_i / \sqrt{1 - h_i},$$

where h_i is the ith diagonal element of the hat matrix $H = X(X^T X)^{-1} X^T$ (Seber, 1977).

The consequences of using raw versus modified residuals are considered in Chapter 3.

Residual resampling assumes that the error terms in the model are independently and identically distributed (i.i.d.). Thus residual resampling has been shown to be not robust against error variance heteroscedasticity (for example, Liu & Singh, 1992; Bose & Chatterjee, 2002).

1.3.3 Score resampling

Score resampling multiplies the residuals by a random variable t^* which has mean zero and variance one. That is:

$$Y_* = X\hat{\beta} + \text{diag}(r)t^*$$

where $E(t^*) = \mathbf{0}$ and $\text{var}(t^*) = I$.

These t^* may be standard normally distributed, or may be resampled from the standardized residuals themselves, or from another suitable distribution. We used the two former choices for the distibution of t_i^*, the second of these suggested by Wu (1986), and the first used by Friedl & Stadlober (1997) in their simulation studies. Score resampling is essentially a trick to produce variance estimates that are unbiased and to ensure robustness against error variance heteroscedasticity: in fact, it was specially devised for the heteroscedastic error case. This can be understood intuitively by noting that the ith residual is retained with the ith observation (i.e. the ith fitted value) in resamples. The method is named score resampling because it is based on a linear estimate of the score equations. It is useful for hypothesis testing, but not for prediction of individual observations. This is because in the prediction setting, the distribution of resampled residuals is critical, but the distribution of t^* is arbitrary.

Wu (1986) investigated the method now known as score resampling, under the label "A general method for resampling residuals", in comparison with other resampling methods, in both homoscedastic and heteroscedastic simulations, as well as comparing their theoretical properties.

1.4 Resampling notation

In this thesis, expectations and variances will sometimes be taken with respect to the sampling distribution of the data, (i.e. the distribution of Y), and at other times with respect to the resampling distribution (conditional on a set of observed Y). These two cases will be distinguished using conventional conditional probability notation, for example $E(\hat{\beta})$ with respect to the sampling distribution, or $E(\hat{\beta}|Y)$ with respect to the bootstrap distribution.

If the resampling method is clear from the context, whether it be residual, score or case resampling, we use $*$ to denote this resampling method. In this thesis, we have taken care to clarify the resampling method in any context, but if the resampling method is in doubt, assume case or score resampling in Chapter 4 and residual resampling in Chapter 5. We would like to be able to distinguish objects which have been resampled directly (for example, the residuals in residual resampling) from objects that are functions of resampled objects (for example, $\hat{\beta}_{*,k}$ for a bootstrap sample). Hence, a superscript * (for example, r^*) denotes a bootstrap resample of a certain random variable or matrix, while a subscript $_*$ (for example, $\hat{\beta}_*$ or t_*) denotes a test statistic or random variable that is a function of resampled values. Further, we use Y_* to denote a case resample of Y, because in case resampling Y and X are resampled jointly, and we also use Y_* to denote a residual or score resample of Y.

Chapter 2

Guidelines for applying the bootstrap

The aim of this Chapter is to review the literature and draw on it to provide guidelines for the linear model bootstrap, and also the reasons for these guidelines. Hence in Section 2.1 we will present the theoretical properties of the three resampling methods (residual, case, score). We describe the assumptions made by each method, and thus define the proper application of each method. In Section 2.2, we describe some currently-known issues in bootstrap hypothesis testing for linear models, based on existing literature. In Section 2.3, we summarize the previous work into three known "rules" for the bootstrapping of linear models, which we will verify via simulation in Chapter 3.

2.1 Properties of the different resampling methods

When resampling, we would like to mimic properties of $\hat{\beta}$, since we use the resampling distribution of $\hat{\beta}_*$ to estimate the sampling distribution of $\hat{\beta}$. It is well known

Table 2.1: Summary of the properties of the different resampling methods

Resampling method	Design	Robust to heteroscedasticity	Assumptions	$E(\hat{\beta}_*) = \hat{\beta}$
Case	X random	yes	independent cases	no
Residual	X fixed	no	i.i.d. residuals	yes
Score	X fixed	yes	independent residuals	yes

that for the "true model" $Y = X\beta + \epsilon$, $E(\hat{\beta}) = \beta$ and:

$$\mathrm{var}(\hat{\beta}) = \begin{cases} \sigma^2(X^TX)^{-1} & \text{homoscedastic}, \mathrm{var}(\epsilon_i) = \sigma^2 \\ (X^TX)^{-1}X^T\mathrm{diag}(\sigma_i^2)X(X^TX)^{-1} & \text{heteroscedastic}, \mathrm{var}(\epsilon_i) = \sigma_i^2 \end{cases}$$

and a Central Limit Theorem argument can be used to show that for each j:

$$\frac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\hat{\beta}_j)} \xrightarrow{D} \mathcal{N}(0,1)$$

for ϵ from any distribution with finite variance. Ideally, we would like the chosen resampling method to also have these properties, and will review the properties of $\hat{\beta}_*$ under different resampling methods as currently understood.

A summary of the properties of the different resampling methods (case, residual and score) is presented in Table 2.1.

2.1.1 Residual resampling

Recall that residual resampling is defined as: $Y_* = X\hat{\beta} + r^*$. Conditional on the data, residual resampling can be shown to have $E(\hat{\beta}_*)$ equal to $\hat{\beta}$ and $\mathrm{var}(\hat{\beta}_*)$ equal to $(1/N)\sum_{i=1}^N r_i^2(X^TX)^{-1}$.
These two statements can be shown as follows.

$$\begin{aligned} E(\hat{\beta}_*|Y) &= E((X^TX)^{-1}X^TY^*) \\ &= E((X^TX)^{-1}X^T(X\hat{\beta} + r^*)) \\ &= \hat{\beta} \end{aligned}$$

since $E(r^*) = 0$.

$$\begin{aligned}
\mathrm{var}_{\mathrm{residual}}(\hat{\beta}_*|Y) &= (X^TX)^{-1}X^T\mathrm{var}(r^*)X(X^TX)^{-1} \\
&= (X^TX)^{-1}X^T\left[(1/N)\sum_{i=1}^{N} r_i^2\right] IX(X^TX)^{-1} \\
&\quad \text{(since the empirical distribution function puts mass } (1/N) \text{ at each } r_i) \\
&= \frac{1}{N}\sum_{i=1}^{N} r_i^2 (X^TX)^{-1} \qquad (2.1) \\
&= \hat{\sigma}^2(X^TX)^{-1} \\
&\quad \text{if modified residuals are used}
\end{aligned}$$

We write the variance (and expectation) as conditional on data Y because, when resampling, the data are the sampling "universe". Also, the two statements may be found in Moulton & Zeger (1991) and elsewhere. What $\hat{\beta}$ is and what r_i are, depend on whether full/null resampling is being applied and whether modified/raw residuals are being used. It follows from equation (2.1) that the variance estimator is biased when raw residuals are used, but that the bias is removed when modified residuals are used instead.

2.1.2 Case resampling

Recall that case resampling selects elements of Y and rows of X using a vector i^* obtained by resampling with replacement from $\{1, 2, \ldots, N\}$. Freedman (1981) showed that:

$$\sqrt{N}(\hat{\beta}_* - \hat{\beta}) \xrightarrow{d} \mathcal{N}(0, J^{-1}MJ^{-1}),$$

where $J = E(X_i^TX_i)$ and $M = E(X_i^TX_ir_i^2)$,
and X_i is defined to be the ith row of X. Note that in the definitions of J and M, we could replace X_i with $X_{*,i}$ because of the definition of case resampling.

An outline of the proof is given in Appendix A.2. Although Freedman's result gives the asymptotic distribution of $\hat{\beta}_*$ it does not give any indication of how quickly $\hat{\beta}_*$ approaches this distribution, with increasing sample size. Moulton & Zeger (1991)

addressed this, at least for $\text{var}(\hat{\beta}_*)$, when they stated that the rate of convergence is of order N^{-2}. That is, they stated that:

$$\text{var}_{\text{case}}(\hat{\beta}_*) = (X^T X)^{-1} X^T \text{diag}(r_i^2) X (X^T X)^{-1} + O(N^{-2}).$$

This is shown in Appendix A.

One key difference from residual resampling is that $E(\hat{\beta}_*)$ is only approximately $\hat{\beta}$, not exact. One way to see this is to consider small N, for which there is a positive chance that the matrix $(X_*^T X_*)$ is singular, and $(X_*^T X_*)^{-1}$ is undefined. In this thesis, the case resampling scheme was constructed so that the singular cases were excluded. Nevertheless, even without the presence of a singular case, the estimator $\hat{\beta}_*$ is a biased estimator (Figure 2.1). Figure 2.1 plots $\hat{\beta}_*$ versus $\hat{\beta}$ for case resampling and full model residual resampling, in comparison with the line $y = x$. Note that for case resampling, the points generally do not lie on the line $y = x$, but instead hover around the line, while for residual resampling, the points are statistically indistinguishable from the line $y = x$. Thus, $E(\hat{\beta}_*)$ is only approximately $\hat{\beta}$ for case resampling, but is exactly $\hat{\beta}$ for residual resampling.

Another difference is that X is treated as random, whereas residual resampling (and conventional regression) condition on X. Hall (1992) relates the "regression model" to residual resampling, and the "correlation model" to case resampling. On page 168 of Hall (1992), the regression model is defined to be inference conditional on the design points (design points are fixed or random but conditioned upon) and errors are random i.i.d.. The correlation model is defined to be inference on slope, with the property that (X_i, Y_i), $i = 1, \ldots, N$ are independent pairs of random vectors, and to be unconditional inference (with the special case of independent errors). Hall (1992, page 183) shows that in regression models, bootstrap confidence interval (C.I.) methods for slope parameters have a special property that for the percentile method, C.I.'s are second-order correct (usually first-order), while for the percentile-t method, the order is smaller than N^{-1} (usually N^{-1}). The reason given is the symmetry of the regression model:

$$E(N^{-1} \sum_{i=1}^{N} (x_i - \bar{x}) e_i^j) = 0.$$

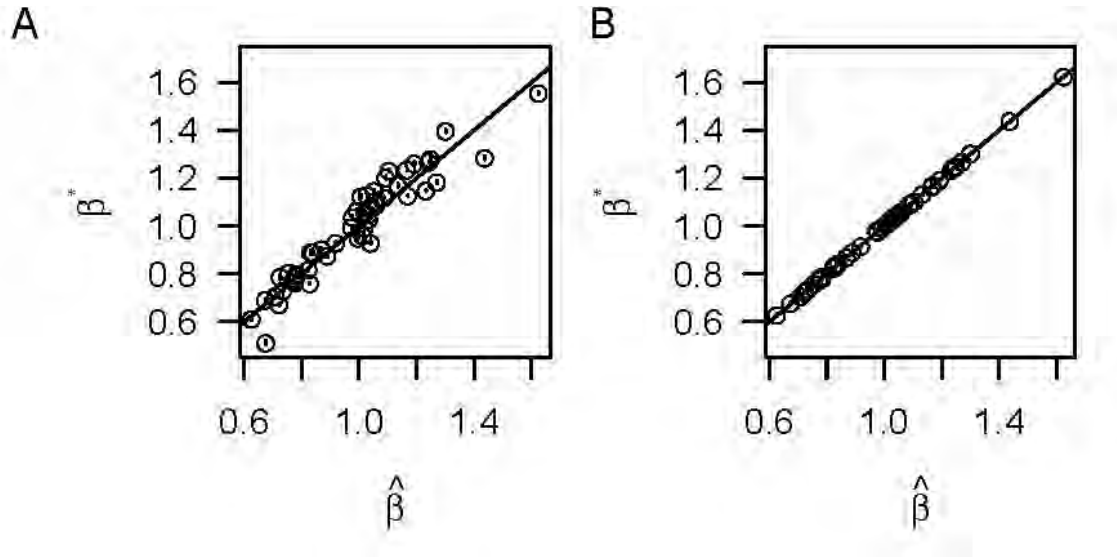

Figure 2.1: Plot of $\hat{\beta}_*$ versus $\hat{\beta}$ for (A) case resampling (left hand plot) and (B) full model residual resampling (right hand plot), with error bars. The simulation used a sample size (N) of 16, exponential(1) distributions for all covariates and true errors, a total of 50 simulations (different points), and a resample size of $B = 10000$. At each point, a vertical error bar (that is, a small vertical line) was drawn (assuming an approximate normal distribution) to see if the error bar crossed the line $y = x$. Note that $E(\hat{\beta}_*)$ is only approximately $\hat{\beta}$ for case resampling, but is exactly $\hat{\beta}$ for residual resampling.

However, the correlation model is not symmetric unless errors are independent (so the above unusual properties of bootstrap C.I.'s in regression are not available under the correlation model but do emerge under correlation with independent errors). With independent errors, the special properties are available for the correlation model: however in this case, the assumptions for residual resampling are satisfied, and various sources (for example, Liu & Singh, 1992) confirm the higher efficiency of residual resampling. Thus for linear models, if errors are i.i.d., residual resampling is expected to have more accurate Type I error than case resampling, as seen in Liu & Singh (1992) for example.

Hjorth (1994) states (p187) that case (vector) resampling is crude, since the fact that the bootstrap design matrix X_* becomes random leads to estimates that will typically exhibit more variability. However, it is sometimes the only feasible method, and exaggerated uncertainties may lead to conservative estimates of the variances. It is claimed (Hjorth, 1994) that case resampling, but not residual resampling, is robust against heteroscedasticity or the presence of non-linearity that is not properly modelled. It is stated (Hjorth, 1994) that case resampling works well for large datasets without very influential observations, while residual resampling is better for smaller datasets, or data with influential observations. If there is heteroscedasticity, score resampling may be used also, but residual resampling may be feasible within groups for which variances are known a priori to be constant within that group. Score and case resampling tend to need a lot of data for good performance in estimating variance.

2.1.3 Score resampling

Recall that score resampling is defined as: $Y_* = X\hat{\beta} + \text{diag}(r)t^*$, where $E(t^*) = \mathbf{0}$ and $\text{var}(t^*) = I$. Conditional on the data, score resampling can be shown to have $E(\hat{\beta}_*)$ equal to $\hat{\beta}$ and $\text{var}(\hat{\beta}_*)$ equal to $(X^TX)^{-1}X^T\text{diag}(r^2)X(X^TX)^{-1}$. The two

statements can be shown as follows.

$$\begin{aligned} E(\hat{\beta}_*|Y) &= E((X^TX)^{-1}X^T(X\hat{\beta} + \text{diag}(r)t^*)) \\ &= (X^TX)^{-1}X^TX\hat{\beta} + (X^TX)^{-1}X^T\text{diag}(r)E(t^*|Y) \\ &= \hat{\beta} \end{aligned}$$

since $E(t^*|Y) = \mathbf{0}$. Also:

$$\begin{aligned} \text{var}_{\text{score}}(\hat{\beta}_*|Y) &= (X^TX)^{-1}X^T\text{var}(\text{diag}(r)t^*)X(X^TX)^{-1} \\ &= (X^TX)^{-1}X^T\text{diag}(r)\text{var}(t^*)\text{diag}(r)X(X^TX)^{-1} \\ &= (X^TX)^{-1}X^T\text{diag}(r^2)X(X^TX)^{-1} \end{aligned}$$

since $\text{var}(t^*) = I$.

According to Wu (1986), the variance estimate taking the above variance form, is bias-robust against error variance heteroscedasticity. This is because no homoscedasticity assumption was made in deriving $E(\hat{\beta}_*)$ and $\text{var}(\hat{\beta}_*)$. This can also be understood intuitively by studying the resampling scheme: the r_i remain fixed under resampling and hence no assumption of homoscedasticity is made. This is in contrast to residual resampling, where the r_i are not fixed but are treated as i.i.d., and so residual resampling is not bias-robust against heteroscedasticity. For both score and residual resampling, $\hat{\beta}_*$ is unbiased for $\hat{\beta}$. However, in the presence of heteroscedasticity, the variance estimator is unbiased for score resampling, but is biased for residual resampling.

Note that any t^* with mean $\mathbf{0}$ and variance I gives the above properties, and there is no guidance on how to choose t^* beyond this. Hence we are assuming that the first and second moments of t^* are all that matters.

2.2 Issues in hypothesis testing for linear models

While there is an abundance of literature on bootstrapping, and even bootstrapping linear models, there is considerably less literature specifically related to the hypothesis testing context. However, there is a symmetry between hypothesis testing and

confidence intervals: a Wald test of $H_0 : \beta_k = 0$ at significance level α is equivalent to checking if β_k is in a $100(1-\alpha)$-percent confidence interval for β_k (Seber, 1977). Hence literature on bootstrap confidence intervals can further our understanding.

In the permutation-testing context, some previous authors have studied hypothesis testing of linear models. The most notable of these is Anderson & Robinson (2001), which describes the theoretical properties of different permutation-testing methods, as well as presenting the results of various simulations. Anderson & Legendre (1999) also provides the results of various simulations, and is therefore mainly empirical. We will extensively use Anderson & Robinson (2001) in Chapter 5 when we compare full and null model residual resampling.

Hall & Wilson (1991) present two "rules" for bootstrapping: to always use pivotal statistics (covered in Section 2.3.1), and to "bootstrap to reflect the null hypothesis". In regard to the second point, if $H_0 : \beta_k = 0$ and $E(\hat{\beta}_{*,k}) = \hat{\beta}_k$, construct the bootstrap using $P_*(|\hat{\beta}_{*,k} - \hat{\beta}_k| > |\hat{\beta}_k|)$ and not using $P_*(|\hat{\beta}_{*,k}| > |\hat{\beta}_k|)$. Or alternatively resample under the null hypothesis, such that $E(\hat{\beta}_{*,k}) = 0$. These ideas were used throughout the thesis, in particular in the construction of non-pivotal and pivotal statistics for each type of resampling method.

In the literature, three general rules of bootstrap hypothesis testing have emerged, as presented in the Section 2.3.

2.3 Three rules of bootstrap hypothesis testing

The three "known" rules of bootstrap hypothesis testing for improving the performance of the test are:

1. Always use a pivotal statistic. (For better Type I error.)

2. For residual and score resampling, use modified rather than raw residuals (For better variance estimation, hence, better Type I error.)

3. For homoscedastic data, use residual resampling. For heteroscedastic data, use case or score resampling. (Residual resampling is more efficient for ho-

moscedastic data, but not robust against heteroscedasticity.)

The reason for these rules will be explained below. We did not include Hall & Wilson (1991)'s second rule in the list, but note that it is essential to bootstrap to reflect the null hypothesis, in order to obtain a valid test.

2.3.1 Always use a pivotal statistic

A pivotal statistic is a statistic whose (asymptotic) distribution is not a function of the model parameters. In this thesis, the pivotal statistic takes the form $t_* = (\hat{\beta}_* - a)/\hat{\text{se}}_*(\hat{\beta}_*)$ where $a = 0$ for resampling methods where $E(\hat{\beta}_*) = 0$ and $a = \hat{\beta}$ for resampling methods where $E(\hat{\beta}_*) = \hat{\beta}$ (exactly or approximately). This statistic is pivotal for linear models in the sense that $t_* \xrightarrow{D} \mathcal{N}(0, 1)$. Any consistent estimator of $\text{se}(\hat{\beta}_*)$ can be used as $\hat{\text{se}}_*(\hat{\beta}_*)$. In Chapter 3, we use the form $\hat{\sigma}_*^2(X^T X)^{-1}$ for residual and score resampling, while we also consider the form $\hat{\sigma}_*^2(X_*^T X_*)^{-1}$ for case resampling. We will consider another choice for standard error estimation in Chapter 4, and we will define the actual bootstrap P-values in Chapter 3.

We expect pivotal statistics to have better Type I properties than non-pivotal statistics, for example as stated in Hall & Wilson (1991). Davison & Hinkley (1997) provide an insight as to why this is the case. Davison & Hinkley (1997) assert that bootstrap-t (i.e. pivotal statistics) have better approximation of coverage probability than the percentile method (i.e. non-pivotal statistics), because the bootstrap-t is second-order accurate (i.e. the coverage probability is accurate to within $O(N^{-1})$), while the percentile method is only first-order accurate, although equi-tailed confidence intervals (i.e. 2-tailed non-pivotal statistics) are second-order accurate. They state that the special case of equi-tailed confidence intervals for the percentile method is second-order accurate only because the first-order ($N^{-1/2}$) terms cancel out (z_α and $z_{1-\alpha}$). Thus for a 1-tailed C.I. (or a 1-tailed test), the percentile method is only first-order accurate ($O(N^{-1/2})$). Also, Hall (1992) showed that for linear models, pivotal statistics have Type I error accurate to within $O(N^{-3/2})$, whereas non-pivotal statistics only have Type I error accurate to within $O(N^{-1})$.

2.3.2 Use modified rather than raw residuals

In Section 2.1.1 we showed that $\text{var}_{\text{residual}}(\hat{\beta}_*) = (1/N)\sum_{i=1}^N r_i^2(X^TX)^{-1}$ where r_i may be raw or modified residuals. In Section 2.1.3 we showed that $\text{var}_{\text{score}}(\hat{\beta}_*) = (X^TX)^{-1}X^T\text{diag}(r^2)X(X^TX)^{-1}$.

However, if the r_i are raw residuals, then these variance estimators are biased. These variance estimators are biased because $\text{var}(r_i) = \sigma^2(1-h_i)$ for the raw residuals r_i and hence:

$$\begin{aligned} E\left(\frac{1}{N}\sum_{i=1}^N r_i^2\right) &= \frac{1}{N}\sigma^2\sum_{i=1}^N(1-h_i) \\ &= \frac{N-p}{N}\sigma^2 \end{aligned}$$

In the context of residual resampling, the un-biased estimator would use instead $\frac{1}{N-p}\sum_{i=1}^N r_i^2$. In the context of score resampling, the un-biased estimator would use instead $X^T\text{diag}(r^2)(I-H)^{-1}X$, where H is the hat-matrix $X(X^TX)^{-1}X^T$ (Seber (1977)). However, if modified residuals $m_i = \frac{r_i}{\sqrt{1-h_i}}$ are used instead, where h_i is the ith diagonal element of H, then since $(1/N)\sum_{i=1}^N m_i^2 = \frac{1}{N-p}\sum_{i=1}^N r_i^2$ and $X^T\text{diag}(m^2)X = X^T\text{diag}(r^2)(I-H)^{-1}X$, then the variance estimators become unbiased.

That is, using modified residuals $m_i = r_i/\sqrt{1-h_i}$, we have:

$$\begin{aligned} E(m_i^2) &= E(r_i^2[1-h_i]^{-1}) \\ &= (1-h_i)^{-1}E(r_i^2) \\ &= (1-h_i)^{-1}\sigma^2(1-h_i) \\ &= \sigma^2 \end{aligned}$$

The above equation holds for homoscedastic errors. For possibly heteroscedastic errors, we have similarly:

$$E(m_i^2) = \sigma_i^2.$$

Thus we have $\text{var}_{\text{residual}}(\hat{\beta}_*) = (1/N)\sum_{i=1}^{N}(m_i^2)(X^TX)^{-1}$ which is unbiased, since:

$$\begin{aligned} E[\text{var}_{\text{residual}}(\hat{\beta}_*)] &= E[(1/N)\sum_{i=1}^{N}(m_i^2)(X^TX)^{-1}] \\ &= (1/N)N\sigma^2(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1} \end{aligned}$$

Also we have $\text{var}_{\text{score}}(\hat{\beta}_*) = (X^TX)^{-1}X^T\text{diag}(m_i^2)X(X^TX)^{-1}$ which is unbiased, since:

$$\begin{aligned} E[\text{var}_{\text{score}}(\hat{\beta}_*)] &= E[(X^TX)^{-1}X^T\text{diag}(m_i^2)X(X^TX)^{-1}] \\ &= (X^TX)^{-1}X^T\text{diag}(E[m_i^2])X(X^TX)^{-1} \\ &= (X^TX)^{-1}X^T\text{diag}(\sigma_i^2)X(X^TX)^{-1} \\ & \quad \text{for possibly heteroscedastic errors} \\ &= (X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1} \\ & \quad \text{for homoscedastic errors} \\ &= \sigma^2(X^TX)^{-1} \end{aligned}$$

Hence for better Type I error of (non-pivotal) statistics, use modified residuals.

Note that since $\frac{N-p}{N} \to 1$ and $(I-H)I^{-1} \to I$ as $N \to \infty$, for large N, the choice of modified versus raw residuals does not matter.

2.3.3 Use residual for homoscedastic data, but case/score for heteroscedastic data

Liu & Singh (1992) state that bootstrap methods are either type E (efficient) or type R (robust versus heteroscedasticity). They state that residual resampling is type E, while case and score resampling are type R. In the homoscedastic case, the asymptotic variance of R-type is larger than that of E-type, so asymptotic relative efficiency (ARE) is larger than 1, and ARE $-1 \to 0$ if and only if $\text{var}(x_i) \to c$ as $N \to \infty$. However, in the heteroscedastic case, the true variance is different from the limit of the E-type variance estimators, so E-type estimators are inconsistent and R-type estimators are actually $\sqrt{N}$-consistent and have asymptotic normal distribution. Liu

& Singh (1992) show that these results can be extended/applied to general linear regression (e.g. Theorem 5, p379). The variance estimators on page 373 imply the use of the correct formula under heteroscedasticity, where the correct or sandwich formula refers to the form of var($\hat{\beta}$) under heteroscedasticity, and the naive formula refers to the form of var($\hat{\beta}$) if errors are i.i.d.. Refer to Chapter 4 of this thesis.

Thus, residual resampling should not be used if heteroscedasticity is suspected. We expect that in this situation, score resampling should outperform case resampling if the data are unbalanced, that is, if there are outliers in the data. According to Wu (1986), case resampling neglects the unbalanced nature of regression data (meaning, presumably, the presence of outliers and/or influential points). This statement is supported by Hjorth (1994). According to Bose & Chatterjee (2002), for case resampling, asymptotic results require the assumption that an influential design point cannot alter the asymptotics. The reason is heuristically clear: if there is an extreme outlier, results will depend on whether that observation is chosen in the case resampling bootstrap resample. However, unbalanced data should not affect score resampling. Bose & Chatterjee (2002) agree with Liu & Singh (1992) that under homoscedastic errors, type E is more efficient than type R, and they state that since both case and score resampling are type R, and have the same asymptotic properties, they are difficult to compare without assuming extra conditions.

In Chapter 3, we will use simulation to assess these rules.

Chapter 3

Verifying the three bootstrap "rules"

3.1 Introduction

The aim of this Chapter is to verify the three known "rules" of the bootstrap, identified in Chapter 2, via simulation. These three "rules" are:

1. Use pivotal rather than non-pivotal statistics. (For better Type I error.)
2. For residual (and score) resampling, use modified rather than raw residuals. (For better Type I error.)
3. Residual (E-type) versus case/score (R-type): if data are homoscedastic, use residual rather than case/score resampling (residual resampling is more efficient for homoscedastic data), but if data are heteroscedastic, use case/score rather than residual resampling (residual resampling is less robust to heteroscedasticity).

3.2 Expectations

Note that to confirm "rule" 2, we only consider residual resampling, since the same argument for the superiority of using modified residuals over raw residuals applies

to score resampling, and further we only consider null model residual resampling, since we will consider full versus null model residual resampling in Chapter 5.

We expect to find the following simulation results concerning the three known "rules" (note that the reasons for the expectations are provided in Section 2.3):

1. We expect pivotal statistics to have accurate Type I error, and non-pivotal statistics to have inflated Type I error. (The latter at least for small sample size (N).)

2. We expect the non-pivotal statistic for null model residual resampling with raw residuals to have inflated Type I error, while we expect the same but with modified residuals to have more accurate Type I error, at least for small sample size (N). This is because the variance of $\hat{\beta}_*$ is biased for the variance of $\hat{\beta}$, in the sense of being smaller, if raw residuals are used, but the bias is removed if modified residuals are used instead.

3. We expect residual resampling to be more efficient (that is, to have greater size-adjusted power, defined in Section 3.4) than case/score resampling for homoscedastic data, but we expect residual resampling to have less accurate Type I error than case/score resampling for heteroscedastic data.

3.3 Simulation design

We conducted simulations to compare the performance of different test statistics and different resampling methods in different situations. We conducted the simulation on R version 2.9.2. Simulations were computationally intensive, and took a total of around 200 hours computing time on my University desktop machine.

In all our simulations, the true form of the model was $Y = 4 + 3X_0 + \beta_1 X_1 + \epsilon$, where ϵ are the "true errors", where we are testing $H_0 : \beta_1 = 0$, and β_1 took the value 0 for Type I simulations and the value 0.5 for Type II simulations.

For each simulation, we generated 1000 random datasets to estimate error rates and power. We also used 1000 resamples to estimate the significance level for each

dataset. These parameters ensure reasonable accuracy of Type I error and power results based on bootstrap P-values. A total of 18 simulations in a $2\times3\times3$ orthogonal design were conducted, varying the following data properties:

- errors: homoscedastic or heteroscedastic
- X design: regular, normal uncorrelated or normal correlated X, as below.
- sample size: $N = 16, 32, 64$

We consider three X designs: regular, normal uncorrelated and normal correlated X. The definition of the regular X design follows. Let $X_{\text{base}} = (1, 2, 3, 4)^T$, N = length of the response variable = number of rows of X. For $N = 16$, $X = [X_0\ X_1] = [X_{\text{base}} \otimes 1_{4\times1}\ 1_{4\times1} \otimes X_{\text{base}}]$
For $N = 32$, $X = [X_0\ X_1] = [X_{\text{base}} \otimes 1_{8\times1}\ 1_{8\times1} \otimes X_{\text{base}}]$
For $N = 64$, $X = [X_0\ X_1] = [X_{\text{base}} \otimes 1_{16\times1}\ 1_{16\times1} \otimes X_{\text{base}}]$
In the normal uncorrelated design, X was comprised of two independent "random" vectors sampled from a $\mathcal{N}(0, \text{var} = 1.25)$ distribution, and in the normal correlated design, X was comprised of two correlated "random" vectors sampled from a $\mathcal{N}(0, \text{var} = 1.25 \times (1 - 0.8^2)^{-0.25})$ distribution, with a correlation of 0.8. The variances of normal X designs were chosen in such a way that $\det(E(X^TX))$ was the same across the three X designs.

For the homoscedastic simulations, the true errors were defined as $\epsilon \sim \mathcal{N}(0, \text{var} = 4)$, whereas for the heteroscedastic simulations, the true errors were defined as $\epsilon = \max(1, X_1) \times \mathcal{N}(0, \text{var} = 4/a)$ where $a = 7.5$ for regular X, $a = 7.551$ for normal uncorrelated X and $a = 8.413$ for normal correlated X. Note that a was chosen so that $E(a^2 \times \max(1, X_1)^2) = 4$, such that average error variance was the same for heteroscedastic and homoscedastic simulations.

We consider three resampling methods: residual, score and case resampling. Both raw and modified residuals were considered for residual resampling. Two methods of score resampling were considered: $t^* \sim \mathcal{N}(0, 1)$ or t^* randomly chosen from the set of modified residuals with equal probability. Results were very similar and so only the results for standard normal score resampling are presented here.

Note that for score resampling, only modified residuals were considered for resampling. For case resampling, we considered statistics based on using design X and X_* in standard error estimation in the calculation of pivotal statistics, but results were similar and so we chose to present results based on design X_*.

For all three resampling methods, we calculated two types of statistics, non-pivotal and pivotal. The pivotal statistic is $t_* = (\hat{\beta}_* - a)/\hat{\text{se}}_*(\hat{\beta}_*)$, where $a = 0$ for null model residual resampling and $a = \hat{\beta}$ for all other resampling methods, and $\hat{\text{se}}_*(\hat{\beta}_*)$ has the form $\hat{\sigma}_*^2(X^TX)^{-1}$ for residual or score resampling, or the form $\hat{\sigma}_*^2(X_*^TX_*)^{-1}$ for case resampling (although another choice will be considered in Chapter 4). We also define t to be $t = \hat{\beta}/\hat{\text{se}}(\hat{\beta})$, where $\hat{\text{se}}(\hat{\beta})$ has the form $\hat{\sigma}^2(X^TX)^{-1}$, as usual. Then the P-value of the pivotal statistic is defined as $p = \sharp[|t_*| \geq |t|]/B$. The P-value of the non-pivotal statistic is defined as $p = \sharp[|\hat{\beta}_* - a| \geq |\hat{\beta}|]/B$ where $a = 0$ for null model residual resampling and $a = \hat{\beta}$ for all other resampling methods.

To summarize, for each simulation, we studied the properties of 3 resampling methods (residual, case or score) $\times$ 2 resample statistics (non-pivotal or pivotal). We considered both Type I error and power simulations.

3.4 Simulation analyses

This section explains the graphical and numerical analysis methods we have employed when comparing the different resampling methods in particular types of simulations.

Error rates have been estimated using Monte Carlo methods from 1000 datasets. This means that at the 0.05 level, for example, the standard error of the Type I error of an exact test is $\sqrt{p(1-p)/n}$, where $p = 0.05$ and n is the number of observations. To aid in interpretation, we have included 95 percent CI bands for a Type I error of 0.05, within which an exact test would fall 95 percent of the time.

To obtain an overall assessment of the accuracy of Type I error across a set of simulations, we applied a global test to the set of P-values from the Type I simulations of a statistic, testing the hypothesis that in each simulation the number of rejections

at 0.05 level is Binomial(1000,0.05). Arranging the number of rejections/retentions of H_0 in a $2\times k$ contingency table (for k simulations), we can calculate a chi-square statistic, denoted as X^2, with k degrees of freedom, to test if Type I error rate is 0.05 for each simulation. We applied this test across all 9 homoscedastic (or sometimes 9 heteroscedastic) simulations to get an overall sense of whether there was any evidence of departure from a test size of 0.05.

In our Type II simulations, the unadjusted power at level 0.05 is simply the proportion of the P-values generated by the simulation which are less than or equal to 0.05. However, this measure is influenced by the size of a test, so we defined size-adjusted power as the proportion of P-values which are less than or equal to the lower five-percent quantile of the P-values from the corresponding Type I simulation. In this way, power is adjusted to the true size of the corresponding Type I simulation, as suggested in Lloyd (2005).

To compare (size-adjusted) power between two statistics for a Type II simulation, we applied McNemar's test to the two sets of p-values. We used the "mcnemar.test()" function in R, with continuity correction. We also applied McNemar's test across multiple simulations by summing McNemar statistics and degrees of freedom for "global" inference of power difference across multiple simulations. As previously, this was done to make global inference across all 9 homoscedastic simulations, or sometimes, across all 9 heteroscedastic simulations.

3.5 Results

We obtained similar results across X design simulations (but not across homoscedastic versus heteroscedastic simulations), so results were averaged in graphs across the 3 simulation designs (for example, for Type I error graphs, the arithmetical mean of the 3 sizes was plotted at each sample size N), and in reporting global tests of Type I error accuracy and power differences, results were combined across the 9 simulations in each of homoscedastic and heteroscedastic scenarios.

Figure 3.1 shows that for homoscedastic data, in general, for all the resampling

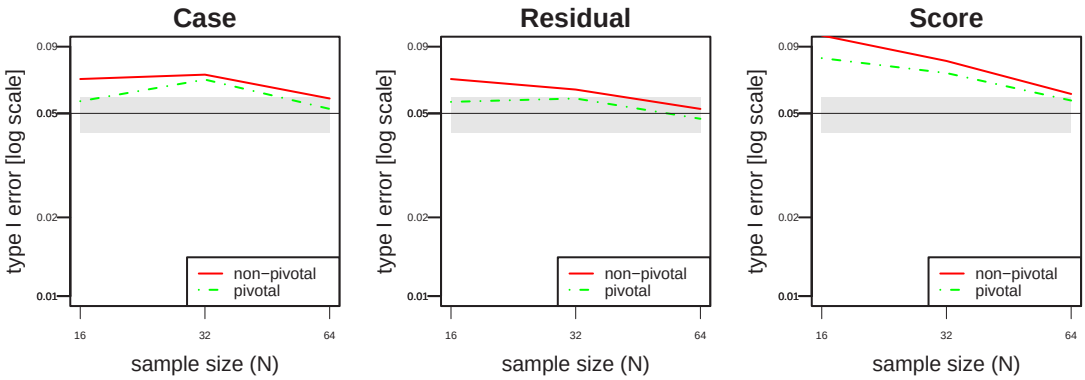

Figure 3.1: A comparison of Type I error of non-pivotal versus pivotal statistics for each of the resampling methods, for combined homoscedastic simulations. Note that in each case, the pivotal statistic is closer to the nominal value (0.05).

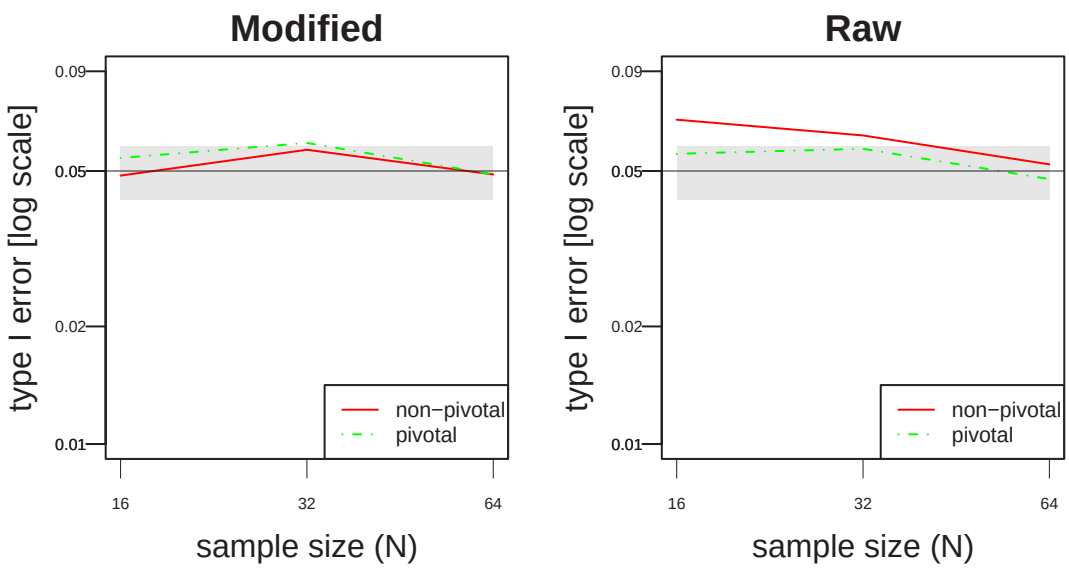

Figure 3.2: A comparison of modified versus raw residual resampling, with regard to Type I error, for homoscedastic data. Note that there is little difference for the pivotal statistics, but the non-pivotal statistic for modified residual resampling has accurate Type I error, while the non-pivotal statistic for raw residual resampling has inflated Type I error

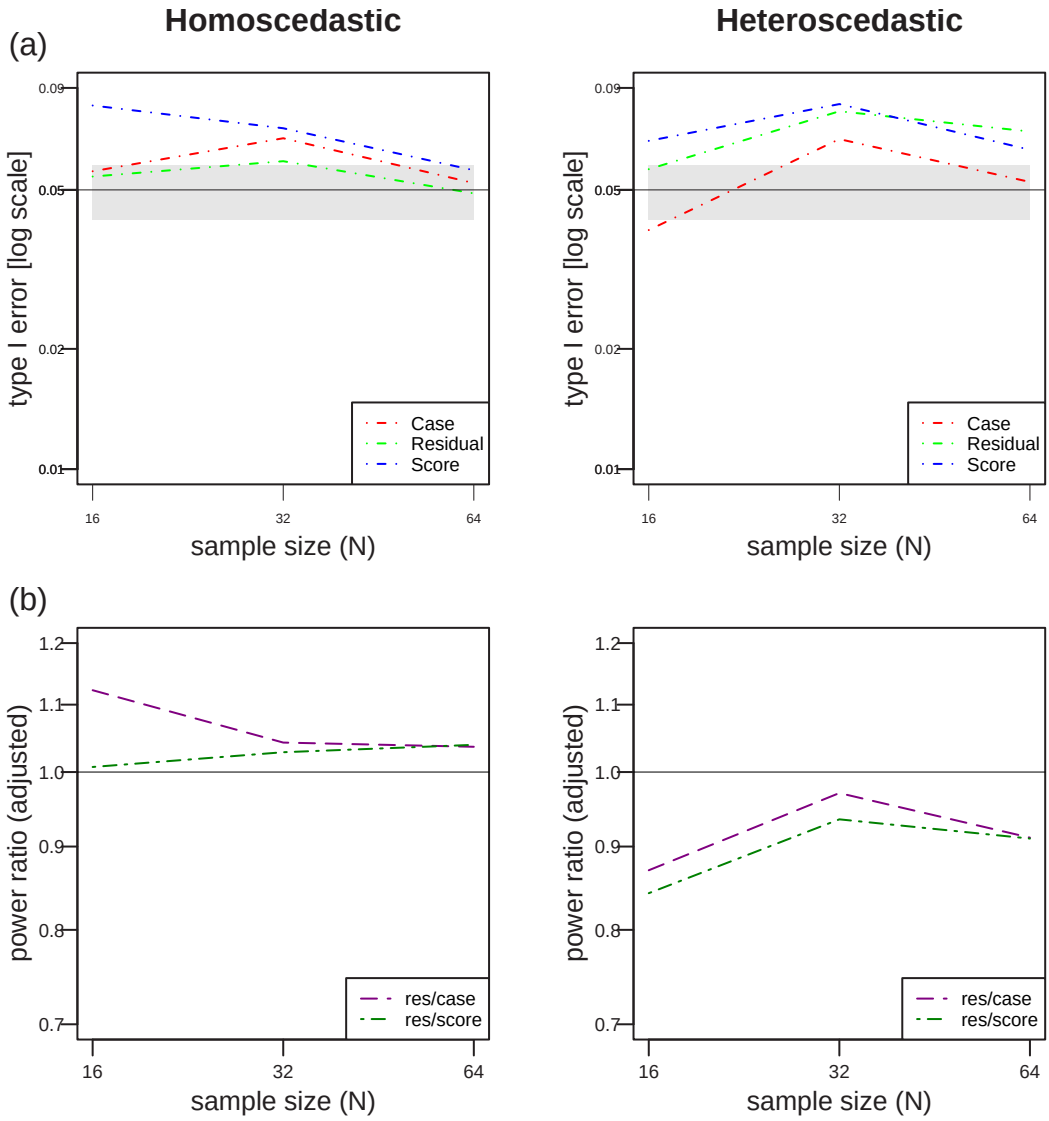

Figure 3.3: A comparison for homoscedastic and heteroscedastic simulations of: (a) Type I error for the pivotal statistics of each resampling method, (b) size-adjusted power (that is, power ratios) of residual versus case resampling and residual versus score resampling (power of residual/power of case and power of residual/power of score). From (a), note that residual resampling has more accurate Type I error for homoscedastic simulations, but less accurate Type I error for heteroscedastic simulations. From (b), note that residual residual resampling has greater power than case/score resampling for homoscedastic simulations (so is more efficient), but has less power for heteroscedastic simulations.

types, pivotal statistics have better Type I error accuracy than non-pivotal statistics. In fact, when we combine results across all 9000 homoscedastic simulation datasets, there is no evidence of significant departure from 0.05 level for the pivotal statistic of residual resampling ($X_9^2 = 8.337$, $p = 0.501$), while there is strong evidence of inflation for the non-pivotal statistic ($X_9^2 = 32.842$, $p = 0.000$). For case resampling, there is evidence of departure from 0.05 level for the pivotal statistic ($X_9^2 = 27.305$, $p = 0.001$), as well as the non-pivotal statistic ($X_9^2 = 58.168$, $p = 0.000$), but there is greater Type I error inflation in the non-pivotal statistic (Figure 3.1). Similarly, for score resampling, there is evidence of departure from 0.05 level for the pivotal statistic ($X_9^2 = 99.326$, $p = 0.000$), as well as the non-pivotal statistic ($X_9^2 = 225.179$, $p = 0.000$), but there is greater Type I error inflation in the non-pivotal statistic (Figure 3.1). This demonstrates the better Type I error characteristics of pivotal statistics for homoscedastic data.

Turning now to the results for heteroscedastic data, when we combine results across all 9000 heteroscedastic simulation datasets, there is evidence of significant departure from 0.05 level for the pivotal statistics for score and case resampling (note that "rule" 3 implies that we do not consider residual resampling for heteroscedastic data in this discussion), but the global test results show that the deviation from 0.05 level is smaller for the pivotal statistics than for the non-pivotal statistics. That is, for case resampling, we have ($X_9^2 = 70.926$, $p = 0.000$) for the pivotal statistic, but ($X_9^2 = 100.190$, $p = 0.000$) for the non-pivotal statistic. Similarly we have, for score resampling ($X_9^2 = 156.168$, $p = 0.000$) for the pivotal statistic, but ($X_9^2 = 283.642$, $p = 0.000$) for the non-pivotal statistic. Thus pivotal statistics also have better Type I error for heteroscedastic data.

Figure 3.2 shows that for the non-pivotal statistic, for residual resampling and homoscedastic data, using modified residuals leads to better Type I error than using raw residuals, at least for small sample sizes. It also shows that for the pivotal statistic, both using modified and raw residuals leads to accurate Type I error. In fact, when we combine results across all 9000 homoscedastic simulation datasets, there is no evidence of significant departure from 0.05 level for the pivotal statistics of both modified ($X_9^2 = 7.368$, $p = 0.599$) and raw ($X_9^2 = 8.337$, $p = 0.501$)

residual resampling, while for the non-pivotal statistics, there is no evidence of significant departure from 0.05 level for the non-pivotal statistic of modified residual resampling ($X_9^2 = 5.663$, $p = 0.773$), while there is evidence of significant departure from 0.05 level for the non-pivotal statistic of raw residual resampling ($X_9^2 = 32.842$, $p = 0.000$). Again, we do not consider heteroscedastic data in this discussion because of rule 3.

Figure 3.3 (a) shows that for homoscedastic data, residual resampling tends to have more accurate Type I error than case and score resampling. In fact, as previously noted, there is no evidence of departure from 0.05 level for the pivotal statistic for residual resampling ($X_9^2 = 7.368$, $p = 0.599$), while there is evidence of departure from 0.05 level for the pivotal statistic for case resampling ($X_9^2 = 27.305$, $p = 0.001$) and score resampling ($X_9^2 = 99.326$, $p = 0.000$). Thus for homoscedastic data, residual resampling has more accurate Type I error.

Figure 3.3 (a) also shows that this pattern does not hold for heteroscedastic data. Residual resampling no longer has accurate Type I error, demonstrating a lack of robustness to heteroscedasticity. Residual resampling ($X_9^2 = 154.463$, $p = 0.000$) tends to be further from nominal size (0.05) than case resampling ($X_9^2 = 70.926$, $p = 0.000$). While residual resampling is not further than score resampling ($X_9^2 = 156.168$, $p = 0.000$) here, in Chapter 4 we will propose an improvement on the pivotal statistic for score resampling, such that it also outperforms residual resampling for heteroscedastic data.

Figure 3.3 (b) shows that for homoscedastic data, residual resampling has greater power than both case and score resampling, after adjusting for size, demonstrating greater efficiency. This is especially true for non-pivotal statistics, but for simplicity, we only present the results for pivotal statistics. At $N = 16$ the residual over case power ratio was approximately 112 percent and the residual over score power ratio was approximately 101 percent. At $N = 32$ the residual over case power ratio was approximately 104 percent and the residual over score power ratio was approximately 103 percent. At $N = 64$ both power ratios were approximately 105 percent. A global McNemar test across all 9 homoscedastic simulations suggests that residual resampling had significantly greater power than case resampling ($X_9^2 = 59.819$,

$p = 0.000$), and also greater power than score resampling ($X_9^2 = 28.939$, $p = 0.001$). So residual resampling seems to have higher size-adjusted power than case and score resampling for homoscedastic simulations. Interestingly, Figure 3.3 (b) also shows that residual resampling has clearly lower size-adjusted power than case and score resampling for heteroscedastic simulations, but this is technically beyond the realm of rule 3.

3.6 Discussion

From the literature, we were able to find three known "rules" for the bootstrap. Simulations have shown the benefit of using pivotal statistics in correcting inflated Type I error for non-pivotal statistics to accurate Type I error, at least for homoscedastic simulations. Simulations have also shown the benefit of using modified residuals rather than raw residuals in residual resampling, at least for the non-pivotal statistic. Finally, simulations have shown the benefit of using residual resampling (rather than case/score resampling) for homoscedastic data, but not for heteroscedastic data.

We expected raw residual resampling to have inflated Type I error (for the non-pivotal statistic) since $(1/N)\sum_{i=1}^{N} r_i^2$ is biased (smaller than) σ^2, while we expected modified residual resampling to have better Type I error (for the non-pivotal statistic) since $(1/N)\sum_{i=1}^{N} m_i^2 = (1/(N-2))\sum_{i=1}^{N} r_i^2 > (1/N)\sum_{i=1}^{N} r_i^2$ is un-biased for σ^2. However, for the pivotal statistic, there is little difference in Type I error accuracy, presumably because of "rule" 1 and because the pivotal statistics for both modified and raw null model residual resampling have standard normal asymptotic distribution.

An interesting pattern in results was that "rule 2", as defined in Section 3.1, only applied for non-pivotal statistics (as expected in Section 3.2, because "rule 1" implies that pivotal statistics generally have accurate size). Using a pivotal statistic appeared to correct for other problems in resampling or construction of test statistic. This pattern will also be seen in Chapter 5, and it suggests that pivoting a test statistic can make a test approximately valid in a range of conditions.

While we have identified the importance of using a pivotal statistic, a question that remains is how to construct one. This question is studied in detail in Chapter 4. Another question that remains is whether residuals from the null model (that is, under H_0) or from the full model (that is, under H_1) should be used for residual resampling, which is studied in detail in Chapter 5.

Chapter 4

"Naive" or "Sandwich" variance estimator?

4.1 Introduction

In Chapter 3, it was shown that pivotal statistics have better Type I error than non-pivotal statistics. A pivotal statistic calculated on a bootstrap sample has the general form $t = (\hat{\beta}_* - a)/\sqrt{\text{var}(\hat{\beta}_*)}$ where $a = \hat{\beta}$ for full model resampling and $a = 0$ for null model resampling (note that we only consider null model residual resampling and not null model case/score resampling). In this Chapter, we focus on the use of pivotal statistics, and consider the question: which standard error estimator should be used when constructing a bootstrap pivotal statistic?

Two choices of estimators of $\text{var}(\hat{\beta}_*)$ are:

- naive $\text{var}(\hat{\beta}_*) = \hat{\sigma}_*^2 (X^T X)^{-1}$
- sandwich $\text{var}(\hat{\beta}_*) = (X^T X)^{-1} X^T \text{diag}(r_{*,i}^2) X (X^T X)^{-1}$

Here X represents the design matrix, $r_{*,i}$ the ith residual, and $\hat{\sigma}_*^2$ the estimated error variance for a bootstrap sample.

The essential difference between these two estimators is that the naive estimator assumes that residuals are homoscedastic, whereas the sandwich estimator does not.

Previously, choice of standard error was considered to be dependent on data. For example, Bose & Chatterjee (2002); Wu (1986); Freedman (1981); Moulton & Zeger (1991); Friedl & Stadlober (1997) recommend the sandwich estimator for robustness to heteroscedasticity. Here, we argue that choice of standard error estimate when bootstrapping depends on *resampling method*.

The aim of this chapter is to determine which variance estimator should be used for each resampling method, and whether this also depends on properties of the data. Details of a new proposal for how it depends on resampling are in Subsection 4.2, and simulation results evaluating this proposal are in Subsection 4.3.

4.2 New proposal

We now propose a "fourth rule" for bootstrapping in regression situations:

> Only use a naive standard error estimator when using a resampling method that resamples residuals independently of explanatory variables. Otherwise, use a sandwich standard error estimator.

Hence for score or case resampling linear models, one should use the "sandwich" variance estimator that does not assume independence of residuals and explanatory variables. But for residual resampling, one should use the "naive" variance estimator.

Note that the proposed rule applies irrespective of actual data properties- that is, the sandwich variance estimator should be used for case/score resampling even for homoscedastic data. We believe this to be a novel result.

The reasoning behind the proposed rule comes from studying the properties of $\hat{\beta}_*$ for different resampling methods, as defined in Davison & Hinkley (1997) and Wu (1986).

Note that in Chapter 2, it was shown that for case and score resampling, the resampling variance takes the form of the sandwich estimator, while for residual resampling, the resampling variance takes the form of the naive estimator. These

Table 4.1: Summary of the properties of the variance estimators of the different resampling methods

Resampling method	$\text{var}(\hat{\beta}_*)$	reference
Score	$(X^TX)^{-1}X^T\text{diag}(r^2)X(X^TX)^{-1}$	Wu (1986)
Case	$(X^TX)^{-1}X^T\text{diag}(r^2)X(X^TX)^{-1} + O(N^{-2})$	Moulton & Zeger (1991)
Residual	$\hat{\sigma}^2(X^TX)^{-1}$	Moulton & Zeger (1991)

results have long been known in the literature (for example, Moulton & Zeger, 1991; Wu, 1986). Hence the novelty of this Chapter is not the standard error results themselves, but their application in bootstrap inference.

The properties of the variance estimators of each resampling method are summarized in Table 4.1.

Note that in deriving the variance results of Chapter 2 (Table 4.1), no assumptions were made about data properties. The variance estimators are conditional on the data and hence true for both homoscedastic and heteroscedastic data.

In Figure 4.1 we use simulation to illustrate the ideas of Table 4.1: that for score resampling, the true variance of $\hat{\beta}_*$ is the sandwich estimated variance (not the naive estimated variance), while for case resampling, the true variance of $\hat{\beta}_*$ is closer to the sandwich estimated variance than the naive estimated variance, and that the sandwich estimated variance approaches the true variance as N increases. These simulations are for homoscedastic data: hence (counterintuitively) the sandwich estimator for case/score resampling is correct even though data are homoscedastic. We also illustrate that for residual resampling, the true variance of $\hat{\beta}_*$ is the naive estimated variance (not the sandwich estimated variance).

Given that the variance of $\hat{\beta}_*$ is well-known (Bose & Chatterjee, 2002; Wu, 1986; Freedman, 1981; Moulton & Zeger, 1991; Friedl & Stadlober, 1997), theory to the effect of our proposed fourth rule has been available in the literature for some time. However, it appears that, when heteroscedasticity is not suspected, the "naive" standard error estimate is currently used for the pivotal statistic for score and case resampling, and equivalently in constructing bootstrap-t confidence intervals, in

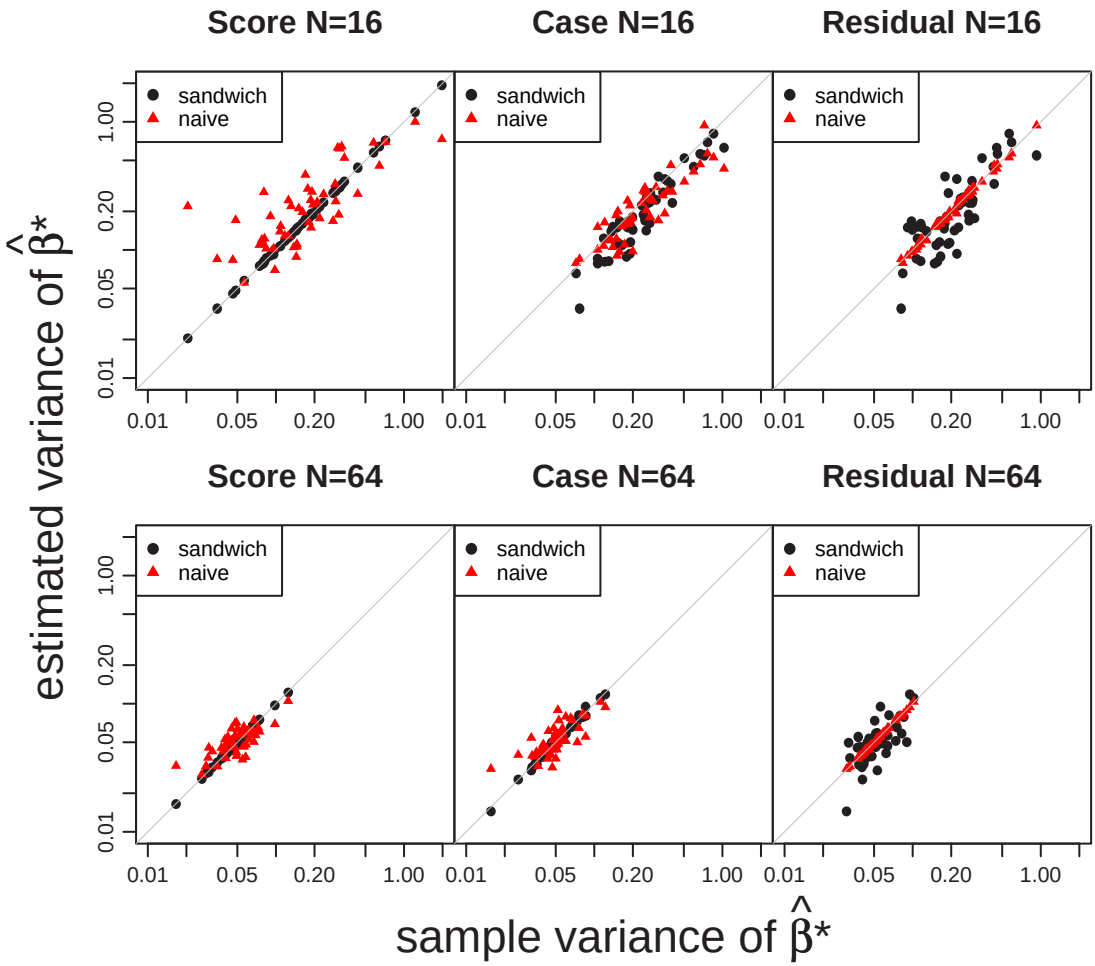

Figure 4.1: A comparison of naive and sandwich estimators to the observed variance of $\hat{\beta}_*$, for different resampling methods. 50 homoscedastic datasets were generated, and the variance of $\hat{\beta}_*$ was estimated from 10000 resamples. Note that for score resampling, $\hat{\text{var}}_{\text{sandwich}}(\hat{\beta}_*) = \text{var}(\hat{\beta}_*)$, for residual resampling $\hat{\text{var}}_{\text{naive}}(\hat{\beta}_*) = \text{var}(\hat{\beta}_*)$, and for case resampling, $\hat{\text{var}}_{\text{sandwich}}(\hat{\beta}_*) \approx \text{var}(\hat{\beta}_*)$ when N is large.

Davison & Hinkley (1997, pages 262,264), Hjorth (1994, page 188), MacKinnon (2006, page S8) and Mammen (1993, pages 268,270).

Here previous literature relates variance estimation to data properties, whereas we argue that the key decision between use of a naive or a sandwich estimator rests with the method of resampling that is used and not on properties of the observed data.

How are the sandwich and naive variance estimators related? We have argued that a sandwich estimator should at times be used in place of the naive estimator, but will this make any difference? In the below, we briefly discuss the relationship between the two estimators.

Under a homoscedastic model, the sandwich and naive estimators are asymptotically equivalent, while under a heteroscedastic model, the sandwich estimator is consistent whereas the naive estimator is not consistent (Freedman, 1981). Hence, for homoscedastic data, for large N, use of sandwich versus naive estimator might not matter. But for small sample sizes, the naive and sandwich estimators can be very different (Figure 4.1). Hence we might expect quite different performance of the two estimators for small samples, or for heteroscedastic data. And we would expect it to make more difference for score resampling than case resampling (based on Figure 4.1).

Another relationship between sandwich and naive estimators is that if we calculate the unconditional expected value of the variances of $\hat{\beta}_*$, assuming homoscedasticity, they equal the naive estimator in all cases. Note that the expectation is for the sampling distribution (with respect to Y), not for the resampling distribution, so in this situation the resampling method that has been applied is irrelevant.

$$
\begin{aligned}
E_Y[\hat{\text{var}}_{\text{sandwich}}(\hat{\beta}_*|Y)] &= (X^TX)^{-1}X^TE_Y[\text{diag}(r_i^2)]X(X^TX)^{-1} \\
&= (X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1} \\
&\quad \text{assuming data are homoscedastic and mod. residuals } m_i \\
&\quad \text{are used} \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}
$$

The final object in the above equation has the form of a naive variance. Note however that it is the actual, that is, conditional variance that is of greater relevance as seen in Figure 4.1, but what this result shows is that we expect sandwich estimators to be centred around the same value as their naive counterparts.

Note that Kauermann and Carroll (2001) have shown that for homoscedastic data, the naive variance estimator is more efficient than the sandwich estimator, in the sense that the variance of the former estimator is less than the variance of the latter estimator. However, for heteroscedastic data, the sandwich estimator is consistent whereas the naive estimator is not. So the sandwich variance estimator gains robustness at the cost of efficiency.

4.3 Simulation results

Simulations were conducted to explore the effect of choice of standard error estimator on Type I error of pivotal statistics.

4.3.1 Simulation design

The same simulation design was used as in Chapter 3. However, we only considered case and score resampling. Also, while we only considered naive pivotal statistics in Chapter 3, we consider both naive and sandwich pivotal statistics in this Chapter. In addition, we do not investigate non-pivotal statistics here, and we do not investigate residual resampling because the sandwich estimator is not appropriate for this resampling method. We only considered Type I error simulations.

While in Chapter 3, we only considered naive statistics for both t_* and t in the construction of pivotal statistics, here we considered sandwich and naive statistics for both t_* and t. So we define $t_{*(\text{naive})} = (\hat{\beta}_* - a)/\hat{\text{se}}_{*(\text{naive})}(\hat{\beta}_*)$ where $a = \hat{\beta}$ for this Chapter. Similarly, we define $t_{*(\text{sandwich})} = (\hat{\beta}_* - a)/\hat{\text{se}}_{*(\text{sandwich})}(\hat{\beta}_*)$ where $a = \hat{\beta}$. Also, $t_{\text{naive}} = \hat{\beta}/\hat{\text{se}}_{\text{naive}}(\hat{\beta})$ and $t_{\text{sandwich}} = \hat{\beta}/\hat{\text{se}}_{\text{sandwich}}(\hat{\beta})$. The pivotal statistic P-value is still defined as $p = \sharp[\text{abs}(t_*) \geq \text{abs}(t)]/B$, but now both t_* and t may take both naive and sandwich form.

Thus, to summarize, for each simulation, we considered 2 resampling methods (case or score) $\times$ 2 resample statistics (naive pivotal or sandwich pivotal) $\times$ 2 sample statistics (naive or sandwich), and we only considered Type I simulations.

As previously, we generated 1000 random datasets to estimate error rates and power. We also used 1000 resamples to estimate the significance level for each dataset. Since we considered three designs for X and three sample sizes N, it follows that the homoscedastic simulations consist of 9000 datasets, and also the heteroscedastic simulations consist of 9000 datasets.

4.3.2 Results

Figure 4.2 displays the Type I error rates for score resampling (Score) and case resampling (Case), averaged across all three homoscedastic designs (Figure 4.2 (a)) and all three heteroscedastic designs (Figure 4.2 (b)). The figure compares the pivotal statistics based on the sandwich and naive variance estimators. There were similar results for all X designs, hence averages are reported here.

Figure 4.2 (a) shows that for homoscedastic simulations, for t_{naive}, for score resampling, the bootstrap pivotal statistic based on the naive estimator ($t_{*(\text{naive})}$) has inflated Type I error, while the bootstrap pivotal statistic based on the sandwich estimator ($t_{*(\text{sandwich})}$) has quite accurate Type I error. In fact, combining results across all 9000 homoscedastic simulation datasets, there is no evidence of significant departure from 0.05 level ($X_9^2 = 11.474$, $p = 0.245$) for the sandwich statistic, in contrast to strong evidence of inflation for the naive statistic ($X_9^2 = 99.326$ $p = 0.000$).

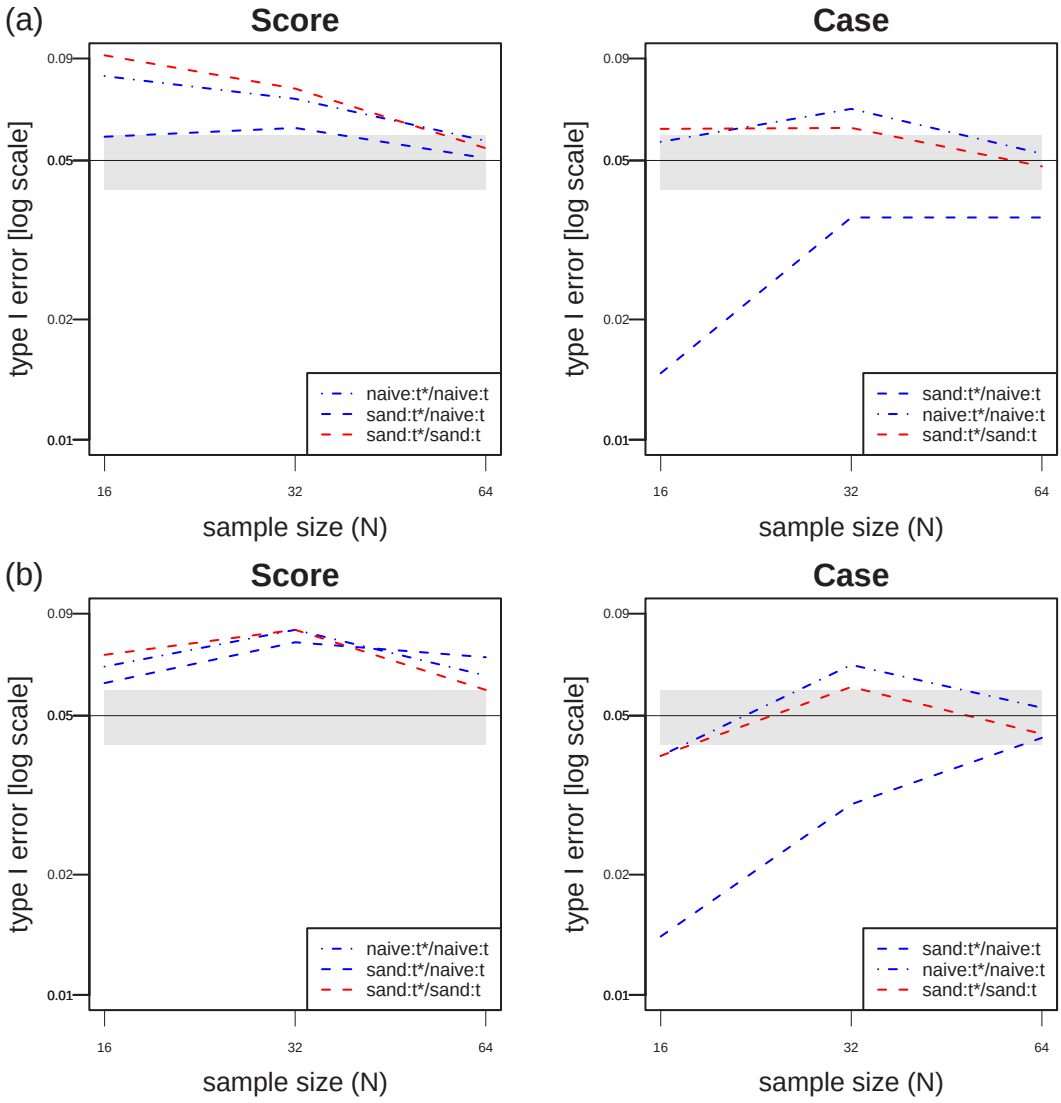

Figure 4.2: A comparison of naive and sandwich estimators for normal score resampling and case resampling for (a) homoscedastic designs and (b) heteroscedastic designs. For both score and case resampling, we consider both t_{naive} and t_{sandwich}. Note that for score resampling and using t_{naive}, the sandwich bootstrap pivotal statistic has more accurate Type I error, but for case resampling, the naive bootstrap pivotal statistic has more accurate Type I error (for t_{naive}). However, if using t_{sandwich}, the sandwich bootstrap pivotal statistic has more accurate Type I error.

Figure 4.2 (a) also shows that for homoscedastic simulations, for case resampling (for t_{naive}), the bootstrap pivotal statistic based on the sandwich estimator ($t_{*(\text{sandwich})}$) has conservative Type I error, while the bootstrap pivotal statistic based on the naive estimator ($t_{*(\text{naive})}$) has surprisingly accurate Type I error. When we combine results across all 9000 homoscedastic simulation datasets, there is strong evidence of departure from 0.05 level ($X_9^2 = 27.305$, $p = 0.001$) for the naive statistic, while there is strong evidence of conservatism for the sandwich statistic ($X_9^2 = 105.810$ $p = 0.000$). In the case of the naive statistic, the magnitude of the departure from 0.05 level appeared to be small (Figure 4.2 (a)).

However, irrespective of whether the sandwich or the naive estimator was used in resampling, the observed test statistic t used the naive estimator in both above cases. If instead the sandwich estimator was used for observed t as well as for t_*, we obtain the opposite result for case resampling. Figure 4.2 (a) shows that for homoscedastic simulations, for case resampling (for t_{sandwich}), the bootstrap pivotal statistic based on the sandwich estimator ($t_{*(\text{sandwich})}$) has generally accurate Type I error. When we combine results across all 9000 homoscedastic simulations, there is strong evidence of a departure from 0.05 level ($X_9^2 = 20.884$, $p = 0.013$), but the magnitude of this departure is small (Figure 4.2 (a)). We do not report results for t_{sandwich} and $t_{*(\text{naive})}$, because we cannot see any situations in which it would make sense to use that approach.

The above results were based on homoscedastic simulations. When we consider the heteroscedastic simulations, the difference between naive and sandwich pivotal statistics disappears for score resampling (with both having inflated Type I error) while the result for case resampling remains the same as previously.

Figure 4.2 (b) shows that for the heteroscedastic simulations, for t_{naive}, for score resampling, the naive and sandwich bootstrap statistics both have inflated Type I error and both perform similarly. There is strong evidence of inflation for both the naive ($X_9^2 = 156.168$ $p = 0.000$) and the sandwich ($X_9^2 = 130.737$ $p = 0.000$) statistics.

Figure 4.2 (b) also shows that for the heteroscedastic simulations, for score resampling, use of a sandwich bootstrap pivotal statistic ($t_{*(\text{sandwich})}$) and a sandwich esti-

mator for observed t (t_{sandwich}) does not lead to accurate Type I error ($X_9^2 = 146.000$ $p = 0.000$). In fact, Figure 4.2 shows that use of a sandwich pivotal statistic with sandwich for observed t does not lead to accurate Type I error even for the homoscedastic simulations ($X_9^2 = 172.695$ $p = 0.000$). This is the opposite result from that for case resampling.

Figure 4.2 (b) shows that for the heteroscedastic simulations, for t_{naive}, for case resampling, the sandwich bootstrap pivotal statistic has conservative Type I error, while the naive bootstrap pivotal statistic has moderately accurate Type I error. Combining results over all 9000 heteroscedastic simulation datasets, there is only moderate evidence of departure from 0.05 level for the naive statistic($X_9^2 = 70.926$ $p = 0.000$), while there is strong evidence of conservatism for the sandwich statistic ($X_9^2 = 117.179$ $p = 0.000$).

If instead, for the heteroscedastic simulations, we use the sandwich estimator for observed t, we again obtain the opposite result. Figure 4.2 (b) again shows that the sandwich bootstrap pivotal statistic has generally accurate Type I error. When we combine results across all 9000 heteroscedastic simulations, there is only moderate evidence of departure from 0.05 level ($X_9^2 = 57.495$, $p = 0.000$). Again, note that it would not make sense to use the sandwich estimator for observed t but the naive estimator in resampling.

4.4 Discussion

In the Chapter, we have proposed a fourth "rule", that for case and score resampling, use of a sandwich estimator in the bootstrap pivotal statistic will improve Type I error. Simulations have shown the benefit of this approach for score resampling, in correcting from inflated Type I error for naive statistic to reasonably accurate Type I error for the sandwich statistic, at least in the homoscedastic simulations. Why this approach did not improve Type I error for heteroscedastic simulations is unclear, and is a possible field for future research. For case resampling, a sandwich estimator did not improve the size accuracy of the test, unless the observed statistic was also of sandwich form. A possible reason for this may be that for case resam-

pling, a correlation model (with random X) applies (Freedman, 1981), in which case the sandwich is the correct sample variance of $\hat{\beta}$ (Seber, 1977). Whereas for score resampling, a regression model (conditional on X) applies (Freedman, 1981; Wu, 1986), so the naive variance is the correct form for $\text{var}(\hat{\beta})$ for homoscedastic simulations.

4.5 Conclusions.

Theory (Table 4.1) suggested that one should use the sandwich variance estimator in the resampled pivotal statistic. Simulations (Figure 4.2) have borne this out to some extent; for score resampling for homoscedastic data, and for case resampling for both homoscedastic and heteroscedastic data, a pivotal statistic using the sandwich variance estimator preserves Type I error closest to 0.05.

An issue however is the choice of standard error estimator for the *observed* test statistic t. In view of our simulation results (Figure 4.2), we recommend the naive estimator for score (conditional on X) and sandwich for case (random X).

Hence we recommend the use of the sandwich estimator for score resampling in general. For case resampling, one should use a sandwich variance estimator in the pivotal statistic if one uses a sandwich estimator of $\text{se}(\hat{\beta})$. Here, we did not discuss power, because theory suggests that method of bootstrapping might affect power more so than method of pivoting.

Chapter 5

Resampling from the full or null model?

In residual resampling, recall that we resample residuals. But in hypothesis testing, there are two types of residuals: those calculated under the full model (assuming H_1 is true) and those calculated under the null model (assuming H_0 is true). A natural question is: do we resample residuals under the full model (under H_1) or the null model (under H_0)?

In this Chapter, we will change notation to be consistent with Anderson & Robinson (2001), a paper which we will follow closely. We write the linear model as:

$$Y = a_1 X + bZ + \epsilon$$

where X and Z are vectors and Y, X and Z have been centred (so no intercept term), and we are testing $H_0 : b = 0$.

Let $R_{Z|X}$ be the residuals from a linear regression of Z on X. A useful "trick" of Anderson & Robinson (2001) is that the above linear model is equivalent to:

$$Y = aX + bR_{Z|X} + \epsilon$$

where we are testing $H_0 : b = 0$. We use X and $R_{Z|X}$ rather than X and Z, because X and $R_{Z|X}$ are orthogonal, which simplifies later working.

Let $R_{Y|XZ}$ be the residuals from a linear regression of Y on both X and Z (or

equivalently on both X and $R_{Z|X}$). Let $R_{Y|X}$ be the residuals from a linear regression of Y on X alone.

In this Chapter, the "true model" is $Y = aX + bR_{Z|X} + \epsilon$, and since we are mainly interested in Type I error (although Subsection 5.4 discusses power properties), we may assume $b = 0$. Also, the "fitted" null model is $Y = \hat{a}X + R_{Y|X}$, and the "fitted" full model is $Y = \hat{a}X + \hat{b}R_{Z|X} + R_{Y|XZ}$.

Hence full model resampling is defined as:

$$Y_{*(\text{full})} = \hat{a}X + \hat{b}R_{Z|X} + R^*_{Y|XZ}$$

and null model resampling is defined as:

$$Y_{*(\text{null})} = \hat{a}X + R^*_{Y|X}.$$

We noted in Chapter 2 that, for full model residual resampling, $E(\hat{\beta}_*) = \hat{\beta}$. In the current notation, for full model residual resampling, $E(\hat{b}_*) = \hat{b}$, and for null model residual resampling, $E(\hat{b}_*) = 0$.

These are both approximations to "true model resampling", defined as:

$$Y_{*(\text{true})} = aX + \epsilon^*_{Y|X}$$

where $\epsilon_{Y|X} = Y - aX$ and a is the coefficient of X in the "true" model (defined above). Since a is generally unknown, true model resampling is a theoretical construct.

We will be primarily interested in $t_{*(\text{null})}$, $t_{*(\text{full})}$ and $t_{*(\text{true})}$, test statistics of the form $\frac{\hat{b}_*}{\sqrt{\hat{\text{var}}_{*(\text{naive})}(\hat{b}_*)}}$ based on $Y_{*(\text{null})}$, $Y_{*(\text{full})}$ and $Y_{*(\text{true})}$, where $\hat{\text{var}}_{*(\text{naive})}(\hat{b}_*)$ is the naive variance estimator of $\hat{b}_*$ for a given bootstrap sample.

The aim of this chapter is to determine what type of residuals (residuals from the null model or residuals from the full model) should be used for resampling. We specifically consider residual resampling, but the same problem arises in score resampling.

This chapter focuses on the properties of pivotal and non-pivotal statistics under full versus null model residual resampling, in particular on the asymptotic correlations of t-statistics generated by the two methods, compared to an ideal "true

bootstrap test" (defined in Appendix B). This work closely follows Anderson & Robinson (2001) who studied properties of different methods of *permutation testing* for linear models. We will extend the results of Anderson & Robinson (2001) to the bootstrapping context.

Since this Chapter mainly deals with asymptotic results, we make the assumption that $\lim_{N\to\infty}(1/N)\sum X^2$ and $\lim_{N\to\infty}(1/N)\sum R^2_{Z|X}$ are of order $O_p(1)$ (Brzezniak & Zastawniak, 1999; Serfling, 1980), and also that ϵ has finite variance σ^2. Since $(N-2)/N \to 1$ as $N \to \infty$, our results hold whether raw or modified residuals are used, but in the following working we use raw residuals.

5.1 Equivalence of r and t statistics under permutation testing

In Anderson & Robinson (2001), π is used to denote permutation, and may be further specified to denote exact, full, or null permutation. For our bootstrap results, we use $*$ to denote bootstrap resampling, i.e. we replace permutation π by bootstrap resample $*$.

Anderson & Robinson (2001) compared three key permutation methods, when using the correlation coefficient as the test statistic:

$$(1) \qquad r^2_{\pi(\text{true})} = \frac{(\sum(Y_{\pi(T)} - a_{\pi(T)}X)R_{Z|X})^2}{\sum(Y_{\pi(T)} - a_{\pi(T)}X)^2 \sum R^2_{Z|X}}$$

where $a_{\pi(T)} = \sum Y_{\pi(T)}X / \sum X^2$ and $Y_{\pi(T)} - a_{\pi(T)}X$ is the residual of $Y_{\pi(T)}$ removing the effect of X and $Y_{\pi(T)} = aX + \epsilon^\pi$. Note that they called it the "exact" method.

$$(2) \qquad r^2_{\pi(\text{null})} = \frac{(\sum(Y_{\pi(F)} - a_{\pi(F)}X)R_{Z|X})^2}{\sum(Y_{\pi(F)} - a_{\pi(F)}X)^2 \sum R^2_{Z|X}}$$

where $a_{\pi(F)} = \sum Y_{\pi(F)}X / \sum X^2$ and $Y_{\pi(F)} = \hat{a}X + R^\pi_{Y|X}$. This statistic was attributed to Freedman & Lane (1983), hence they called it $r^2_{\pi(\text{FreedmanLane})}$.

$$(3) \qquad r^2_{\pi(\text{full})} = \frac{(\sum(R^\pi_{Y|XZ} - k_\pi X)R_{Z|X})^2}{\sum(R^\pi_{Y|XZ} - k_\pi X)^2 \sum R^2_{Z|X}}$$

where $k_\pi = \sum R^\pi_{Y|XZ} X / \sum X^2$ and $R^\pi_{Y|XZ}$ are the permuted least-squares residuals of the full model. This statistic was attributed to ter Braak (1992), hence they called it $r^2_{\pi(\text{terBraak})}$.

Theorem 1. *Let* $f(r) = sign(r)\sqrt{\frac{(N-p)r^2}{1-r^2}}$, *which is a monotonic function. Let* $t = \hat{b}/\hat{se}(\hat{b})$ *and let* $t_{\pi(null)}$ *and* $t_{\pi(full)}$ *be as defined in Appendix B. Then (A)* $t = f(r)$, *(B)* $t_{\pi(null)} = f(r_{\pi(null)})$, *(C)* $t_{\pi(full)} = f(r_{\pi(full)})$, *(D)* $t_{\pi(true)} = f(r_{\pi(true)})$. *Hence the* t*-statistics are equivalent to their* r*-statistic counterparts.*

Proof: see Appendix B.

Anderson & Robinson (2001) found that for the permutation test, the Freedman & Lane (1983) statistic ($r^2_{\pi(\text{null})}$) has asymptotic correlation one with the true method statistic ($r^2_{\pi(\text{true})}$), but the ter Braak (1992) statistic ($r^2_{\pi(\text{full})}$) does not.

In particular, they showed the following result.

The statistics $\sqrt{N} r_{\pi(null)}$, $\sqrt{N} r_{\pi(full)}$, $\sqrt{N} r_{\pi(true)}$ *all converge in distribution to* $\mathcal{N}(0,1)$, *and:*

$$\begin{bmatrix} 1 & 1 & \sqrt{1-r^2} \\ 1 & 1 & \sqrt{1-r^2} \\ \sqrt{1-r^2} & \sqrt{1-r^2} & 1 \end{bmatrix}^{-1} \operatorname{cor}(r_{\pi(\text{true})}, r_{\pi(\text{null})}, r_{\pi(\text{full})}) \to I \tag{5.1}$$

where r *is the partial correlation calculated from the original sample.*

These results were derived by finding expressions for r_{null} and r_{full} in terms of a third quantity, r_{Kennedy}. The method of Kennedy (1995) is an approximation to null model resampling of theoretical interest. The Kennedy (1995) method does not account for all sources of uncertainty, and thus produces inflated Type I error. However, Anderson & Robinson (2001) used r_{Kennedy} as part of the working for relating r_{null} and r_{full}, but in our analogous proofs for the bootstrap case, we relate $t_{*(\text{null})}$ and $t_{*(\text{full})}$ directly.

5.2 Relation between permutation-testing t-statistics

The Delta method can be used to show how $t_{\pi(\text{null})}$, $t_{\pi(\text{full})}$ and $t_{\pi(\text{true})}$ are asymptotically related for permutation testing, building on the Anderson & Robinson (2001) results quoted in the above Subsection.

Corollary 1.

$$\begin{bmatrix} 1 & 1 & \sqrt{1-r^2} \\ 1 & 1 & \sqrt{1-r^2} \\ \sqrt{1-r^2} & \sqrt{1-r^2} & 1 \end{bmatrix}^{-1} cor(t_{\pi(true)}, t_{\pi(null)}, t_{\pi(full)}) \to I$$

Proof. From the Delta Method, we know that:

$$\text{var}(f(x)) \approx f'(x)^2 \text{var}(x)$$

$$\text{cov}(f(x), f(y)) \approx f'(x)f'(y)\text{cov}(x, y)$$

Thus for any monotonic differentiable function f,

$$\begin{aligned} \text{correlation}(f(x), f(y)) &= \text{cov}(f(x), f(y))/\sqrt{\text{var}(f(x))\text{var}(f(y))} \\ &\approx f'(x)f'(y)\text{cov}(x, y)/\sqrt{f'(x)^2\text{var}(x)f'(y)^2\text{var}(y)} \\ &= \text{cov}(x, y)/\sqrt{\text{var}(x)\text{var}(y)} \\ &= \text{correlation}(x, y) \end{aligned}$$

Hence the Anderson & Robinson (2001) result of equation 5.1 applies to t-statistics.

□

This Corollary implies that for permutation testing, the t-statistics for true model and null model permutation have asymptotic correlation one, whereas the t-statistics for null model and full model permutation do not. Anderson & Robinson (2001) argued that null model permutation is in this sense "superior" to full model permutation.

But note that we are interested in the *bootstrap*, considered in the following section.

5.3 Application to the bootstrap

Having shown the asymptotic correlation for permutation t-statistics, we now consider the asymptotic correlation of *bootstrap* t-statistics.

Theorem 2. $t_{*(true)}, t_{*(null)}, t_{*(full)} \xrightarrow{D} \mathcal{N}(0,1)$ *with correlation matrix:*

$$\begin{bmatrix} 1 & 1 & \sqrt{1-r^2} \\ 1 & 1 & \sqrt{1-r^2} \\ \sqrt{1-r^2} & \sqrt{1-r^2} & 1 \end{bmatrix}^{-1} cor(t_{*(true)}, t_{*(null)}, t_{*(full)}) \to I$$

This is the same result as Anderson and Robinson's for r^2-type statistics under permutation testing, which we have already extended to permutation t-statistics using the Delta Method in the previous subsection.

For Theorem 2 however we derived the result from first principles rather than building on the results of Anderson & Robinson (2001).

For the following, let all t-statistics refer to the bootstrap context. An outline of the proof follows. See Appendix B for full details.

First, we use the Central Limit Theorem to show that $\frac{\hat{b}_*}{\text{se}(\hat{b}_*)} \xrightarrow{D} \mathcal{N}(0,1)$ for true, null and full resampling (Appendix B , "Proof of Theorem 2"). Since $\frac{\hat{\text{se}}_*(\hat{b}_*)}{\text{se}(\hat{b}_*)} \xrightarrow{p} 1$ in each case (Appendix B, Lemma 1), we use Slutsky's Theorem to show that $t_* \xrightarrow{D} \mathcal{N}(0,1)$ for true, null and full resampling. An important difference however for $t_{*(\text{full})}$ is that $\text{se}(\hat{b}_{*(\text{full})}) = (1/N)\frac{\sum R^2_{Y|X}}{\sum R^2_{Z|X}}(1-r^2)$ whereas $\text{se}(\hat{b}_{*(\text{null})}) = (1/N)\frac{\sum R^2_{Y|X}}{\sum R^2_{Z|X}}$ (Appendix B, Lemma 1).

Next, we compute the correlation matrix after deriving identities relating $t_{*(\text{null})}$ to both $t_{*(\text{full})}$ and $t_{*(\text{true})}$ (Appendix B, Lemma 2). That is, we first show that $t_{*(\text{null})}$ is related to $t_{*(\text{full})}$ by the identity:

$$t_{*(\text{full})} = t_{*(\text{null})} \times \frac{\hat{\text{se}}_*(\hat{b}_{*(\text{null})})}{\hat{\text{se}}_*(\hat{b}_{*(\text{full})})} - \hat{b}\frac{\sum R^*_{Z|X} R_{Z|X}}{\sum R^2_{Z|X} \times \hat{\text{se}}_*(\hat{b}_{*(\text{full})})}. \tag{5.2}$$

And similarly, $t_{*(\text{null})}$ is related to $t_{*(\text{true})}$ by the identity:

$$t_{*(\text{true})} = t_{*(\text{null})}\frac{\hat{\text{se}}_*(\hat{b}_{*(\text{null})})}{\hat{\text{se}}_*(\hat{b}_{*(\text{true})})} + (\hat{a} - a)\frac{\sum X^* R_{Z|X}}{\hat{\text{se}}_*(\hat{b}_{*(\text{true})})}. \tag{5.3}$$

In computing $\text{cor}(t_{*(\text{null})}, t_{*(\text{full})}) = E(t_{*(\text{null})} t_{*(\text{full})})$, we substituted the following convergence results into equation (5.2), giving the desired result:

$$\frac{\hat{\text{se}}_*(\hat{b}_{*(\text{null})})}{\hat{\text{se}}_*(\hat{b}_{*(\text{full})})}\sqrt{1-r^2} \xrightarrow{p} 1$$

and

$$E(t_{*(\text{null})} \times \frac{\hat{b}\sum R^*_{Z|X} R_{Z|X}}{\sum R^2_{Z|X}\hat{\text{se}}_*(\hat{b}_{*(\text{full})})})\frac{\sqrt{1-r^2}}{r^2} \xrightarrow{p} 1.$$

Hence we derived $\text{cor}(t_{*(\text{null})}, t_{*(\text{full})})\frac{1}{\sqrt{1-r^2}} \xrightarrow{p} 1$.

For $\text{cor}(t_{*(\text{null})}, t_{*(\text{true})}) = E(t_{*(\text{null})} t_{*(\text{true})})$, however, we derived the following convergence results:

$$\frac{\hat{\text{se}}_*(\hat{b}_{*(\text{null})})}{\hat{\text{se}}_*(\hat{b}_{*(\text{true})})} \xrightarrow{p} 1$$

and

$$(\hat{a}-a)E(\frac{(\sum R^*_{Y|X} R_{Z|X})(\sum X^* R_{Z|X})}{\hat{\text{se}}_*(\hat{b}_{*(\text{null})}) \times \hat{\text{se}}_*(\hat{b}_{*(\text{true})})}) \xrightarrow{p} 0.$$

Substituting these into equation (5.3), we derived $\text{cor}(t_{*(\text{null})}, t_{*(\text{true})}) \xrightarrow{p} 1$.

The outline of the proof is now complete.

Note that we used the same strategy (relate $t_{*(\text{null})}$ and $t_{*(\text{full})}$ and then calculate correlation) as Anderson & Robinson (2001), but different detail. Also, some results used in permutation testing ($\sum R^{\pi 2} = \sum R^2$) only apply asymptotically for the bootstrap ($\sum R^{*2}/N \xrightarrow{p} \sum R^2/N$). The end result is identical: as for Anderson & Robinson (2001), $t_{*(\text{null})}$ has asymptotic correlation 1 with $t_{*(\text{true})}$ but $t_{*(\text{full})}$ does not, which Anderson & Robinson (2001) argue makes null model resampling superior.

5.4 Power of full versus null model residual resampling

As a consequence of the above working, we predict that full model residual resampling will have greater power than null model residual resampling, at least for the non-pivotal statistic, which would suggest that full model residual resampling is preferable to null model residual resampling.

The reason we know this is that in Appendix B, Lemma 1:

$$\frac{\mathrm{se}(\hat{b}_{*(\mathrm{full})})}{\mathrm{se}(\hat{b}_{*(\mathrm{null})})}\frac{1}{\sqrt{1-r^2}} \xrightarrow{p} 1$$

meaning that $\mathrm{se}(\hat{b}_{*(\mathrm{full})})$ is smaller than $\mathrm{se}(\hat{b}_{*(\mathrm{null})})$ by a factor of $\sqrt{1-r^2}$ for Type II simulations.

So we predict that when non-pivotal statistics are used, the test based on null model residual resampling is less powerful than that based on full model residual resampling. However, when pivotal statistics are used, since they have the same asymptotic distribution, we expect no difference in power.

5.5 Application from pivotal to non-pivotal statistics

While the results of Anderson & Robinson (2001) for permutation testing apply to pivotal statistics, and our bootstrap results apply also to pivotal statistics, we now show that the correlation result of Theorem 2 also applies to non-pivotal statistics. In fact, as in Chapter 3, differences between resampling methods often occur for non-pivotal rather than pivotal statistics, as will be shown in the simulation sections below.

Corollary 2.

$$\begin{bmatrix} 1 & 1 & \sqrt{1-r^2} \\ 1 & 1 & \sqrt{1-r^2} \\ \sqrt{1-r^2} & \sqrt{1-r^2} & 1 \end{bmatrix}^{-1} cor(\hat{b}_{*(true)}, \hat{b}_{*(null)}, (\hat{b}_{*(full)} - \hat{b})) \to I$$

Proof. The argument is as follows. We showed in the proof of Theorem 2 that

$\frac{\hat{b}_*}{\text{se}(\hat{b}_*)} \xrightarrow{D} \mathcal{N}(0,1)$. Hence:

$$\begin{aligned}
\text{cor}\left(\frac{\hat{b}_{*(\text{null})}}{\text{se}(\hat{b}_{*(\text{null})})}, \frac{\hat{b}_{*(\text{full})}}{\text{se}(\hat{b}_{*(\text{full})})}\right) &= E\left(\frac{\hat{b}_{*(\text{null})}}{\text{se}(\hat{b}_{*(\text{null})})} \times \frac{\hat{b}_{*(\text{full})}}{\text{se}(\hat{b}_{*(\text{full})})}\right) \\
&= E\left(t_{*(\text{null})} \times t_{*(\text{full})} \times \frac{\hat{\text{se}}_*(\hat{b}_{*(\text{null})})}{\text{se}(\hat{b}_{*(\text{null})})} \times \frac{\hat{\text{se}}_*(\hat{b}_{*(\text{full})})}{\text{se}(\hat{b}_{*(\text{full})})}\right) \\
&\xrightarrow{p} E(t_{*(\text{null})} \times t_{*(\text{full})}) \text{ by Slutsky's Theorem} \\
&= \text{cor}(t_{*(\text{null})}, t_{*(\text{full})})
\end{aligned}$$

We finally note that the LHS refers to the correlation of non-pivotal statistics, while the RHS, being the correlation of pivotal statistics, uses the correlation result of Theorem 2. A similar argument also applies for null versus true model resampling.

□

5.6 Simulation results

Simulations were conducted to explore the effect of choice of residual (full/null model) in residual resampling on Type I error and power of pivotal and non-pivotal statistics.

This point is pertinent because we predict that full model resampling will be more computationally efficient than null model resampling if one wishes to test many different null hypotheses with the same alternate hypothesis. Also, our theoretical results suggest that null model resampling should have better Type I properties than full model resampling, based on being closer to the ideal true bootstrap test. With regard to power, our working predicts that full model resampling should have greater power than null model resampling.

5.6.1 Simulation design

The same simulation design is used as explained in Chapter 3, but note that we will only consider the nine homoscedastic simulations here: 3 sample sizes ($N = 16, 32, 64$) $\times$ 3 design matrices (regular, normal uncorrelated, normal correlated).

Because residual resampling is not appropriate for heteroscedastic data (Rule 3), we do not consider heteroscedastic simulations.

As before, raw power is defined as the rate of rejections at level 0.05 for Type II simulations, while adjusted power is defined as the rate of rejections for Type II simulations at the level defined as the lower 5 percent quantile of the corresponding Type I simulation (where we test H_0: $b = 0$), along the lines of Lloyd (2005), as in Chapter 3. Both pivotal and non-pivotal statistics were compared.

5.6.2 Results

Figure 5.1 displays the Type I error rates, the raw power ratios, and the size-adjusted power ratios for null and full model resampling for the homoscedastic simulations (i.e. averaged over the homoscedastic simulations).

Figure 5.1 (a) shows that with regard to Type I error, the pivotal statistics for both full and null model resampling have quite accurate size. In fact, combining results across all 9000 homoscedastic datasets, there is no evidence of significant departure from 0.05 level for either full ($X_9^2 = 6.968$, $p = 0.640$) or null model ($X_9^2 = 7.368$, $p = 0.599$) resampling. So there is little difference in size-accuracy between the pivotal statistics for full and null model resampling.

However, Figure 5.1 (a) shows that with regard to Type I error, the non-pivotal statistic for null model resampling had quite accurate size but the non-pivotal statistic for full model resampling had inflated size. In fact, combining results across all 9000 homoscedastic datasets, there is no evidence of any significant departure from 0.05 level for null model resampling ($X_9^2 = 5.663$, $p = 0.773$), but there is strong evidence of significant departure from 0.05 level for full model resampling ($X_9^2 = 69.011$, $p = 0.000$), with Type I error tending to be inflated by about 50 percent for small sample sizes. So for the non-pivotal statistic, null model resampling has clearly better Type I error than full model resampling.

Figure 5.1 (b) also shows that with regard to raw power, there was no evidence of a difference for the pivotal statistics. In fact combining results across all 9000 homoscedastic datasets, there is no evidence of a significant difference in power

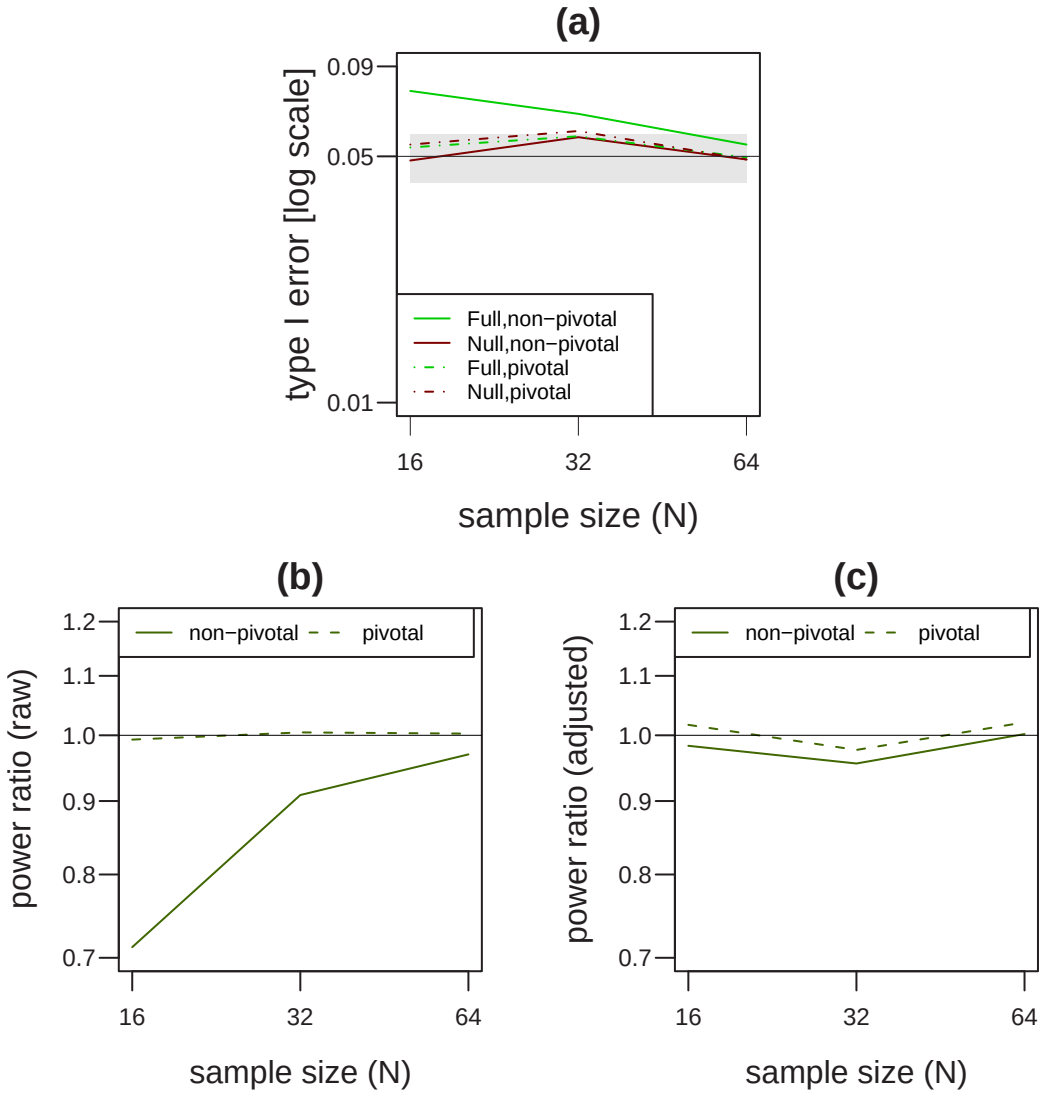

Figure 5.1: A comparison of null versus full model residual resampling, with regard to (a) Type I error, (b) raw power and (c) size-adjusted power. Ratios of power are presented (power for null/power for full) for clarity. Note that: (a) only the non-pivotal statistic for full model resampling has inaccurate size; (b) full model resampling has greater raw power for the non-pivotal statistic; (c) after size correction the power increase is very small.

($X_9^2 = 2.122$, $p = 0.989$).

However, Figure 5.1 (b) shows that the non-pivotal statistic for full model resampling had significantly greater raw power than for null model resampling. In fact combining results across all 9000 homoscedastic datasets, there is strong evidence of significant difference in power ($X_9^2 = 250.328$, $p = 0.000$). At $N = 16$ the null over full power ratio was approximately 72 percent, while at $N = 32$, the null over full power ratio was approximately 90 percent. This is not unexpected given the higher Type I error rate of full model resampling.

When adjusting for differences in Type I error, the power advantage of full model resampling almost disappeared, as in Figure 5.1 (c). Overall, there was a slight increase in power (At $N = 16$, null/full was approximately 99 percent, at $N = 32$, null/full was approximately 96 percent) which was however significant ($X_9^2 = 40.648$, $p = 0.000$). But while there were significant differences, there was no general pattern: for example, for the normal uncorrelated simulations, the size-adjusted powers for $N = 16, 32, 64$ were:

full model, non-pivotal: 0.109, 0.313, 0.547

null model, non-pivotal: 0.117, 0.281, 0.555

Thus for these simulations, full model resampling had slightly higher power (average for full: 0.323, average for null: 0.318), but null model resampling had higher power for $N = 16, 64$. These small differences in power look to be due to sample variation, but were significant on McNemar's test ($X_3^2 = 32.73693$, $p = 0.000$). The reason we think this happened is that adjusted P-values are calculated as a function of a sample quantile from Type I simulations (proportion of P-values less than or equal to the lower 5 percent quantile), which introduces sample variation not accounted for in McNemar's test.

Hence the difference in raw power for non-pivotal statistics observed in Figure 5.1 (b) is most likely entirely due to Type I error inflation of full model resampling.

5.7 New theorem for non-pivotal statistics

Our simulations showed that with regard to Type I error, the pivotal statistics of both null and full model resampling had accurate size, while the non-pivotal statistic of null model resampling had accurate size, but the non-pivotal statistic of full model resampling had inflated size. This result was not expected for the pivotal statistics, given our asymptotic correlation results. To explain this, we re-considered the situation, and we now present a Theorem which proposes that Type I error accuracy depends not on the asymptotic correlation with the ideal true bootstrap test, but on the asymptotic (marginal) distribution of the test statistic.

Theorem 3. *Under $H_0 : b = 0$, while $t_{*(null)}, t_{*(full)} \xrightarrow{D} t$ and $\frac{\hat{b}_{*(null)}}{se(\hat{b})} \xrightarrow{D} \frac{\hat{b}}{se(\hat{b})}$, $\frac{\hat{b}_{*(full)}-\hat{b}}{se(\hat{b})}\frac{1}{\sqrt{1-r^2}} \xrightarrow{D} \frac{\hat{b}}{se(\hat{b})}$.*
Hence, under $H_0 : b = 0$, while the resampling distributions of t_* and $\frac{\hat{b}_{*(null)}}{se(\hat{b})}$ are consistent for the null distribution of t and $\frac{\hat{b}}{se(\hat{b})}$ respectively, the distribution of the non-pivotal statistic under full model resampling differs from its desired null distribution by a factor of $\sqrt{1-r^2}$.

Proof. Note that Theorem 3 only applies under the condition $H_0 : b = 0$.

We know from the Central Limit Theorem (and Slutsky's Theorem) that t has asymptotic standard normal distribution, while we showed in the proof of Theorem 2 that $t_{*(\text{null})}$ and $t_{*(\text{full})}$ also have asymptotic standard normal distribution, that is $t_{*(\text{null})}, t_{*(\text{full})} \xrightarrow{D} t$.

However, the situation is different for the non-pivotal statistic. We know from the Central Limit Theorem that $\frac{\hat{b}}{\text{se}(\hat{b})}$ has asymptotic standard normal distribution. We now show that $\frac{\hat{b}_{*(\text{null})}}{\text{se}(\hat{b})}$ has asymptotic standard normal distribution, while $\frac{\hat{b}_{*(\text{full})}-\hat{b}}{\text{se}(\hat{b})} \times \frac{1}{\sqrt{1-r^2}}$ has asymptotic $\mathcal{N}(0,1)$ distribution.

In the following, we use the result (R1) that $\text{var}(\hat{b}) = \sigma^2/\sum R^2_{Z|X}$, from Seber (1977), but adjusted to the notation of this chapter. We also use Result 3 and Lemma 1 of Appendix B, see this appendix for more details. Now, for null model

resampling:

$$\begin{aligned}
\frac{\hat{\text{var}}_*(\hat{b}_{*(\text{null})})}{\text{var}(\hat{b})} &\xrightarrow{p} \frac{(1/N)\sum R^2_{Y|X}/\sum R^2_{Z|X}}{\text{var}(\hat{b})} \text{ Lemma 1 Part 1} \\
&= \frac{(1/N)\sum R^2_{Y|X}/\sum R^2_{Z|X}}{\sigma^2/\sum R^2_{Z|X}} \text{ from (R1)} \\
&\xrightarrow{p} \frac{(1/N)\sum \epsilon^2_{Y|X}}{\sigma^2} \text{ Result 3} \\
&\xrightarrow{p} 1 \text{ by Weak Law of Large Numbers}
\end{aligned}$$

Noting that $\frac{\hat{b}_{*(\text{null})}}{\text{se}(\hat{b})} = t_{*(\text{null})} \times \frac{\hat{\text{se}}_*(\hat{b}_{*(\text{null})})}{\text{se}(\hat{b})}$, and using Slutsky's Theorem, it follows that since $t_{*(\text{null})}$ has an asymptotic standard normal distribution, so does $\frac{\hat{b}_{*(\text{null})}}{\text{se}(\hat{b})}$. Hence $\frac{\hat{b}_{*(\text{null})}}{\text{se}(\hat{b})} \xrightarrow{D} \frac{\hat{b}}{\text{se}(\hat{b})}$.

However, for full model resampling:

$$\begin{aligned}
\frac{\hat{\text{var}}_*(\hat{b}_{*(\text{full})} - \hat{b})}{\text{var}(\hat{b})}\frac{1}{1-r^2} &\xrightarrow{p} \frac{(1/N)\sum R^2_{Y|X}/\sum R^2_{Z|X}}{\text{var}(\hat{b})} \text{ Lemma 1 Part 2} \\
&= \frac{(1/N)\sum R^2_{Y|X}/\sum R^2_{Z|X}}{\sigma^2/\sum R^2_{Z|X}} \text{ from (R1)} \\
&\xrightarrow{p} \frac{(1/N)\sum \epsilon^2_{Y|X}}{\sigma^2} \text{ Result 3} \\
&\xrightarrow{p} 1 \text{ by Weak Law of Large Numbers}
\end{aligned}$$

Noting that $\frac{\hat{b}_{*(\text{full})}-\hat{b}}{\text{se}(\hat{b})} = t_{*(\text{full})} \times \frac{\hat{\text{se}}_*(\hat{b}_{*(\text{full})}-\hat{b})}{\text{se}(\hat{b})}$, and using Slutsky's Theorem, it follows that, since $t_{*(\text{full})}$ has an asymptotic standard normal distribution, $\frac{\hat{b}_{*(\text{full})}-\hat{b}}{\text{se}(\hat{b})}\frac{1}{\sqrt{1-r^2}}$ has an asymptotic $\mathcal{N}(0,1)$ distribution. Hence $\frac{\hat{b}_{*(\text{full})}-\hat{b}}{\text{se}(\hat{b})}\frac{1}{\sqrt{1-r^2}} \xrightarrow{D} \frac{\hat{b}}{\text{se}(\hat{b})}$.

We have now derived all the asymptotic distribution results necessary for the proof. □

A consequence of the above Theorem is that we expect the pivotal statistics of both null and full model resampling to have accurate Type I error, while we expect the non-pivotal statistic for null model resampling to have accurate Type I error, but the non-pivotal statistic for full model resampling to have inflated Type I error. Since this was observed in the simulations, we believe that size-accuracy depends more on the asymptotic (marginal) distribution of the test statistic, and less on the correlation of the test statistic with the ideal true bootstrap method.

The argument follows. The definition of the P-values for the pivotal statistics of null and full model resampling are $\sharp(|t_{*(\text{null})}| \geq |t|)/B$ and $\sharp(|t_{*(\text{full})}| \geq |t|)/B$, where t is the observed test statistic. Accurate Type I error is achieved if the LHS and RHS of the inequalities have the same (asymptotic) distribution. We showed in Theorem 3 that this condition holds, so we expect accurate Type I error in this case.

The definition of the P-values for the non-pivotal statistics of null and full model resampling are:

$$\begin{aligned} P_{*(\text{null})} &= \sharp(|\hat{b}_{*(\text{null})}| \geq |\hat{b}|)/B \\ &= \sharp\left(\frac{|\hat{b}_{*(\text{null})}|}{\text{se}(\hat{b})} \geq \frac{|\hat{b}|}{\text{se}(\hat{b})}\right)/B \end{aligned}$$

And:

$$\begin{aligned} P_{*(\text{full})} &= \sharp(|\hat{b}_{*(\text{full})} - \hat{b}| \geq |\hat{b}|)/B \\ &= \sharp\left(\frac{|\hat{b}_{*(\text{full})} - \hat{b}|}{\text{se}(\hat{b})} \geq \frac{|\hat{b}|}{\text{se}(\hat{b})}\right)/B \end{aligned}$$

From Theorem 3, we know that the LHS of the inequality of the definition of the P-value of the non-pivotal statistic for null model resampling, has the same asymptotic distribution as the RHS of the inequality, so we expect accurate Type I error.

But for full model resampling, from Theorem 3, we know that the LHS of the inequality of the definition of the P-value of the non-pivotal statistic for full model resampling does not have the same asymptotic distribution as the RHS of the inequality. Noting that the variance of the LHS of the inequality is less than the variance of the RHS of the inequality, since $1 - r^2 < 1$, we expect inflated Type I error. These observations are consistent with what we saw via simulation.

5.8 Discussion

As expected, full model resampling had inflated Type I error for the non-pivotal statistic. This is predicted by theory since the (asymptotic) variance of $\hat{b}_*$ for full

model resampling is less than that for null (and the "near-exact" true) model resampling. Contrary to expectation, we did not find the expected Type I error advantage of null over full, for pivotal statistics. We believe that there is negligible practical difference, having found none in 9000 simulation datasets, and given that this was also found for permutation testing by Anderson & Robinson (2001) and Anderson & Legendre (1999).

We propose that having a correct (marginal) asymptotic distribution may be more important than having better correlation with an exact test, in determining size accuracy. It was differences in asymptotic marginal distribution that explains the pattern in the simulation results (with null and full model resampling different only for the *non-pivotal* statistic), not asymptotic correlation.

A problem with our approach, and that of Anderson & Robinson (2001), is that for Type I simulations, $r \to 0$ as $N \to \infty$. Hence the factor of $\sqrt{1-r^2}$, the key point of difference between null and full model resampling in Theorems 2 and 3, disappears in large samples: $\mathrm{cor}(t_{*(\mathrm{null})}, t_{*(\mathrm{full})}) \to \sqrt{1-r^2} \to 1$ and $\hat{b}_{*(\mathrm{full})} - \hat{b} \to \sqrt{1-r^2}\hat{b} \to \hat{b}$, so it could be argued that these theorems do not imply any difference in asymptotic properties between null and full model resampling. Perhaps future research may involve deriving the order of approximation of Type I error for the non-pivotal statistics of null and full model resampling, as it is likely that the $\sqrt{1-r^2}$ factor in $\hat{b}_{*(\mathrm{full})} - \hat{b}$ implies a lower order of approximation in the full model resampling case.

Finally, there seems to be little difference in size-adjusted power for the two methods, for both pivotal and non-pivotal statistics. Our argument predicted greater power for full model resampling, but a possible reason as stated above is that null and full resampling have similar asymptotic properties. Another possible reason is that the argument was based on the relative size of $\hat{\mathrm{se}}_*(\hat{b}_{*(\mathrm{full})})$ and $\hat{\mathrm{se}}_*(\hat{b}_{*(\mathrm{null})})$, which can evidently be handled via size-correction.

5.9 Conclusions

If using pivotal statistics, you can use *either* full or null model residual resampling. In other settings beyond linear models where pivoting may be difficult, null model resampling is advised.

Chapter 6

Conclusions

The objective of the thesis was to establish "rules" for the bootstrapping of linear models: to determine which resampling method is optimal for each situation.

The existing literature on bootstrapping linear models was reviewed, and three "rules" were found in the literature. We confirmed these via simulation. We also identified two outstanding issues. Firstly, which variance estimator should be used when constructing a bootstrap test statistic? Secondly, if resampling residuals, should this be done using the model that was fitted under the null hypothesis ("null model") or under the alternative hypothesis ("full model")? To our knowledge, these two questions have not been previously addressed. We provided theoretical results to answer these questions, and subsequently confirmed these via simulation. Our simulations were designed to evaluate both the size and (size-adjusted) power characteristics of the proposed bootstrap schemes.

We proposed the use of a sandwich variance estimator for case and score resampling, rather than the naive statistic that is commonly used in practice. Via simulation, we showed that bootstrap test statistics using the sandwich estimator tend to have superior Type I error for case and score resampling, but there was still an issue of which estimator (naive or sandwich) to use for the observed test statistic (t). Best results were achieved when using t-naive for score resampling and t-sandwich for case resampling. One possible explanation for this result is that score resampling conditions on X whereas case resampling does not, and instead treats

X as random.

We also studied full versus null model residual resampling. We showed that null model resampling has better Type I error in theory, having an asymptotic correlation of one with a "true bootstrap" procedure, analogous to a result derived in the permutation testing case by Anderson & Robinson (2001). However in practice, this superiority holds only for non-pivotal statistics: for pivotal statistics, both null and full model resampling had accurate Type I error. We showed that the reason for this is to do with the asymptotic distribution of the test statistic: it appears that the critical issue for test size is the (marginal) asymptotic distribution of the statistic and not its correlation with an ideal test.

When this thesis was reviewed, an examiner mentioned a method described in Huh & Jhun (2001). While they focussed on permutation testing, they also briefly discussed the bootstrap. We now briefly discuss this method and its advantages. The key to their method is to use a N by $(N-p)$ matrix of orthogonal columns V, where V is defined to satisfy $VV^T = I_N - X(X^TX)^{-1}X^T = I - H$. The key difference between our residual resampling method and theirs is that they resample V^Tr rather than r. In the case where the null hypothesis is not that all $\beta_k = 0$, they suggest resampling $V^Tr = V^T(Y - X\hat{\beta})$ giving $Y_* = X\hat{\beta} + V(V^Tr)^*$. Therefore, the method of Huh & Jhun (2001) is to use null model residual resampling with V^Tr instead of r.

The two differences between their method and our thesis method are that we standardized residuals to have equal variance, whereas they standardized them to have equal variance and be uncorrelated (which is an advantage), and that they back-transform their resampled residuals to re-introduce unequal variance and correlation, where we did not (which is also an advantage). Interestingly, the simulations of Huh & Jhun (2001) show that their method compares well with the bootstrap pivotal test, thus showing no improvement over the null model residual resampling bootstrap pivotal statistic. Thus we emphasize that their method shows no improvement over the pivotal bootstrap statistics considered in the thesis.

The thesis considered linear models with two explanatory variables, where the aim is to test a hypothesis about the relationship of one explanatory variable with

the response, that is, a hypothesis test concerning a one-dimensional parameter. A slightly more general model was considered in the investigation of variance estimators, although the hypothesis test still involved a one-dimensional parameter. We expect that our results apply more generally to models containing multi-dimensional parameter components, but we have not done theory or simulation for this case. This would be one area in which the research could usefully be extended. Another limitation was that in our power comparisons, we dealt with the problem of comparing methods with different Type I error by adjusting for test size, but other methods could be investigated (Lloyd, 2005).

Another possible field of further study is extending the results to generalized linear model resampling. In this context the definition of residuals is less straightforward, so the issue of which type of residual to use would need to be addressed.

Appendix A

Properties of $\hat{\beta}_*$ under case resampling

In this Appendix, we derive some properties of $\hat{\beta}_*$ under case resampling, originally due to Freedman (1981) and Moulton & Zeger (1991).

A.1 Asymptotic distribution of $\hat{\beta}_*$

We wish to prove the result from Freedman (1981) that under case resampling:

$$\sqrt{N}(\hat{\beta}_* - \hat{\beta}) \xrightarrow{d} \mathcal{N}(0, J^{-1}MJ^{-1})$$

where J and M are defined as:

$J = E(X_{*,i}^T X_{*,i})$

$M = E(X_{*,i}^T X_{*,i} r_{*,i}^2)$

An outline of the convergence part of the proof follows.

$$\begin{aligned}
1.\ \sqrt{N}(\hat{\beta}_* - \hat{\beta}) &= \sqrt{N}((X_*^T X_*)^{-1} X_*^T Y_* - (X_*^T X_*)^{-1} X_*^T X_* \hat{\beta}) \\
&= \sqrt{N}(X_*^T X_*)^{-1} X_*^T (Y_* - X_* \hat{\beta}) \\
&= \sqrt{N}(X_*^T X_*)^{-1} X_*^T r_* \\
&= ((1/N) X_*^T X_*)^{-1} . ((1/\sqrt{N}) X_*^T r_*) \\
&= W_{*(p\times p)}^{-1} . Z_{*(p\times 1)} \qquad \text{(A.1)}
\end{aligned}$$

where $W_* = (1/N)X_*^T X_*$ and $Z_* = (1/\sqrt{N})X_*^T r_*$.

2. Now $E(W_*) = J$ and W_* is a sample mean, so by the Weak Law of Large Numbers,

$$W_* \xrightarrow{p} J \tag{A.2}$$

3. We derive the expected value and variance of Z_*:

$$\begin{aligned} E(Z_*) &= (1/\sqrt{N})\sum_{i=1}^{N} E(X_{*,i}^T r_{*,i}) \\ &= \sqrt{N}(1/N)\sum_{i=1}^{N} X_i^T r_i \\ &= 0 \text{ by orthogonality} \end{aligned}$$

Also:

$$\begin{aligned} \text{var}(Z_*) &= (1/N)\sum_{i=1}^{N} E(X_{*,i}^T X_{*,i} r_{*,i}^2) \\ &= E(X_{*,i}^T X_{*,i} r_{*,i}^2) \\ &= M \end{aligned}$$

And so by the Central Limit Theorem:

$$Z_* = (1/\sqrt{N})X_*^T r_* \xrightarrow{D} \mathcal{N}(0, M) \tag{A.3}$$

4. From equation (A.1),

$$\begin{aligned} \sqrt{N}(\hat{\beta}^* - \hat{\beta}) &= W_*^{-1} Z_* \\ & \quad \text{and from equations (A.2) and (A.3),} \\ &\xrightarrow{D} J^{-1}\mathcal{N}(0, M) \\ &\sim \mathcal{N}(0, J^{-1} M J^{-1}) \end{aligned}$$

using Slutsky's Theorem (Serfling, 1980).

A.2 Proof of Moulton and Zeger result for case resampling

We want to prove the result from Moulton & Zeger (1991) that:

$$\mathrm{var}_{\mathrm{case}}(\hat{\beta}_*) = (X^T X)^{-1} X^T \mathrm{diag}(r^2) X (X^T X)^{-1} + O(N^{-2})$$

We know that, from the Central Limit Theorem:

$$\begin{aligned} A_* &:= \frac{1}{N}(X_*^T X_*) \\ &= \frac{1}{N}(X^T X) + O_p(N^{-1/2}) \end{aligned}$$

Because A_* has the form of a mean of N terms. Since A_* is a matrix, we mean that the (i,j)th element of A_* converges to the (i,j)th element of $(1/N)(X^T X)$ at rate $O_p(N^{-1/2})$.

Also, by the Central Limit Theorem:

$$\begin{aligned} B_* &:= \frac{1}{N} X_*^T \mathrm{diag}(r_*^2) X_* \\ &= \frac{1}{N} X^T \mathrm{diag}(r^2) X + O_p(N^{-1/2}) \end{aligned}$$

Because B_* has the form of a mean of N terms. Note that the same meaning of stochastic convergence of a matrix applies, as in the note attached to the corresponding equation for A_*.

In order to apply the Central Limit Theorem, we must show that A_* and B_* have finite variance. But A_* may be written as $(1/N)\sum_i X_{*,i}^T X_{*,i}$, where $X_{*,i}$ is the ith row of X_*, and we can assume that $X_{*,i}^T X_{*,i}$ has finite variance. Similarly, B_* may be written as $(1/N)\sum_i X_{*,i}^T r_{*,i}^2 X_{*,i}$, and we can assume that $X_{*,i}^T r_{*,i}^2 X_{*,i}$ has finite variance. Thus the use of the Central Limit Theorem is justified in this case.

Then:

$$\begin{aligned} S &:= \frac{1}{N} A_*^{-1} B_* A_*^{-1} \\ &= (1/N) N (X_*^T X_*)^{-1} (1/N) X_*^T \mathrm{diag}(r_*^2) X_* N (X_*^T X_*)^{-1} \\ &= (X_*^T X_*)^{-1} X_*^T \mathrm{diag}(r_*^2) X_* (X_*^T X_*)^{-1} \\ &= \mathrm{var}_{\mathrm{case}}(\hat{\beta}_*) \end{aligned}$$

But:

$$\begin{aligned}
S &:= \frac{1}{N} A_*^{-1} B_* A_*^{-1} \\
&= \frac{1}{N}[N(X^T X)^{-1} + O(N^{-1/2})][\frac{1}{N} X^T \mathrm{diag}(r^2) X + O(N^{-1/2})][N(X^T X)^{-1} + O(N^{-1/2})] \\
&= (X^T X)^{-1} X^T \mathrm{diag}(r^2) X (X^T X)^{-1} + (1/N) O(N^{-1}) \\
&= (X^T X)^{-1} X^T \mathrm{diag}(r^2) X (X^T X)^{-1} + O(N^{-2})
\end{aligned}$$

From which we deduce that:

$$\mathrm{var}_{\mathrm{case}}(\hat{\beta}_*) = (X^T X)^{-1} X^T \mathrm{diag}(r^2) X (X^T X)^{-1} + O(N^{-2})$$

.

Appendix B

Proofs of Chapter 5 results

In this appendix, we present proofs of asymptotic relations between null and full t_* of Chapter 5.

We relate null and full model resampling *directly*, in contrast to Anderson & Robinson (2001), who related them via the Kennedy (1995) method for permutation testing. Note, however, that the extension from pivotal to non-pivotal statistics still applies.

B.1 Definitions

Recall that under Chapter 5 notation, the true model being assumed is: $Y = aX + bR_{Z|X} + \epsilon$ where X and Y are centred.

We will relate test statistics for three resampling methods: full, null and true, defined as follows:

Full model:

$$Y = \hat{a}X + \hat{b}R_{Z|X} + R_{Y|XZ} \tag{B.1}$$

Full model residual resampling:

$$\begin{aligned} Y_{*(\text{full})} &= \hat{a}X + \hat{b}R_{Z|X} + R^*_{Y|XZ} && \text{(B.2)} \\ &= \hat{a}_{*(\text{full})}X + \hat{b}_{*(\text{full})}R_{Z|X} + R_{Y_*|XZ} && \text{(B.3)} \end{aligned}$$

Null model: $Y = \hat{a}X + R_{Y|X}$

Null model residual resampling:

$$Y_{*(\text{null})} = \hat{a}X + R^*_{Y|X} \tag{B.4}$$

$$= \hat{a}_{*(\text{null})}X + \hat{b}_{*(\text{null})}R_{Z|X} + R_{Y_*|XZ} \tag{B.5}$$

True model: $Y = aX + \epsilon_{Y|X}$

True model residual resampling:

$$Y_{*(\text{true})} = aX + \epsilon^*_{Y|X}$$

$$= \hat{a}_{*(\text{true})}X + \hat{b}_{*(\text{true})}R_{Z|X} + R_{Y_*|XZ}$$

The pivotal statistics for true, null and full model residual resampling are defined as:

$$t_{*(\text{full})} = \frac{\hat{b}_{*(\text{full})} - \hat{b}}{\hat{\text{se}}_*(\hat{b}_{*(\text{full})})}$$

Similarly,

$$t_{*(\text{null})} = \frac{\hat{b}_{*(\text{null})}}{\hat{\text{se}}_*(\hat{b}_{*(\text{null})})}$$

And:

$$t_{*(\text{true})} = \frac{\hat{b}_{*(\text{true})}}{\hat{\text{se}}_*(\hat{b}_{*(\text{true})})}$$

And in each case, $\hat{\text{se}}_*(\hat{b}_*) = \sqrt{\frac{\sum R^2_{Y_*|XZ}}{(N-2)\sum R^2_{Z|X}}}$.

Since Theorem 2 concerns asymptotic results, we now make the following assumptions:

- $\lim_{N\to\infty}(1/N)\sum R^2_{Z|X}, \lim_{N\to\infty}(1/N)\sum X^2 = O_p(1)$.
- ϵ has finite variance $\sigma^2 > 0$.

Since our results are asymptotic, they apply whether raw or modified residuals are used (since $\frac{N-2}{N} \to 1$) but in the following raw residuals are assumed.

B.2 Statement of Main Theorems

Theorem 1. *Let* $f(r) = sign(r)\sqrt{\frac{(N-p)r^2}{1-r^2}}$, *which is a monotonic function. Then (A)* $t = f(r)$, *(B)* $t_{\pi(null)} = f(r_{\pi(null)})$, *(C)* $t_{\pi(full)} = f(r_{\pi(full)})$, *(D)* $t_{\pi(true)} = f(r_{\pi(true)})$. *Hence the* t*-statistics are equivalent to their* r*-statistic counterparts.*

Theorem 2. $t_{*(true)}, t_{*(null)}, t_{*(full)} \xrightarrow{D} \mathcal{N}(0,1)$ *with correlation matrix:*

$$\begin{bmatrix} 1 & 1 & \sqrt{1-r^2} \\ 1 & 1 & \sqrt{1-r^2} \\ \sqrt{1-r^2} & \sqrt{1-r^2} & 1 \end{bmatrix}^{-1} cor(t_{*(true)}, t_{*(null)}, t_{*(full)}) \to I$$

B.3 Useful results

Before proving the Theorems, we state some useful algebraic and convergence results which will be used in deriving the Theorem proofs.

Result 1. *Note that due to orthogonality,* $\sum R_{Y|XZ}X = \sum R_{Y|X}X = 0$.

Result 2. *For full model resampling,*

$$\hat{a}_* - \hat{a} = \frac{\sum R^*_{Y|XZ}X}{\sum X^2} \text{ and } \hat{b}_* - \hat{b} = \frac{\sum R^*_{Y|XZ}R_{Z|X}}{\sum R^2_{Z|X}}.$$

For null model resampling,

$$\hat{a}_* - \hat{a} = \frac{\sum R^*_{Y|X}X}{\sum X^2} \text{ and } \hat{b}_* = \frac{\sum R^*_{Y|X}R_{Z|X}}{\sum R^2_{Z|X}}.$$

For true model resampling,

$$\hat{a}_* - a = \frac{\sum \epsilon^*_{Y|X}X}{\sum X^2} \text{ and } \hat{b}_* = \frac{\sum \epsilon^*_{Y|X}R_{Z|X}}{\sum R^2_{Z|X}}.$$

Proof. For full model resampling,

$$\begin{aligned} \hat{a}_* - \hat{a} &= \frac{\sum(Y_* - Y)X}{\sum X^2} \\ &= \frac{\sum(R^*_{Y|XZ} - R_{Y|XZ})X}{\sum X^2} \\ &\quad \text{from (B.1) and (B.2)} \\ &= \frac{\sum R^*_{Y|XZ}X}{\sum X^2} \\ &\quad \text{since} \sum R_{Y|XZ}X = 0 \text{ (Result 1)} \end{aligned}$$

Similarly for the other cases. □

Result 3. $\frac{\sum \epsilon^2_{Y|X}}{\sum R^2_{Y|X}} \xrightarrow{p} 1$

Proof.

$$\sum \epsilon^2_{Y|X} = \sum (R_{Y|X} + (\hat{a} - a)X)^2 = \sum R^2_{Y|X} + (\hat{a} - a)^2 \sum X^2.$$

But $\sqrt{\sum X^2}(\hat{a} - a) = \frac{\sum \epsilon_{Y|X} X}{\sqrt{\sum X^2}} \xrightarrow{D} \mathcal{N}(0, \sigma^2)$.
So $\frac{\sum \epsilon^2_{Y|X}}{\sum R^2_{Y|X}} = 1 + \frac{O_p(1)}{\sum R^2_{Y|X}} \xrightarrow{p} 1$. □

Result 4. *If* $a = \frac{\sum YX}{\sum X^2}$ *then* $\sum (Y - aX)^2 = \sum Y^2 - \frac{(\sum YX)^2}{\sum X^2}$. *Similarly, if* $b = \frac{\sum YR_{Z|X}}{\sum R^2_{Z|X}}$ *then* $\sum (Y - bR_{Z|X})^2 = \sum Y^2 - \frac{(\sum YR_{Z|X})^2}{\sum R^2_{Z|X}}$.

Proof.

$$\begin{aligned}
\sum (Y - aX)^2 &= \sum [Y^2 + a^2X^2 - 2aYX] \\
&= \sum Y^2 + \frac{(\sum YX)^2}{(\sum X^2)^2} \sum X^2 - 2\frac{\sum YX}{\sum X^2} \sum YX \\
&= \sum Y^2 - \frac{(\sum YX)^2}{\sum X^2}
\end{aligned}$$

by cancellation. A similar proof applies for the second part of the Result. □

Result 5.

$$R_{Y|XZ} = R_{Y|X} - \hat{b} R_{Z|X},$$

where: $\hat{b} = \sum R_{Y|X} R_{Z|X} / \sum R^2_{Z|X}$.

This is an important result used in later proofs. It is stated and used in Anderson & Robinson (2001).

Proof. Note that $\sum X R_{Z|X} = 0$, by construction of linear model $Z = \gamma X + R_{Z|X}$.

$$\begin{aligned}
\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} &= \begin{bmatrix} X^T X & X^T R_{Z|X} \\ R_{Z|X}^T X & R_{Z|X}^T R_{Z|X} \end{bmatrix}^{-1} \begin{bmatrix} X^T Y \\ R_{Z|X}^T Y \end{bmatrix} \\
&= \begin{bmatrix} X^T X & 0 \\ 0 & R_{Z|X}^T R_{Z|X} \end{bmatrix}^{-1} \begin{bmatrix} X^T Y \\ R_{Z|X}^T Y \end{bmatrix} \\
&= \begin{bmatrix} (X^T X)^{-1} X^T Y \\ (R_{Z|X}^T R_{Z|X})^{-1} R_{Z|X}^T Y \end{bmatrix}
\end{aligned}$$

But note that $\hat{a} = (X^T X)^{-1} X^T Y$ is the estimated coefficient of X in the null model, so under both the null and the full model, the estimated coefficient of X is $\hat{a}$.

Therefore:

$$\begin{aligned}
Y &= \hat{a} X + \hat{b} R_{Z|X} + R_{Y|XZ} \\
&= \hat{a} X + R_{Y|X}
\end{aligned}$$

From which the result follows. □

B.4 Proof of Theorem 1

Proof of Part (A).

First note that the relationship between t and r is well-known for linear models, as in for example Draper & Smith (1966) and Seber (1977).

Note:

$$\begin{aligned}
t^2 &= \left(\frac{\hat{b}}{\hat{\text{se}}(\hat{b})} \right)^2 \\
&= (N-2) \frac{(\sum Y R_{Z|X})^2}{\sum R_{Y|XZ}^2 \sum R_{Z|X}^2} \qquad \text{(B.6)}
\end{aligned}$$

Now the numerator of (B.6), ignoring the factor $N-2$, is:

$$\begin{aligned}(\sum YR_{Z|X})^2 &= (\sum(Y-\hat{a}X)R_{Z|X})^2 \\ &= (\sum R_{Y|X}R_{Z|X})^2 \text{ Result 1}\end{aligned}$$

Also, the denominator (B.6) is:

$$\begin{aligned}\sum R^2_{Y|XZ}\sum R^2_{Z|X} &= \sum(R_{Y|X}-\hat{b}R_{Z|X})^2\sum R^2_{Z|X} \text{ from Result 5} \\ &= (\sum R^2_{Y|X}-(\sum R_{Y|X}R_{Z|X})^2/\sum R^2_{Z|X})\sum R^2_{Z|X} \\ &\quad \text{from Result 4, since } \hat{b}=\sum YR_{Z|X}/\sum R^2_{Z|X} \\ &= \sum R^2_{Y|X}\sum R^2_{Z|X}-(\sum R_{Y|X}R_{Z|X})^2\end{aligned}$$

Therefore (B.6) can be written as:

$$\begin{aligned}t^2 &= (N-2)\frac{(\sum R_{Y|X}R_{Z|X})^2}{\sum R^2_{Y|X}\sum R^2_{Z|X}-(\sum R_{Y|X}R_{Z|X})^2} \\ &= (N-2)\frac{r^2}{1-r^2}\end{aligned}$$

since $r^2=\frac{(\sum R_{Y|X}R_{Z|X})^2}{\sum R^2_{Y|X}\sum R^2_{Z|X}}$.

Proof of Part (B).

We want to show the equivalence of the Freedman & Lane (1983) method and $t_{\pi(\text{null})}$ for permutation testing. That is:

$$t^2_{\pi(\text{null})}=(N-2)\frac{r^2_{\pi(\text{null})}}{(1-r^2_{\pi(\text{null})})}$$

Note:

$$\begin{aligned}t^2_{\pi(\text{null})} &= \left(\frac{\hat{b}_{\pi(\text{null})}}{\hat{\text{se}}_\pi(\hat{b}_{\pi(\text{null})})}\right)^2 \\ &= (N-2)\frac{(\sum Y_{\pi(\text{null})}R_{Z|X})^2}{\sum R^2_{Y_{\pi(\text{null})}|XZ}\sum R^2_{Z|X}} \qquad \text{(B.7)}\end{aligned}$$

Now the numerator of (B.7), ignoring the factor $N-2$, is:

$$(\sum Y_{\pi(\text{null})}R_{Z|X})^2=[\sum(Y_{\pi(\text{null})}-a_{\pi(\text{null})}X)R_{Z|X}]^2,$$

from Result 1.

Also, the denominator of (B.7) is:

$$\begin{aligned}
\sum R^2_{Y_{\pi(\text{null})}|XZ} \sum R^2_{Z|X} &= \sum (R_{Y_{\pi(\text{null})}|X} - \hat{b}_{\pi(\text{null})} R_{Z|X})^2 \sum R^2_{Z|X} \\
&\quad \text{from Result 5} \\
&= \sum (\{Y_{\pi(\text{null})} - a_{\pi(\text{null})} X\} - \hat{b}_{\pi(\text{null})} R_{Z|X})^2 \sum R^2_{Z|X} \\
&= [\sum \{Y_{\pi(\text{null})} - a_{\pi(\text{null})} X\}^2 \\
&\quad -(\sum \{Y_{\pi(\text{null})} - a_{\pi(\text{null})} X\} R_{Z|X})^2 / \sum R^2_{Z|X}] \sum R^2_{Z|X} \\
&\quad \text{from Result 4}
\end{aligned}$$

Since $\hat{b}_{\pi(\text{null})} = \frac{\sum Y_{\pi(\text{null})} R_{Z|X}}{\sum R^2_{Z|X}} = \frac{\{\sum Y_{\pi(\text{null})} - a_{\pi(\text{null})} X\} R_{Z|X}}{\sum R^2_{Z|X}}$ from Result 1.

So:

$$\sum R^2_{Y_{\pi(\text{null})}|XZ} \sum R^2_{Z|X} = \sum \{Y_{\pi(\text{null})} - a_{\pi(\text{null})} X\}^2 \sum R^2_{Z|X} - (\sum \{Y_{\pi(\text{null})} - a_{\pi(\text{null})} X\} R_{Z|X})^2$$

Therefore, (B.7) can be written as:

$$\begin{aligned}
t^2_{\pi(\text{null})} &= (N-2) \frac{(\sum \{Y_{\pi(\text{null})} - a_{\pi(\text{null})} X\} R_{Z|X})^2}{\sum \{Y_{\pi(\text{null})} - a_{\pi(\text{null})} X\}^2 \sum R^2_{Z|X} - (\sum \{Y_{\pi(\text{null})} - a_{\pi(\text{null})} X\} R_{Z|X})^2} \\
&= (N-2) \frac{r^2_{\pi(\text{null})}}{1 - r^2_{\pi(\text{null})}}
\end{aligned}$$

Since $r^2_{\pi(\text{null})} = \frac{(\sum \{Y_{\pi(\text{null})} - a_{\pi(\text{null})} X\} R_{Z|X})^2}{\sum \{Y_{\pi(\text{null})} - a_{\pi(\text{null})} X\}^2 \sum R^2_{Z|X}}$ from the definition in Anderson & Robinson (2001).

Proof of Part (C).

We want to show the equivalence of the ter Braak (1992) method and t_{full} for permutation testing. That is:

$$t^2_{\pi(\text{full})} = (N-2) \frac{r^2_{\pi(\text{full})}}{(1 - r^2_{\pi(\text{full})})}$$

Note that:

$$\begin{aligned}
t_{\pi(\text{full})} &= (\hat{b}_{\pi(\text{full})} - \hat{b}) / \hat{\text{se}}_\pi(\hat{b}_{\pi(\text{full})}) \\
&= \sqrt{N-2} \frac{\sum (Y_{\pi(\text{full})} - Y) R_{Z|X}}{\sqrt{\sum R^2_{Y_{\pi(\text{full})}|XZ} \sum R^2_{Z|X}}} \qquad \text{(B.8)}
\end{aligned}$$

This proof is more algebraically complicated than that for null model permutation testing, hence it will divided into six steps.

1. We define objects A_π and k_π critical for the proof.
2. We show that the numerator of B.8, ignoring the factor $\sqrt{N-2}$, may be written as $\sum A_\pi R_{Z|X}$.
3. We show that $R_{Y_{\pi(\text{full})}|XZ}$ may be written as $A_\pi + (\hat{b} - \hat{b}_\pi)R_{Z|X}$.
4. We state a simplification for $\sum(A_\pi + (\hat{b} - \hat{b}_\pi)R_{Z|X})^2$.
5. We use steps 3 and 4 to show that the square of the denominator of B.8 may be written as: $\sum A_\pi^2 \sum R_{Z|X}^2 - (\sum A_\pi R_{Z|X})^2$.
6. We use steps 2 and 5 to complete the proof.

Step 1.

We now define A_π and k_π.

$A_\pi = R^\pi_{Y|XZ} - k_\pi X$

$k_\pi = \sum R^\pi_{Y|XZ} X / \sum X^2$

Step 2.

$$\begin{aligned}
\sum A_\pi R_{Z|X} &= \sum (R^\pi_{Y|XZ} - k_\pi X) R_{Z|X} \\
&= \sum R^\pi_{Y|XZ} R_{Z|X} \text{ since } \sum X R_{Z|X} = 0 \text{ by Result 1} \\
&= \sum (Y_{\pi(\text{full})} - \hat{a}X - \hat{b}R_{Z|X}) R_{Z|X} \\
&= \sum Y_{\pi(\text{full})} R_{Z|X} - \hat{b} \sum R^2_{Z|X} \text{ since } \sum X R_{Z|X} = 0 \text{ by Result 1} \\
&= \sum (Y_{\pi(\text{full})} - Y) R_{Z|X} \text{ since } Y = \hat{a}X + \hat{b}R_{Z|X} + R_{Y|XZ} \qquad \text{(B.9)}
\end{aligned}$$

Note that the last line in the above equation is the numerator of (B.8), ignoring the factor $\sqrt{N-2}$.

Step 3.

First, we show that $k_\pi = \hat{a}_\pi - \hat{a}$. Note that $\hat{a}_\pi - \hat{a} = \sum(Y_{\pi(\text{full})} - Y)X / \sum X^2$ and

$k_\pi = \sum R^\pi_{Y|XZ} X / \sum X^2$. Now:

$$\begin{aligned}\sum (Y_{\pi(\text{full})} - Y)X &= \sum (R^\pi_{Y|XZ} - R_{Y|XZ})X \\ &= \sum R^\pi_{Y|XZ} X \text{ because } \sum R_{Y|XZ} X = 0 \text{ from Result 1}\end{aligned}$$

Thus $k_\pi = \hat{a}_\pi - \hat{a}$ is proven.

Then, we relate A_π to $R_{Y_{\pi(\text{full})}|XZ}$:

$$\begin{aligned}R_{Y_{\pi(\text{full})}|XZ} &= A_\pi + (\hat{b} - \hat{b}_\pi) R_{Z|X} + (\hat{a} - \hat{a}_\pi) X + k_\pi X \\ &= A_\pi + (\hat{b} - \hat{b}_\pi) R_{Z|X}\end{aligned}$$

Step 4.

It can be shown that the following simplification holds.

$$\begin{aligned}&\sum (A_\pi + (\hat{b} - \hat{b}_\pi) R_{Z|X})^2 \\ &\quad = \sum (A_\pi + \hat{b} R_{Z|X})^2 - (\sum (A_\pi + \hat{b} R_{Z|X}) R_{Z|X})^2 / \sum R^2_{Z|X} \\ &= [(\sum (A_\pi + \hat{b} R_{Z|X})^2)(\sum R^2_{Z|X}) - (\sum (A_\pi + \hat{b} R_{Z|X}) R_{Z|X})^2] / \sum R^2_{Z|X} \\ &\quad = [\sum A^2_\pi \sum R^2_{Z|X} - (\sum A_\pi R_{Z|X})^2] / \sum R^2_{Z|X}\end{aligned}$$

Step 5.

From steps 3 and 4, the square of the denominator of (B.8) may be written as:

$$\begin{aligned}\sum R^2_{Y_{\pi(\text{full})}|XZ} \sum R^2_{Z|X} &= \sum [A_\pi + (\hat{b} - \hat{b}_\pi) R_{Z|X}]^2 (\sum R^2_{Z|X}) \\ &= \sum A^2_\pi \sum R^2_{Z|X} - (\sum A_\pi R_{Z|X})^2 \qquad \text{(B.10)}\end{aligned}$$

where the first equality is from step 3 and the second equality is from step 4.

Step 6.

Therefore combining equations (B.9) (step 2) and (B.10) (step 5) into equation (B.8):

$$\begin{aligned} t_{\pi(\text{full})} &= \frac{\sqrt{N-2}\sum(Y_{\pi(\text{full})}-Y)R_{Z|X}}{\sqrt{\sum R^2_{Y_{\pi(\text{full})}|XZ}\sum R^2_{Z|X}}} \\ &= \frac{\sqrt{N-2}\sum A_\pi R_{Z|X}}{\sqrt{\sum A_\pi^2 \sum R^2_{Z|X} - (\sum A_\pi R_{Z|X})^2}} \text{ from (B.9) and (B.10)} \\ &= \sqrt{N-2}\sqrt{\frac{r^2_{\pi(\text{full})}}{1-r^2_{\pi(\text{full})}}} \end{aligned}$$

because $r^2_{\pi(\text{full})} = \frac{(\sum A_\pi R_{Z|X})^2}{\sum A_\pi^2 \sum R^2_{Z|X}} = \frac{(\sum\{R^\pi_{Y|XZ}-k_\pi X\}R_{Z|X})^2}{\sum\{R^\pi_{Y|XZ}-k_\pi X\}^2 \sum R^2_{Z|X}}$, from definition of A_π and Anderson & Robinson (2001).

Proof of Part (D).

We want to show the equivalence of $r_{\pi(\text{true})}$ and $t_{\pi(\text{true})}$ for permutation testing. But the same proof as for null model permutation test holds for the true permutation test, with $\pi(\text{null})$ replaced by $\pi(\text{true})$. So $t^2_{\pi(\text{true})} = (N-2)\frac{r^2_{\pi(\text{true})}}{1-r^2_{\pi(\text{true})}}$ as required.

So Parts (A), (B), (C) and (D) have been proven. QED.

B.5 Introduction to Proof of Theorem 2

We now prove Theorem 2, for the bootstrap. In Lemma 1, we show that $N\hat{\text{var}}_*(\hat{b}_{*(a)}) \xrightarrow{p} N\text{var}(\hat{b}_{*(a)})$ where a denotes true, null or full resampling. In Lemma 2, we derive direct relations between $t_{*(\text{null})}$ and $t_{*(\text{full})}$ and between $t_{*(\text{null})}$ and $t_{*(\text{true})}$. In "Proof of Theorem 2", we show that $t_{*(\text{true})}$, $t_{*(\text{null})}$, and $t_{*(\text{full})}$ all converge in distribution to a standard normal. Then we show that the asymptotic correlation between $t_{*(\text{null})}$ and $t_{*(\text{full})}$ is $\sqrt{1-r^2}$, where r is the partial correlation from the original dataset, while the asymptotic correlation between $t_{*(\text{null})}$ and $t_{*(\text{true})}$ is one, thus completing the proof.

B.6 Convergence results for variance estimators

Lemma 1. *The following convergence results hold.*

1. $N\hat{var}_*(\hat{b}_{*(null)}) \xrightarrow{p} \frac{\sum R^2_{Y|X}}{\sum R^2_{Z|X}} = N var(\hat{b}_{*(null)})$

2. $N\hat{var}_*(\hat{b}_{*(full)}) \xrightarrow{p} \frac{\sum R^2_{Y|X}}{\sum R^2_{Z|X}}(1-r^2) = N var(\hat{b}_{*(full)})$

3. $N\hat{var}_*(\hat{b}_{*(true)}) \xrightarrow{p} \frac{\sum \epsilon^2_{Y|X}}{\sum R^2_{Z|X}} = N var(\hat{b}_{*(true)})$

Proof. Part 1:

$$\hat{\text{var}}_*(\hat{b}_{*(\text{null})}) = \frac{1}{N-2}\sum[(Y_* - \hat{a}_* X - \hat{b}_* R_{Z|X})^2]\frac{1}{\sum R^2_{Z|X}}$$

But $\hat{b}_{*(\text{null})} = \frac{\sum(Y_* - \hat{a}_* X)R_{Z|X}}{\sum R^2_{Z|X}}$ where $Y_* = \hat{a}X + R^*_{Y|X}$ so we can apply Result 4:

$$\begin{aligned}(N-2)(\sum R^2_{Z|X})\hat{\text{var}}_*(\hat{b}_{*(\text{null})}) &= \frac{1}{N-2}\left(\sum\{(Y_* - \hat{a}_* X)^2\} - \frac{\{\sum(Y_* - \hat{a}_* X)R_{Z|X}\}^2}{\sum R^2_{Z|X}}\right) \\ &= \sum\{(\hat{a} - \hat{a}_*)X + R^*_{Y|X}\}^2 \\ &\quad -\frac{1}{\sum R^2_{Z|X}}\left(\sum\{(\hat{a} - \hat{a}_*)X + R^*_{Y|X}\}R_{Z|X}\right)^2\end{aligned}$$

From equations (B.4) and (B.5) from the Definitions section.

But $\hat{a}_* - \hat{a} = \frac{\sum R^*_{Y|X}X}{\sum X^2}$ from Result 2 and $\sum XR_{Z|X} = 0$ from Result 1, so we simplify the above as follows:

Applying Result 4, $\sum\{(\hat{a} - \hat{a}_*)X + R^*_{Y|X}\}^2 = \sum R^{*2}_{Y|X} - \frac{(\sum R^*_{Y|X}X)^2}{\sum X^2}$. Applying Result 1, $\frac{1}{\sum R^2_{Z|X}}[\sum\{((\hat{a} - \hat{a}_*)X + R^*_{Y|X})R_{Z|X}\}]^2 = \frac{1}{\sum R^2_{Z|X}}(\sum R^*_{Y|X}R_{Z|X})^2$ because the $\sum XR_{Z|X}$ term is zero. So:

$$(N-2)\hat{\text{var}}_*(\hat{b}_{*(\text{null})}) = \frac{\sum R^{*2}_{Y|X}}{\sum R^2_{Z|X}} - \frac{1}{\sum R^2_{Z|X}}\frac{(\sum R^*_{Y|X}X)^2}{\sum X^2} - \frac{1}{\sum R^2_{Z|X}}\frac{(\sum R^*_{Y|X}R_{Z|X})^2}{\sum R^2_{Z|X}}.$$

But $\frac{\sum R^*_{Y|X}X}{\sqrt{\sum X^2}}, \frac{\sum R^*_{Y|X}R_{Z|X}}{\sqrt{\sum R^2_{Z|X}}} \xrightarrow{D} \mathcal{N}(0, (1/N)\sum R^2_{Y|X})$ by the Central Limit Theorem, and so: $\frac{(\sum R^*_{Y|X}X)^2}{\sum X^2}, \frac{(\sum R^*_{Y|X}R_{Z|X})^2}{\sum R^2_{Z|X}} = O_p(1)$, and these terms vanish when divided by $\sum R^2_{Z|X}$, as N increases.

Whereas: $\frac{\sum R^{*2}_{Y|X}}{\sum R^2_{Z|X}} = \frac{(1/N)\sum R^{*2}_{Y|X}}{(1/N)\sum R^2_{Z|X}} \xrightarrow{p} \frac{\sum R^2_{Y|X}}{\sum R^2_{Z|X}}$ by the Weak Law of Large Numbers. Therefore:

$$(N-2)\hat{\text{var}}_*(\hat{b}_{*(\text{null})}) \xrightarrow{p} \frac{\sum R^2_{Y|X}}{\sum R^2_{Z|X}}.$$

Noting that $(N-2)/N \to 1$, the convergence result in Part 1 follows.

Also: $\hat{b}_{*(\text{null})} = \frac{\sum Y_* R_{Z|X}}{\sum R^2_{Z|X}} = \frac{\sum R^*_{Y|X} R_{Z|X}}{\sum R^2_{Z|X}}$ so $E(\hat{b}_{*(\text{null})}) = \frac{\sum E(R^*_{Y|X}) R_{Z|X}}{\sum R^2_{Z|X}} = 0$ and:

$$\begin{aligned}
\text{var}(\hat{b}_{*(\text{null})}) &= E(\sum_i R^{*2}_{Y|X,i} R^2_{Z|X,i} + \sum_i \sum_{j \neq i} R^*_{Y|X,i} R_{Z|X,i} R^*_{Y|X,j} R_{Z|X,j})/(\sum R^2_{Z|X})^2 \\
&= (1/N) \sum R^2_{Y|X} \frac{\sum R^2_{Z|X}}{(\sum R^2_{Z|X})^2} \\
&= (1/N) \frac{\sum R^2_{Y|X}}{\sum R^2_{Z|X}}
\end{aligned}$$

So $N\hat{\text{var}}_*(\hat{b}_{*(\text{null})}) \xrightarrow{p} N\text{var}(\hat{b}_{*(\text{null})})$.

Part 2:

$$Y_* - \hat{a}_* X - \hat{b}_* R_{Z|X} = (\hat{a} - \hat{a}_*)X + (\hat{b} - \hat{b}_*)R_{Z|X} + R^*_{Y|XZ}.$$

Therefore, using the same approach as for null model resampling:

$$\begin{aligned}
(N-2)(\sum R^2_{Z|X})\hat{\text{var}}_*(\hat{b}_{*(\text{full})}) &= \sum (Y_* - \hat{a}_* X - \hat{b}_* R_{Z|X})^2 \\
&\quad - \frac{[\sum (Y_* - \hat{a}_* X - \hat{b}_* R_{Z|X}) R_{Z|X}]^2}{\sum R^2_{Z|X}} \\
&= \sum [(\hat{a} - \hat{a}_*)X + (\hat{b} - \hat{b}_*)R_{Z|X} + R^*_{Y|XZ}]^2 \\
&\quad - \frac{1}{\sum R^2_{Z|X}} \left(\sum \{(\hat{a} - \hat{a}_*)X + (\hat{b} - \hat{b}_*)R_{Z|X} + R^*_{Y|XZ}\} R_{Z|X} \right)^2
\end{aligned}$$

from equations (B.2) and (B.3) from the Definitions Section.

Recall Result 2: $\hat{a}_* - \hat{a} = \frac{\sum R^*_{Y|XZ} X}{\sum X^2}$ and $\hat{b}_* - \hat{b} = \frac{\sum R^*_{Y|XZ} R_{Z|X}}{\sum R^2_{Z|X}}$. So we can apply Result 4 and Result 1, using a similar argument as for null, but with $R^*_{Y|X}$ replaced by $R^*_{Y|XZ}$:

$$(N-2)\hat{\text{var}}_*(\hat{b}_{*(\text{full})}) = \frac{\sum R^{*2}_{Y|XZ}}{\sum R^2_{Z|X}} - \frac{1}{\sum R^2_{Z|X}} \frac{(\sum R^*_{Y|XZ} X)^2}{\sum X^2} - \frac{1}{\sum R^2_{Z|X}} \frac{(\sum R^*_{Y|XZ} R_{Z|X})^2}{\sum R^2_{Z|X}}.$$

But as before, $\frac{(\sum R^*_{Y|XZ}X)^2}{\sum X^2}$, $\frac{(\sum R^*_{Y|XZ}R_{Z|X})^2}{\sum R^2_{Z|X}}$ are $O_p(1)$, so these terms vanish when divided by $\sum R^2_{Z|X}$ as N increases.

Whereas: $\frac{\sum R^{*2}_{Y|XZ}}{\sum R^2_{Z|X}} = \frac{(1/N)\sum R^{*2}_{Y|XZ}}{(1/N)\sum R^2_{Z|X}} \xrightarrow{p} \frac{\sum R^2_{Y|XZ}}{\sum R^2_{Z|X}}$ by Weak Law of Large Numbers. Therefore:

$$(N-2)\hat{\text{var}}_*(\hat{b}_{*(\text{full})}) \xrightarrow{p} \frac{\sum R^2_{Y|XZ}}{\sum R^2_{Z|X}}.$$

From the RHS:

$$\begin{aligned}
\frac{\sum R^2_{Y|XZ}}{\sum R^2_{Z|X}} &= \frac{\sum (R_{Y|X} - \hat{b}R_{Z|X})^2}{\sum R^2_{Z|X}} \\
&= \frac{1}{\sum R^2_{Z|X}}\left(\sum R^2_{Y|X} - \frac{(\sum R_{Y|X}R_{Z|X})^2}{\sum R^2_{Z|X}}\right) \\
&= \frac{\sum R^2_{Y|X}}{\sum R^2_{Z|X}}(1-r^2)
\end{aligned}$$

This completes the convergence result stated in part 2.

Also: $\hat{b}_{*(\text{full})} - \hat{b} = \frac{\sum (Y_* - Y)R_{Z|X}}{\sum R^2_{Z|X}} = \frac{\sum R^*_{Y|XZ}R_{Z|X}}{\sum R^2_{Z|X}}$ from Lemma 2 so $E(\hat{b}_{*(\text{full})} - \hat{b}) = \frac{\sum E(R^*_{Y|XZ})R_{Z|X}}{\sum R^2_{Z|X}} = 0$ and:

$$\begin{aligned}
\text{var}(\hat{b}_{*(\text{full})}) &= E\left\{\sum_i R^{*2}_{Y|XZ,i}R^2_{Z|X,i} + \sum_i \sum_{j\neq i} R^*_{Y|XZ,i}R_{Z|X,i}R^*_{Y|XZ,j}R_{Z|X,j}\right\} / \left(\sum R^2_{Z|X}\right)^2 \\
&= (1/N)\sum R^2_{Y|XZ}\frac{\sum R^2_{Z|X}}{(\sum R^2_{Z|X})^2} \\
&= (1/N)\frac{\sum R^2_{Y|XZ}}{\sum R^2_{Z|X}}
\end{aligned}$$

So $N\hat{\text{var}}_*(\hat{b}_{*(\text{full})}) \xrightarrow{p} N\text{var}(\hat{b}_{*(\text{full})})$.

Part 3:

We use the same argument as for null model resampling, but with $\hat{a}$ replaced by a and $R_{Y|X}$ replaced by $\epsilon_{Y|X}$. Again, we use Result 4 and Result 1.

Therefore:

$$(N-2)\hat{\text{var}}_*(\hat{b}_{*(\text{true})}) = \frac{\sum \epsilon^{*2}_{Y|X}}{\sum R^2_{Z|X}} - \frac{1}{\sum R^2_{Z|X}}\frac{(\sum \epsilon^*_{Y|X}X)^2}{\sum X^2} - \frac{1}{\sum R^2_{Z|X}}\frac{(\sum \epsilon^*_{Y|X}R_{Z|X})^2}{\sum R^2_{Z|X}}$$

Using the same argument as before:

$$\frac{(\sum \epsilon^*_{Y|X} X)^2}{\sum X^2}, \frac{(\sum \epsilon^*_{Y|X} R_{Z|X})^2}{\sum R^2_{Z|X}} = O_p(1),$$

so the second and third terms vanish when divided through by $\sum R^2_{Z|X}$ as N increases.

Again using the Weak Law of Large Numbers:

$$(N-2)\hat{\text{var}}_*(\hat{b}_{*(\text{true})}) \xrightarrow{p} \frac{\sum \epsilon^{*2}_{Y|X}}{\sum R^2_{Z|X}} \xrightarrow{p} \frac{\sum \epsilon^2_{Y|X}}{\sum R^2_{Z|X}}.$$

This completes the convergence result stated in part 3.

Using the same argument as for null model resampling,

$$\text{var}(\hat{b}_{*(\text{true})}) = (1/N)\frac{\sum \epsilon^2_{Y|X}}{\sum R^2_{Z|X}}$$

So $N\hat{\text{var}}_*(\hat{b}_{*(\text{true})}) \xrightarrow{p} N\text{var}(\hat{b}_{*(\text{true})})$. And so parts 1,2 and 3 are proven.

□

B.7 Relations between statistics

Lemma 2. *The following relations between statistics hold.*

1. $t_{*(full)} = t_{*(null)} \times \frac{\hat{se}_*(\hat{b}_{*(null)})}{\hat{se}_*(\hat{b}_{*(full)})} - \hat{b}\frac{\sum R^*_{Z|X} R_{Z|X}}{\sum R^2_{Z|X} \times \hat{se}_*(\hat{b}_{*(full)})}$

2. $t_{*(true)} = t_{*(null)} \frac{\hat{se}_*(\hat{b}_{*(null)})}{\hat{se}_*(\hat{b}_{*(true)})} + (\hat{a} - a)\frac{\sum X^* R_{Z|X}}{\hat{se}_*(\hat{b}_{*(true)})}$

Proof. Part 1:

$$
\begin{aligned}
t_{*(\text{full})} &= \frac{\sum\{R^*_{Y|XZ} - (\hat{a} - \hat{a}_*)X\}R_{Z|X} / \sum R^2_{Z|X}}{\hat{\text{se}}_*(\hat{b}_{*(\text{full})})} \\
&= \frac{\sum R^*_{Y|XZ} R_{Z|X}}{\sum R^2_{Z|X} \times \hat{\text{se}}_*(\hat{b}_{*(\text{full})})} \\
&= \frac{\sum (R^*_{Y|X} - \hat{b} R^*_{Z|X}) R_{Z|X}}{\sum R^2_{Z|X} \times \hat{\text{se}}_*(\hat{b}_{*(\text{full})})} \\
&= t_{*(\text{null})} \times \frac{\hat{\text{se}}_*(\hat{b}_{*(\text{null})})}{\hat{\text{se}}_*(\hat{b}_{*(\text{full})})} - \hat{b} \frac{\sum R^*_{Z|X} R_{Z|X}}{\sum R^2_{Z|X} \times \hat{\text{se}}_*(\hat{b}_{*(\text{full})})}
\end{aligned}
$$

Part 2:

$$
\begin{aligned}
t_{*(\text{true})} &= \frac{\sum \epsilon^*_{Y|X} R_{Z|X}}{\hat{\text{se}}_*(\hat{b}_{*(\text{true})})} \\
&= \frac{\sum \{R_{Y|X} + (\hat{a} - a)X\}^* R_{Z|X}}{\hat{\text{se}}_*(\hat{b}_{*(\text{true})})} \\
&= \frac{\sum R^*_{Y|X} R_{Z|X}}{\hat{\text{se}}_*(\hat{b}_{*(\text{true})})} + (\hat{a} - a) \frac{\sum X^* R_{Z|X}}{\hat{\text{se}}_*(\hat{b}_{*(\text{true})})} \\
&= t_{*(\text{null})} \frac{\hat{\text{se}}_*(\hat{b}_{*(\text{null})})}{\hat{\text{se}}_*(\hat{b}_{*(\text{true})})} + (\hat{a} - a) \frac{\sum X^* R_{Z|X}}{\hat{\text{se}}_*(\hat{b}_{*(\text{true})})}
\end{aligned}
$$

□

B.8 Proof of Theorem 2

First we will show that $t_{*(\text{null})}, t_{*(\text{full})}, t_{*(\text{true})} \xrightarrow{D} \mathcal{N}(0,1)$.

We know from the Central Limit Theorem that for null model resampling, $\hat{b}_{*(\text{null})} \xrightarrow{D} \mathcal{N}(0, \text{var}(\hat{b}_{*(\text{null})})$ and so $\frac{\hat{b}_{*(\text{null})}}{\text{se}(\hat{b}_{*(\text{null})})} \xrightarrow{D} \mathcal{N}(0,1)$. Exactly the same argument applies for true model resampling, and for full model resampling, we note that $E(\hat{b}_{*(\text{null})}) - \hat{b} = 0$.

We know from linear model theory that if the linear model is correct, then the variance estimate is consistent for the true variance. This is confirmed in Lemma 1 for bootstrap samples. So from Lemma 1, $\frac{\hat{\text{se}}_*(\hat{b}_{*(\text{null})})}{\text{se}(\hat{b}_{*(\text{null})})} \xrightarrow{p} 1$, and similarly for full and

true model resampling. Therefore from Slutsky's Theorem, $t_{*(\text{null})}, t_{*(\text{full})}, t_{*(\text{true})} \xrightarrow{D} \mathcal{N}(0,1)$.

Now we will find the correlation matrix of $t_{*(\text{null})}, t_{*(\text{full})}, t_{*(\text{true})}$.

Following from Lemma 2:

$$
\begin{aligned}
\text{cor}(t_{*(\text{null})}, t_{*(\text{full})}) &= E(t_{*(\text{null})} \times t_{*(\text{full})}) \\
&= E\left\{ t_{*(\text{null})} \times \left[t_{*(\text{null})} \times \frac{\hat{\text{se}}_*(\hat{b}_{*(\text{null})})}{\hat{\text{se}}_*(\hat{b}_{*(\text{full})})} - \hat{b} \frac{\sum R^*_{Z|X} R_{Z|X}}{\sum R^2_{Z|X} \hat{\text{se}}_*(\hat{b}_{*(\text{full})})} \right] \right\} \\
&= \frac{\hat{\text{se}}_*(\hat{b}_{*(\text{null})})}{\hat{\text{se}}_*(\hat{b}_{*(\text{full})})} E(t^2_{*(\text{null})}) - E\left\{ t_{*(\text{null})} \times \frac{\hat{b} \sum R^*_{Z|X} R_{Z|X}}{\sum R^2_{Z|X} \hat{\text{se}}_*(\hat{b}_{*(\text{full})})} \right\} \\
&= \frac{\hat{\text{se}}_*(\hat{b}_{*(\text{null})})}{\hat{\text{se}}_*(\hat{b}_{*(\text{full})})} E(t^2_{*(\text{null})}) - E\left[\frac{\hat{b} \sum R^*_{Y|X} R_{Z|X} \sum R^*_{Z|X} R_{Z|X}}{(\sum R^2_{Z|X})^2 \hat{\text{se}}_*(\hat{b}_{*(\text{null})}) \hat{\text{se}}_*(\hat{b}_{*(\text{full})})} \right] \\
&\quad \text{from Result 2} \\
&\xrightarrow{p} \frac{1}{\sqrt{1-r^2}} E(t^2_{*(\text{null})}) - \frac{1}{\sqrt{1-r^2}} \frac{\hat{b}}{(1/N) \sum R^2_{Y|X} \sum R^2_{Z|X}} \\
&\quad \times E\left(\sum R^*_{Y|X} R_{Z|X} \sum R^*_{Z|X} R_{Z|X} \right) \\
&\quad \text{from Lemma 1} \\
&= \frac{1}{\sqrt{1-r^2}} - \frac{1}{\sqrt{1-r^2}} \frac{\hat{b}}{(1/N) \sum R^2_{Y|X} \sum R^2_{Z|X}} \\
&\quad \times \{ \sum \left(\frac{1}{N} \sum R_{Y|X} R_{Z|X} \right) R^2_{Z|X} \\
&\quad + \sum_i \left(\frac{1}{N} \sum R_{Y|X} \right) R_{Z|X,i} \times \sum_{j \neq i} \left(\frac{1}{N} \sum R_{Z|X} \right) R_{Z|X,j} \} \\
&= \frac{1}{\sqrt{1-r^2}} \left\{ 1 - \frac{\hat{b}}{(1/N) \sum R^2_{Y|X} \sum R^2_{Z|X}} \right\} \\
&\quad \times \frac{\sum R_{Y|X} R_{Z|X} \sum R^2_{Z|X}}{N} \\
&= \frac{1}{\sqrt{1-r^2}} \left(1 - \frac{(\sum R_{Y|X} R_{Z|X})^2}{\sum R^2_{Y|X} \sum R^2_{Z|X}} \right)
\end{aligned}
$$

Since:

$$
\begin{aligned}
\hat{b} &= \frac{\sum (\hat{a} X + R_{Y|X}) R_{Z|X}}{\sum R^2_{Z|X}} \\
&= \frac{\sum R_{Y|X} R_{Z|X}}{\sum R^2_{Z|X}} \text{ (Result 1)}
\end{aligned}
$$

So:

$$\begin{aligned}\mathrm{cor}(t_{*(\mathrm{null})}, t_{*(\mathrm{full})})\frac{1}{\sqrt{1-r^2}} &\xrightarrow{p} \frac{1}{1-r^2}\left(1-\frac{(\sum R_{Y|X}R_{Z|X})^2}{\sum R^2_{Y|X}\sum R^2_{Z|X}}\right)\\ &= 1\end{aligned}$$

From the definition of r.

Using Result 3, $\frac{\sum \epsilon^2_{Y|X}}{\sum R^2_{Y|X}} \xrightarrow{p} 1$, we can infer that (using Slutsky's Theorem):

$$N\hat{\mathrm{var}}_*(\hat{b}_{*(\mathrm{true})}) \xrightarrow{p} \frac{\sum R^2_{Y|X}}{\sum R^2_{Z|X}} \tag{B.11}$$

From Lemma 2:

$$\begin{aligned}\mathrm{cor}(t_{*(\mathrm{null})}, t_{*(\mathrm{true})}) &= E(t_{*(\mathrm{null})} \times t_{*(\mathrm{true})})\\ &= E\left(t^2_{*(\mathrm{null})}\frac{\hat{\mathrm{se}}_*(\hat{b}_{*(\mathrm{null})})}{\hat{\mathrm{se}}_*(\hat{b}_{*(\mathrm{true})})}\right) + (\hat{a}-a)E\left(\frac{(\sum R^*_{Y|X}R_{Z|X})(\sum X^*R_{Z|X})}{\hat{\mathrm{se}}_*(\hat{b}_{*(\mathrm{null})}) \times \hat{\mathrm{se}}_*(\hat{b}_{*(\mathrm{true})})}\right)\end{aligned}$$

But $\frac{\hat{\mathrm{se}}_*(\hat{b}_{*(\mathrm{null})})}{\hat{\mathrm{se}}_*(\hat{b}_{*(\mathrm{true})})} \xrightarrow{p} 1$ (from Lemma 1 and (B.11), and:
$\sqrt{N}\hat{\mathrm{se}}_*(\hat{b}_{*(\mathrm{null})}), \sqrt{N}\hat{\mathrm{se}}_*(\hat{b}_{*(\mathrm{true})}) \xrightarrow{p} \sqrt{(1/N)\frac{\sum R^2_{Y|X}}{\sum R^2_{Z|X}}}$ (from Lemma 1 and (B.11)) so:

$$\begin{aligned}\mathrm{cor}(t_{*(\mathrm{null})}, t_{*(\mathrm{true})}) &\xrightarrow{p} 1 + N(\hat{a}-a)\frac{\sum R^2_{Y|X}}{\sum R^2_{Z|X}} \times E\left((\sum R^*_{Y|X}R_{Z|X})(\sum X^*R_{Z|X})\right)\\ &= 1 + N(\hat{a}-a)\frac{\sum R^2_{Y|X}}{\sum R^2_{Z|X}} \times \{\sum_i E(R_{Y|X}X)^*R^2_{Z|X}\\ &\quad + \sum_i\sum_{j\neq i} E(R^*_{Y_i|X_i}X^*_j)R_{Z_i|X_i}R_{Z_j|X_j}\}\\ &= 1\end{aligned}$$

The proof of the Theorem 2 is now complete.

Appendix C

Bootstrap and simulation code

C.1 Bootstrap and simulation definition code

```
### case resampling ###

boot.cases <- function(R, y, x1, x2, resid.type, het, eps=1.e-8)
{
n                 <- length(y)
fit.alt           <- lm(y~x1+x2)
raw.resid         <- resid(fit.alt)
sigma.hat         <- summary(fit.alt)$sigma
mod.resid         <- sigma.hat * rstandard(fit.alt)
if (resid.type==0) {used.resid <- raw.resid}
if (resid.type==1) {used.resid <- mod.resid}
x.dat             <- cbind(1,x1,x2)
xtx.inv           <- solve(t(x.dat)%*%x.dat)
h2          <- ( xtx.inv %*% t(x.dat) )^2
se.sand           <- sqrt( h2[3,] %*% used.resid^2 ) #faster sandwich est

b.star            <- rep(NA,R)
se.sand.star      <- rep(NA,R)
```

```
se.sand.x         <- rep(NA,R)
se.naive.star     <- rep(NA,R)
se.naive.x        <- rep(NA,R)
# Generate matrix of random indices before loop
set       <- 1:n
ivec      <- matrix( sample(set,n*R,replace=TRUE), n, R)
r         <- 1
while (r <= R)
{
    y.star         <- y[ ivec[,r] ]
    x1.star        <- x1[ ivec[,r] ]
    x2.star        <- x2[ ivec[,r] ]
    fit.star       <- lm(y.star~x1.star+x2.star)
      b.star[r]      <- coef(fit.star)[3]
    x.dat.star     <- x.dat[ ivec[,r] , ]
      xtx.inv.star   <- solve( t(x.dat.star) %*% x.dat.star )
    sigma.hat.star <- summary(fit.star)$sigma
    if (resid.type==0) {s.i.star <- resid(fit.star)}
    if (resid.type==1)
    {
        mod.resid.star <- sigma.hat.star * rstandard(fit.star)
        s.i.star <- mod.resid.star
    }
    se.sand.star[r]  <- sqrt((xtx.inv.star%*%t(x.dat.star)%*%diag(s.i.star^2)
    %*%x.dat.star%*%xtx.inv.star)[3,3])
    se.sand.x[r]     <- sqrt( h2[3,] %*% (s.i.star^2) )
    se.naive.star[r] <- sigma.hat.star * sqrt( xtx.inv.star[3,3] )
    se.naive.x[r]    <- sigma.hat.star * sqrt( xtx.inv[3,3])
    if (is.na(b.star[r])==FALSE) r <- r + 1
}

b.hat            <- coef(fit.alt)[3]
```

```
if (het==TRUE) { z.0 <- as.vector(b.hat/se.sand) }
else           { z.0 <- summary(fit.alt)$coef[3,3] }

p.case.non       <- mean( abs(b.star-b.hat) > abs(b.hat)-eps )

z.sand.star      <- (b.star - b.hat)/se.sand.star
p.case.sand      <- mean( abs(z.sand.star) > abs(z.0)-eps )

z.sand.x         <- (b.star - b.hat)/se.sand.x
p.case.sand.x    <- mean( abs(z.sand.x) > abs(z.0)-eps )

z.naive.star     <- (b.star - b.hat)/se.naive.star
p.case.naive     <- mean( abs(z.naive.star) > abs(z.0)-eps )

z.naive.x        <- (b.star - b.hat)/se.naive.x
p.case.naive.x   <- mean( abs(z.naive.x) > abs(z.0)-eps )

c( p.case.non, p.case.sand, p.case.sand.x, p.case.naive, p.case.naive.x )
}

### residual resampling ###

boot.resid <- function(R, y, x1, x2, resid.type=0, het=F, reduced=FALSE, eps=1
{
n                 <- length(y)

fit.alt           <- lm(y~x1+x2)
if (reduced==TRUE)
{
    fit         <- lm(y~x1)
}
```

```
else
{
    fit         <- fit.alt
}
fit.fitted        <- fit$fitted
sigma.hat         <- summary(fit)$sigma

x.dat             <- cbind(1,x1,x2)
xtx.inv           <- solve( t(x.dat) %*% x.dat )

b.hat             <- coef(fit.alt)[3]

if (resid.type==1)
{
    resid.vector <- sigma.hat * rstandard(fit)
}
else
{
    resid.vector <- resid(fit)
}
fitted.mat      <- matrix( rep(fit.fitted,R), nrow=n )
r.star.mat      <- matrix( sample(resid.vector, n*R, replace=TRUE), nrow=n )
y.star.mat      <- fitted.mat + r.star.mat
fit.star        <- lm(y.star.mat~x1+x2)
b.star          <- coef(fit.star)[3,]
resids          <- resid(fit.star)
sigmasq         <- colSums(resids^2) / (n-3)
se.naive.star   <- sqrt( sigmasq*xtx.inv[3,3] )

if (reduced==FALSE)
{
    p.resid.non <- mean( abs( b.star - b.hat ) > abs(b.hat) - eps )
```

```
    z.star.one  <- ( b.star - b.hat ) / se.naive.star
}
if (reduced==TRUE)
{
    p.resid.non <- mean( abs(b.star) > abs(b.hat) - eps )
    z.star.one  <- ( b.star - 0 ) / se.naive.star
}

if (het==TRUE) {
    if (resid.type==1)
    {
        sigma.hat    <- summary(fit.alt)$sigma
        resid.vector <- sigma.hat * rstandard(fit.alt)
    }
    else
    {
        resid.vector <- resid(fit.alt)
    }
    cov.mat.estimator <-
xtx.inv%*%t(x.dat)%*%diag(resid.vector^2)%*%x.dat%*%xtx.inv
    se.sand  <- sqrt(cov.mat.estimator[3,3])
    z.0      <- as.vector(b.hat / se.sand)
}
else
{
    z.0      <- summary(fit.alt)$coef[3,3]
}

p.resid.naive <- mean( abs(z.star.one) > abs(z.0) - eps )
c(p.resid.non, p.resid.naive)
}
```

```
### score resampling ###

boot.score <- function(R, y, x1, x2, t.type, het, eps=1.e-8)
{
n                 <- length(y)
fit.alt           <- lm(y~x1+x2)
b.hat             <- coef(fit.alt)[3]
sigma.hat         <- summary(fit.alt)$sigma
mod.resid         <- sigma.hat * rstandard(fit.alt)
x.dat             <- cbind(1,x1,x2)
xtx.inv           <- solve(t(x.dat)%*%x.dat)

h2            <- ( xtx.inv %*% t(x.dat) )^2
se.sand           <- sqrt( h2[3,] %*% mod.resid^2 ) #faster sand est

#generate resampled t* hence y*
if (t.type==1)
{
    t.star     <- matrix(rnorm(n*R,mean=0,sd=1),nrow=n)
}
else
{
    aj         <- mod.resid/sqrt((1/n)*sum(mod.resid^2))
    t.star     <- matrix(sample(aj,n*R,replace=TRUE),nrow=n)
}
fitted.mat       <- matrix( rep( fitted(fit.alt), R ), nrow=n )
r.mat            <- matrix( rep( mod.resid, R), nrow=n )
y.star           <- fitted.mat + r.mat * t.star

#fit model and store stats
fit.regress.full <- lm(y.star~x1+x2)
```

```
b.star           <- coef(fit.regress.full)[3,]
resids           <- resid(fit.regress.full)
sigmasq          <- colSums(resids^2) / (n-3)
se.naive.star    <- sqrt( sigmasq*xtx.inv[3,3] )

H                <- diag(rep(1,n)) - x.dat %*% xtx.inv %*% t(x.dat)
mod.resids       <- diag( 1/sqrt( diag(H) ) ) %*% resids
se.sand.star     <- sqrt( h2[3,] %*% mod.resids^2 ) #faster sand est

#get p-values
if (het==TRUE) { z.0 <- as.vector(b.hat/se.sand) }
else           { z.0 <- summary(fit.alt)$coef[3,3] }

p.score.non      <- mean( abs( b.star - b.hat ) > abs(b.hat) - eps )

z.star.naive     <- (b.star - b.hat) / se.naive.star
p.score.naive    <- mean( abs(z.star.naive) > abs(z.0) - eps)

z.star.sand      <- (b.star - b.hat) / se.sand.star
p.score.sand     <- mean( abs(z.star.sand) > abs(z.0) - eps )
c(p.score.non, p.score.naive, p.score.sand)
}



### simulation design ####

explore.diff <- function(R,nsim,beta.two,sed=0,n,fixed,correlated,het)
{
if (sed > 0) { set.seed(sed) }

if (fixed==TRUE)
{
```

```
    nrep <- n/16
    x1   <- c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4)
    x2   <- c(1,2,3,4,1,2,3,4,1,2,3,4,1,2,3,4)
    x1   <- rep(x1,nrep)
    x2   <- rep(x2,nrep)
}

if (fixed==FALSE)
{
    x1   <- rnorm(n,mean=0,sd=1)
    x2   <- rnorm(n,mean=0,sd=1)
    a    <- sqrt(1.25)
    if (correlated==TRUE)
    {
        alpha <- 0.8
        x2    <- alpha*x1 + sqrt(1-alpha^2)*x2
        a     <- a * (1-alpha^2)^(-0.25) #to ensure that det(X'X) is constant
    }
    x1   <- a*x1 + 2.5
    x2   <- a*x2 + 2.5
}

if (het==FALSE)
{
error <- rnorm((nsim*n),mean=0,sd=2)
}

if (het==TRUE)
{
if (fixed==TRUE)
    { error <- rep(pmax(1,x2),nsim) * rnorm((nsim*n),mean=0,sd=(2/sqrt(7.5))) }
if (fixed==FALSE & correlated==FALSE)
```

```
    { error <- rep(pmax(1,x2),nsim) *
    rnorm((nsim*n),mean=0,sd=(2/sqrt(7.550634))) }
if (fixed==FALSE & correlated==TRUE)
    { error <- rep(pmax(1,x2),nsim) *
    rnorm((nsim*n),mean=0,sd=(2/sqrt(8.41257))) }
}

y.mat <- 4 + 3*rep(x1,nsim) + beta.two*rep(x2,nsim)+ error
y.mat <- matrix(y.mat,nrow=n,ncol=nsim)
p.value.mat <- array(NA,c(nsim,17,2))
dimnames(p.value.mat)[[2]]=c("case.non","case.sand",
"case.sand.x","case.naive","case.naive.x","resid.full.non",
"resid.full.naive","resid.null.non","resid.null.naive","score.non",
"score.naive","score.sand","scoreR.non","scoreR.naive","scoreR.sand",
"resid.raw.null.non","resid.raw.null.naive")

for (j in 1:nsim)
{
    y                   <- y.mat[,j]
      p.value.mat[j,1:5,1]   <- boot.cases(R,y,x1,x2,resid.type=0,het=F)
    p.value.mat[j,6:7,1]   <- boot.resid(R,y,x1,x2,resid.type=1,het=F,
    reduced=FALSE)
    p.value.mat[j,8:9,1]   <- boot.resid(R,y,x1,x2,resid.type=1,het=F,
    reduced=TRUE)
      p.value.mat[j,10:12,1] <- boot.score(R,y,x1,x2,t.type=1,het=F)
    p.value.mat[j,13:15,1] <- boot.score(R,y,x1,x2,t.type=0,het=F)
    p.value.mat[j,16:17,1] <- boot.resid(R,y,x1,x2,resid.type=0,het=F,
    reduced=TRUE)

      p.value.mat[j,1:5,2]   <- boot.cases(R,y,x1,x2,resid.type=0,het=T)
      p.value.mat[j,10:12,2] <- boot.score(R,y,x1,x2,t.type=1,het=T)
    p.value.mat[j,13:15,2] <- boot.score(R,y,x1,x2,t.type=0,het=T)
```

```
}

p.value.mat
}
```

C.2 Simulation execution code

```
# example script file for running bootstrap simulations

source("sim.functions.R")

test.run = explore.diff(2,2,0,1,16,TRUE, FALSE,FALSE) #To get the
dimensions and names of columns of output

R=1000 nsim=1000 n=c(16,32,64) options = c(TRUE, FALSE, FALSE)
  fix.homo.I = array(NA,c(nsim,dim(test.run)[2:3],3))
colnames(fix.homo.I)=colnames(test.run)
  fix.homo.I[,,,1] = explore.diff(R,nsim,0,1+10,n[1],options[1],options[2],options[3])
  fix.homo.I[,,,2] = explore.diff(R,nsim,0,2+10,n[2],options[1],options[2],options[3])
  fix.homo.I[,,,3] = explore.diff(R,nsim,0,3+10,n[3],options[1],options[2],options[3])

  fix.homo.II = array(NA,c(nsim,dim(test.run)[2:3],3))
colnames(fix.homo.II)=colnames(test.run)
  fix.homo.II[,,,1] = explore.diff(R,nsim,0.5,4+10,n[1],options[1],options[2],
  options[3])
  fix.homo.II[,,,2] = explore.diff(R,nsim,0.5,5+10,n[2],options[1],options[2],
  options[3])
  fix.homo.II[,,,3] = explore.diff(R,nsim,0.5,6+10,n[3],options[1],options[2],
  options[3])

save(fix.homo.I, fix.homo.II, nsim, R, n, options,
```

```
file="fix.homo.run2.RData")

# randu

options = c(FALSE, FALSE, FALSE)
  randu.homo.I = array(NA,c(nsim,dim(test.run)[2:3],3))
colnames(randu.homo.I)=colnames(test.run)
  randu.homo.I[,,,1] = explore.diff(R,nsim,0,11+10,n[1],options[1],options[2],
  options[3])
  randu.homo.I[,,,2] = explore.diff(R,nsim,0,12+10,n[2],options[1],options[2],
  options[3])
  randu.homo.I[,,,3] = explore.diff(R,nsim,0,13+10,n[3],options[1],options[2],
  options[3])

  randu.homo.II = array(NA,c(nsim,dim(test.run)[2:3],3))
colnames(randu.homo.II)=colnames(test.run)
  randu.homo.II[,,,1] = explore.diff(R,nsim,0.5,14+10,n[1],options[1],
  options[2],options[3])
  randu.homo.II[,,,2] = explore.diff(R,nsim,0.5,15+10,n[2],options[1],
  options[2],options[3])
  randu.homo.II[,,,3] = explore.diff(R,nsim,0.5,16+10,n[3],options[1],
  options[2],options[3])

save(randu.homo.I, randu.homo.II, nsim, R, n, options,
file="randu.homo.run2.RData")

# randc

options = c(FALSE, TRUE, FALSE) randc.homo.I =
array(NA,c(nsim,dim(test.run)[2:3],3))
colnames(randc.homo.I)=colnames(test.run) randc.homo.I[,,,1] =
explore.diff(R,nsim,0,21+10,n[1],options[1],options[2],options[3])
```

```
randc.homo.I[,,,2] =
explore.diff(R,nsim,0,22+10,n[2],options[1],options[2],options[3])
randc.homo.I[,,,3] =
explore.diff(R,nsim,0,23+10,n[3],options[1],options[2],options[3])

randc.homo.II = array(NA,c(nsim,dim(test.run)[2:3],3))
colnames(randc.homo.II)=colnames(test.run) randc.homo.II[,,,1] =
explore.diff(R,nsim,0.5,24+10,n[1],options[1],options[2],options[3])
randc.homo.II[,,,2] =
explore.diff(R,nsim,0.5,25+10,n[2],options[1],options[2],options[3])
randc.homo.II[,,,3] =
explore.diff(R,nsim,0.5,26+10,n[3],options[1],options[2],options[3])

save(randc.homo.I, randc.homo.II, nsim, R, n, options,
file="randc.homo.run2.RData")

# fixedh

options = c(TRUE, FALSE, TRUE) fix.het.I =
array(NA,c(nsim,dim(test.run)[2:3],3))
colnames(fix.het.I)=colnames(test.run) fix.het.I[,,,1] =
explore.diff(R,nsim,0,31+10,n[1],options[1],options[2],options[3])
fix.het.I[,,,2] =
explore.diff(R,nsim,0,32+10,n[2],options[1],options[2],options[3])
fix.het.I[,,,3] =
explore.diff(R,nsim,0,33+10,n[3],options[1],options[2],options[3])

fix.het.II = array(NA,c(nsim,dim(test.run)[2:3],3))
colnames(fix.het.II)=colnames(test.run) fix.het.II[,,,1] =
explore.diff(R,nsim,0.5,34+10,n[1],options[1],options[2],options[3])
fix.het.II[,,,2] =
explore.diff(R,nsim,0.5,35+10,n[2],options[1],options[2],options[3])
```

```
fix.het.II[,,,3] =
explore.diff(R,nsim,0.5,36+10,n[3],options[1],options[2],options[3])

save(fix.het.I, fix.het.II, nsim, R, n, options,
file="fix.het.run2.RData")

# randuh

options = c(FALSE, FALSE, TRUE) randu.het.I =
array(NA,c(nsim,dim(test.run)[2:3],3))
colnames(randu.het.I)=colnames(test.run) randu.het.I[,,,1] =
explore.diff(R,nsim,0,41+10,n[1],options[1],options[2],options[3])
randu.het.I[,,,2] =
explore.diff(R,nsim,0,42+10,n[2],options[1],options[2],options[3])
randu.het.I[,,,3] =
explore.diff(R,nsim,0,43+10,n[3],options[1],options[2],options[3])

randu.het.II = array(NA,c(nsim,dim(test.run)[2:3],3))
colnames(randu.het.II)=colnames(test.run) randu.het.II[,,,1] =
explore.diff(R,nsim,0.5,44+10,n[1],options[1],options[2],options[3])
randu.het.II[,,,2] =
explore.diff(R,nsim,0.5,45+10,n[2],options[1],options[2],options[3])
randu.het.II[,,,3] =
explore.diff(R,nsim,0.5,46+10,n[3],options[1],options[2],options[3])

save(randu.het.I, randu.het.II, nsim, R, n, options,
file="randu.het.run2.RData")

# randch

options = c(FALSE, TRUE, TRUE) randc.het.I =
array(NA,c(nsim,dim(test.run)[2:3],3))
```

```
colnames(randc.het.I)=colnames(test.run) randc.het.I[,,,1] =
explore.diff(R,nsim,0,51+10,n[1],options[1],options[2],options[3])
randc.het.I[,,,2] =
explore.diff(R,nsim,0,52+10,n[2],options[1],options[2],options[3])
randc.het.I[,,,3] =
explore.diff(R,nsim,0,53+10,n[3],options[1],options[2],options[3])

randc.het.II = array(NA,c(nsim,dim(test.run)[2:3],3))
colnames(randc.het.II)=colnames(test.run) randc.het.II[,,,1] =
explore.diff(R,nsim,0.5,54+10,n[1],options[1],options[2],options[3])
randc.het.II[,,,2] =
explore.diff(R,nsim,0.5,55+10,n[2],options[1],options[2],options[3])
randc.het.II[,,,3] =
explore.diff(R,nsim,0.5,56+10,n[3],options[1],options[2],options[3])

save(randc.het.I, randc.het.II, nsim, R, n, options,
file="randc.het.run2.RData")
```

Appendix D

Simulation results

The following tables present simulation results from the 18 different simulations described in Chapter 3. Results are reported for Type I error, raw power, and adjusted power.

Results are reported for 17 different resampling test statistics, as follows:

Case resampling non-pivotal (*non*), sandwich estimator calculated from X_* (*sand*) and calculated from the original design matrix X (*sand.x*), naive estimator calculated from X_* (*naive*) and calculated from the original design matrix X (*naive.x*).

Residual resampling full model non-pivotal (*full.non*) and naive estimator (*full.naive*), null model non-pivotal (*null.non*) and naive estimator (*null.naive*), (null model) raw residual non-pivotal (*raw.non*) and naive estimator (*raw.naive*).

Score resampling normal method non-pivotal (*non*), naive estimator (*naive*) and sandwich estimator (*sand*), modified residuals method non-pivotal (*R.non*), naive estimator (*R.naive*) and sandwich estimator (*R.sand*).

When using pivotal statistics for case and score resampling, the observed test statistic was calculated using either the naive estimator (*naive-t*) or the sandwich estimator (*sand-t*). This was not done for residual resampling because in that case the true variance estimator takes the form of the naive estimator.

Individual simulation Type I error results

Regular design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.078	0.019	0.012	0.051	0.029	0.07	0.052	0.053	0.056	0.064	0.057	0.09	0.074	0.059	0.095	0.072	0.056
N=32	0.072	0.042	0.035	0.058	0.05	0.068	0.061	0.059	0.06	0.061	0.058	0.072	0.065	0.064	0.071	0.064	0.065
N=64	0.052	0.038	0.036	0.045	0.041	0.043	0.043	0.042	0.044	0.044	0.04	0.054	0.05	0.048	0.05	0.047	0.043
sand-*t*																	
N=16		0.05	0.046	0.102	0.073								0.105	0.077		0.103	0.077
N=32		0.053	0.053	0.087	0.075								0.08	0.062		0.078	0.063
N=64		0.044	0.048	0.054	0.052								0.053	0.045		0.048	0.042

Normal uncorrelated design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.073	0.013	0.01	0.055	0.026	0.076	0.057	0.049	0.054	0.068	0.059	0.092	0.079	0.06	0.093	0.08	0.065
N=32	0.067	0.031	0.024	0.07	0.041	0.062	0.053	0.056	0.057	0.057	0.056	0.082	0.074	0.059	0.081	0.077	0.06
N=64	0.061	0.033	0.043	0.059	0.052	0.063	0.057	0.054	0.053	0.06	0.054	0.064	0.063	0.053	0.067	0.063	0.055
sand-*t*																	
N=16		0.058	0.05	0.114	0.078								0.101	0.088		0.107	0.086
N=32		0.063	0.061	0.113	0.085								0.098	0.08		0.098	0.08
N=64		0.052	0.054	0.079	0.064								0.064	0.06		0.068	0.063

Normal correlated design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.052	0.012	0.003	0.061	0.015	0.084	0.05	0.044	0.052	0.071	0.05	0.116	0.091	0.053	0.122	0.094	0.054
N=32	0.072	0.035	0.036	0.074	0.052	0.068	0.057	0.055	0.06	0.067	0.057	0.084	0.075	0.058	0.087	0.076	0.061
N=64	0.058	0.037	0.038	0.052	0.047	0.056	0.049	0.051	0.05	0.052	0.049	0.06	0.055	0.051	0.063	0.057	0.053
sand-*t*																	
N=16		0.072	0.051	0.157	0.096								0.142	0.11		0.146	0.112
N=32		0.065	0.067	0.107	0.089								0.096	0.085		0.099	0.086
N=64		0.049	0.048	0.079	0.066								0.063	0.056		0.065	0.05

Regular heteroscedastic design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.064	0.018	0.013	0.039	0.026	0.071	0.047	0.046	0.049	0.065	0.051	0.084	0.064	0.06	0.081	0.069	0.061
N=32	0.07	0.031	0.026	0.05	0.035	0.068	0.054	0.056	0.058	0.065	0.059	0.071	0.066	0.06	0.069	0.068	0.064
N=64	0.057	0.039	0.031	0.047	0.041	0.069	0.065	0.063	0.061	0.066	0.066	0.061	0.057	0.062	0.052	0.054	0.064
sand-*t*																	
N=16		0.04	0.033	0.071	0.056								0.077	0.071		0.081	0.075
N=32		0.04	0.047	0.074	0.056								0.069	0.062		0.072	0.068
N=64		0.039	0.035	0.05	0.042								0.05	0.049		0.053	0.052

Normal uncorrelated heteroscedastic design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.086	0.023	0.012	0.069	0.034	0.129	0.097	0.087	0.092	0.113	0.1	0.125	0.103	0.089	0.12	0.103	0.086
N=32	0.086	0.034	0.025	0.077	0.043	0.12	0.108	0.101	0.107	0.111	0.106	0.105	0.096	0.1	0.1	0.099	0.1
N=64	0.066	0.043	0.042	0.056	0.053	0.074	0.069	0.073	0.077	0.076	0.075	0.062	0.066	0.072	0.061	0.061	0.076
sand-*t*																	
N=16		0.066	0.053	0.123	0.084								0.118	0.099		0.12	0.1
N=32		0.062	0.053	0.111	0.082								0.1	0.089		0.095	0.091
N=64		0.046	0.049	0.067	0.06								0.063	0.058		0.063	0.06

Normal correlated heteroscedastic design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.031	0.001	0.001	0.011	0.003	0.046	0.033	0.027	0.028	0.043	0.033	0.051	0.032	0.032	0.046	0.032	0.036
N=32	0.078	0.025	0.024	0.074	0.05	0.081	0.073	0.062	0.071	0.075	0.071	0.102	0.084	0.069	0.088	0.081	0.073
N=64	0.061	0.05	0.029	0.054	0.046	0.075	0.069	0.067	0.072	0.073	0.071	0.066	0.066	0.076	0.065	0.066	0.075
sand-*t*																	
N=16		0.013	0.007	0.042	0.02								0.055	0.043		0.055	0.048
N=32		0.075	0.078	0.146	0.117								0.139	0.095		0.13	0.107
N=64		0.05	0.034	0.064	0.049								0.056	0.067		0.057	0.067

Individual size-adjusted power results

Regular design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.155	0.158	0.157	0.153	0.153	0.164	0.167	0.16	0.169	0.15	0.148	0.155	0.161	0.155	0.143	0.152	0.161
N=32	0.331	0.312	0.304	0.31	0.314	0.317	0.317	0.311	0.306	0.314	0.317	0.31	0.312	0.321	0.319	0.316	0.313
N=64	0.606	0.609	0.608	0.607	0.618	0.62	0.621	0.622	0.629	0.623	0.623	0.6	0.602	0.615	0.612	0.614	0.62
sand-*t*																	
N=16		0.138	0.129	0.13	0.128								0.122	0.147		0.125	0.134
N=32		0.302	0.289	0.296	0.309								0.293	0.294		0.275	0.305
N=64		0.619	0.572	0.583	0.581								0.581	0.609		0.593	0.617

Normal uncorrelated design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.039	0.085	0.057	0.117	0.068	0.109	0.105	0.117	0.115	0.113	0.11	0.139	0.122	0.099	0.139	0.127	0.088
N=32	0.3	0.289	0.298	0.275	0.288	0.313	0.298	0.281	0.283	0.306	0.298	0.261	0.267	0.28	0.274	0.274	0.284
N=64	0.539	0.544	0.52	0.534	0.529	0.547	0.544	0.555	0.569	0.555	0.554	0.55	0.543	0.553	0.537	0.54	0.546
sand-*t*																	
N=16		0.133	0.089	0.162	0.103								0.159	0.124		0.17	0.122
N=32		0.246	0.245	0.238	0.254								0.215	0.241		0.208	0.228
N=64		0.519	0.498	0.5	0.5								0.487	0.527		0.492	0.516

Normal correlated design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.077	0.146	0.095	0.105	0.093	0.146	0.142	0.135	0.137	0.134	0.136	0.121	0.135	0.141	0.131	0.142	0.139
N=32	0.151	0.169	0.159	0.144	0.152	0.162	0.163	0.165	0.171	0.167	0.175	0.143	0.16	0.173	0.147	0.148	0.171
N=64	0.307	0.316	0.316	0.317	0.314	0.313	0.315	0.306	0.313	0.32	0.313	0.306	0.309	0.31	0.295	0.296	0.295
sand-*t*																	
N=16		0.116	0.067	0.104	0.07								0.125	0.119		0.131	0.125
N=32		0.143	0.135	0.118	0.125								0.122	0.138		0.122	0.141
N=64		0.302	0.3	0.284	0.286								0.294	0.31		0.3	0.319

Regular heteroscedastic design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.201	0.183	0.194	0.192	0.199	0.206	0.209	0.204	0.211	0.213	0.206	0.2	0.195	0.207	0.176	0.199	0.197
N=32	0.325	0.322	0.309	0.306	0.303	0.363	0.362	0.338	0.336	0.345	0.35	0.316	0.335	0.337	0.297	0.32	0.344
N=64	0.562	0.559	0.548	0.554	0.55	0.557	0.555	0.565	0.564	0.554	0.558	0.545	0.556	0.556	0.562	0.572	0.563
sand-*t*																	
N=16		0.161	0.167	0.17	0.174								0.183	0.193		0.184	0.189
N=32		0.287	0.27	0.262	0.266								0.278	0.306		0.284	0.314
N=64		0.556	0.522	0.527	0.53								0.528	0.558		0.538	0.553

Normal uncorrelated heteroscedastic design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.132	0.101	0.127	0.115	0.125	0.081	0.077	0.081	0.08	0.08	0.076	0.132	0.126	0.087	0.117	0.116	0.092
N=32	0.275	0.223	0.26	0.253	0.266	0.202	0.203	0.206	0.194	0.195	0.202	0.245	0.249	0.207	0.237	0.267	0.222
N=64	0.567	0.587	0.557	0.57	0.566	0.532	0.538	0.542	0.533	0.541	0.554	0.556	0.577	0.541	0.558	0.584	0.555
sand-*t*																	
N=16		0.089	0.117	0.107	0.122								0.138	0.098		0.135	0.115
N=32		0.225	0.253	0.226	0.257								0.223	0.221		0.237	0.228
N=64		0.567	0.534	0.534	0.538								0.549	0.569		0.543	0.565

Normal correlated heteroscedastic design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.137	0.112	0.139	0.148	0.142	0.108	0.105	0.109	0.105	0.111	0.106	0.146	0.149	0.124	0.139	0.149	0.127
N=32	0.293	0.298	0.301	0.26	0.279	0.266	0.257	0.269	0.265	0.276	0.268	0.231	0.266	0.284	0.235	0.272	0.28
N=64	0.402	0.304	0.401	0.4	0.405	0.27	0.289	0.289	0.292	0.304	0.295	0.387	0.393	0.303	0.393	0.399	0.324
sand-*t*																	
N=16		0.141	0.161	0.169	0.169								0.174	0.16		0.171	0.166
N=32		0.216	0.214	0.159	0.197								0.178	0.196		0.174	0.199
N=64		0.406	0.462	0.453	0.461								0.473	0.415		0.468	0.418

Individual raw power results

Regular design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.213	0.079	0.059	0.153	0.103	0.236	0.173	0.165	0.173	0.213	0.17	0.247	0.212	0.189	0.241	0.212	0.195
N=32	0.394	0.274	0.25	0.351	0.307	0.376	0.354	0.349	0.357	0.372	0.357	0.392	0.37	0.364	0.397	0.373	0.37
N=64	0.609	0.563	0.551	0.599	0.576	0.613	0.603	0.597	0.598	0.613	0.607	0.605	0.602	0.605	0.612	0.606	0.605
sand-*t*																	
N=16		0.138	0.113	0.23	0.177								0.229	0.214		0.233	0.214
N=32		0.314	0.297	0.403	0.36								0.379	0.371		0.377	0.361
N=64		0.582	0.566	0.61	0.586								0.586	0.601		0.59	0.595

Normal uncorrelated design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.086	0.033	0.009	0.132	0.033	0.162	0.122	0.114	0.122	0.149	0.121	0.203	0.169	0.114	0.209	0.176	0.114
N=32	0.338	0.207	0.195	0.319	0.263	0.348	0.319	0.316	0.318	0.338	0.321	0.351	0.332	0.321	0.353	0.33	0.321
N=64	0.58	0.494	0.48	0.575	0.538	0.584	0.573	0.576	0.577	0.571	0.572	0.583	0.572	0.569	0.577	0.569	0.558
sand-*t*																	
N=16		0.142	0.089	0.281	0.156								0.243	0.187		0.246	0.182
N=32		0.274	0.277	0.385	0.338								0.348	0.336		0.349	0.341
N=64		0.532	0.522	0.587	0.562								0.556	0.564		0.567	0.568

Normal correlated design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.079	0.049	0.012	0.13	0.034	0.175	0.142	0.129	0.139	0.167	0.136	0.216	0.19	0.149	0.222	0.192	0.146
N=32	0.206	0.127	0.117	0.179	0.152	0.207	0.182	0.181	0.184	0.198	0.187	0.213	0.201	0.189	0.214	0.191	0.187
N=64	0.329	0.26	0.259	0.323	0.296	0.328	0.308	0.306	0.313	0.322	0.308	0.336	0.33	0.314	0.332	0.32	0.314
sand-*t*																	
N=16		0.161	0.067	0.263	0.135								0.252	0.22		0.259	0.214
N=32		0.166	0.17	0.235	0.206								0.215	0.2		0.213	0.191
N=64		0.301	0.292	0.355	0.327								0.337	0.318		0.334	0.319

Regular heteroscedastic design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.245	0.074	0.06	0.163	0.113	0.248	0.209	0.194	0.21	0.233	0.207	0.263	0.237	0.217	0.252	0.237	0.229
N=32	0.376	0.239	0.212	0.306	0.276	0.396	0.365	0.354	0.365	0.379	0.361	0.373	0.366	0.367	0.357	0.37	0.376
N=64	0.576	0.534	0.494	0.545	0.528	0.614	0.598	0.587	0.593	0.612	0.603	0.564	0.562	0.594	0.562	0.583	0.608
sand-*t*																	
N=16		0.149	0.126	0.226	0.183								0.251	0.232		0.252	0.242
N=32		0.261	0.249	0.321	0.29								0.339	0.349		0.356	0.357
N=64		0.511	0.471	0.527	0.51								0.528	0.555		0.544	0.566

Normal uncorrelated heteroscedastic design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.195	0.038	0.04	0.154	0.093	0.184	0.141	0.125	0.142	0.173	0.144	0.251	0.221	0.167	0.247	0.223	0.172
N=32	0.327	0.192	0.2	0.288	0.258	0.32	0.298	0.286	0.291	0.319	0.295	0.335	0.318	0.291	0.332	0.325	0.304
N=64	0.623	0.534	0.525	0.607	0.579	0.633	0.621	0.622	0.625	0.626	0.621	0.622	0.621	0.626	0.61	0.629	0.632
sand-*t*																	
N=16		0.141	0.131	0.28	0.202								0.294	0.232		0.293	0.239
N=32		0.255	0.261	0.347	0.313								0.33	0.308		0.333	0.314
N=64		0.549	0.53	0.606	0.579								0.602	0.617		0.602	0.613

Normal correlated heteroscedastic design

	Case resampling					**Residual resampling**						**Score resampling**					
naïve-*t*	non	sand	sand.x	naive	naive.x	full.non	full.naive	null.non	null.naive	raw.non	raw.naive	non	naive	sand	R.non	R.naive	R.sand
N=16	0.096	0.013	0.01	0.056	0.023	0.1	0.074	0.067	0.079	0.093	0.08	0.146	0.109	0.096	0.136	0.116	0.101
N=32	0.365	0.213	0.209	0.323	0.279	0.345	0.327	0.317	0.329	0.342	0.326	0.363	0.347	0.324	0.347	0.342	0.335
N=64	0.436	0.304	0.321	0.416	0.374	0.384	0.37	0.362	0.365	0.372	0.372	0.442	0.447	0.397	0.434	0.441	0.404
sand-*t*																	
N=16		0.07	0.044	0.163	0.104								0.182	0.148		0.177	0.157
N=32		0.271	0.286	0.38	0.34								0.367	0.335		0.36	0.339
N=64		0.406	0.408	0.485	0.455								0.482	0.446		0.48	0.462

Bibliography

Anderson, M. & Legendre, P. (1999). An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation* **62**, 271–303.

Anderson, M. & Robinson, J. (2001). Permutation tests for linear models. *Aust. N.Z. J. Stat.* **43**, 75–88.

Bose, A. & Chatterjee, S. (2002). Comparison of bootstrap and jackknife variance estimators in linear regression: Second order results. *Statistica Sinica* **12**, 575–598.

Brzezniak, Z. & Zastawniak, T. (1999). *Basic Stochastic Processes*. Springer.

Chernick, M. (2008). *Bootstrap Methods: a Guide for Practitioners and Researchers*. second edition. John Wiley and Sons, Inc.

Davison, A. & Hinkley, D. (1997). *Bootstrap Methods and their Application*. Cambridge University Press. Chapters 4 and 6.

Draper, N. & Smith, H. (1966). *Applied Regression Analysis*. John Wiley and Sons, Inc.

Freedman, D. (1981). Bootstrapping regression models. *The Annals of Statistics* **9**, 1218–1228.

Freedman, D. & Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *J. Bus. Econom. Statist.* **1**, 292–298.

Friedl, H. & Stadlober, E. (1997). Resampling methods in generalized linear models useful in environmetrics. *Environmetrics* **8**, 441–457.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag. Chapter 4.

Hall, P. & Wilson, S. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757–762.

Hjorth, J. (1994). *Computer Intensive Statistical Methods: Validation Model Selection and Bootstrap*. Chapman and Hall. Chapter 6.

Huh, M.-H. & Jhun, M. (2001). Random permutation testing in multiple linear regression. *Commun. Statist.- Theory Meth.* **30**, 2023–2032.

Kennedy, P. (1995). Randomization tests in econometrics. *J. Bus. Econom. Statist.* **13**, 85–94.

Liu, R. & Singh, K. (1992). Efficiency and robustness in resampling. *The Annals of Statistics* **20**, 370–384.

Lloyd, C. (2005). On comparing the accuracy of competing tests of the same hypotheses from simulation data. *The Journal of Statistical Planning and Inference* **128**, 497–508.

MacKinnon, J. (2006). Bootstrap methods in econometrics. *The Economic Record* **82**, S2–S18.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics* **21**, 255–285.

Manly, B. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman and Hall. 2nd edition.

Moulton, L. & Zeger, S. (1991). Bootstrapping generalized linear models. *Computational Statistics and Data Analysis* **11**, 53–63. North-Holland.

Seber, G. (1977). *Linear Regression Analysis*. John Wiley and Sons.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley and Sons, Inc.

ter Braak, C. (1992). Permutation versus bootstrap significance tests in multiple regression and anova. *In: Bootstrapping and Related Techniques* pp. 79–86.

Wu, C. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics* **14**, 1261–1295.