Senior Thesis in Mathematics

Permutation Tests for Multiple Linear Regression Models

Author:
Julie Ye

Advisor:
Dr. Johanna S. Hardin

Submitted to Pomona College in Partial Fulfillment
of the Degree of Bachelor of Arts

April 15, 2023

Abstract

This paper examines the implications of different permutation methods in the context of testing the significance of a single partial regression coefficient in a multiple linear regression model. The methods compared were permutation of the response variable, permutation of the predictor variable in question, and two methods of permutation of residuals—one under the reduced model and one under the full model. We used simulations to empirically explore the effects of (1) the sample size, (2) the magnitude of the covariable's parameter, and (3) the correlation between the predictor variables. It was found that there were no substantial differences in size or power among the four permutation methods under a normal errors structure. The simulation results showed that power increases with larger sample sizes, larger predictor variable coefficient magnitudes, and smaller correlation between the predictor variables.

Contents

Chapter 1

Statistical Inference

Statistics is the science of collecting, wrangling, visualizing, and analyzing data from a smaller sample in order to say something about a bigger population as a whole. In particular, *inferential statistics* use a sample intended to produce results that can be generalized to a larger population of interest.

1.1 Simple Linear Regression

In this chapter, we will be talking about linear models, which describe a linear relationship between a predictor (or explanatory) variable and the expected value of an outcome variable. The simplest form of this model, otherwise known as simple linear regression (SLR), involves exactly two variables and assumes the following model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where Y is the response variable, X is the predictor variable, and ε is the random error term. The parameters β_0 and β_1 describe the "true" intercept and slope respectively, but for this type of question, we are interested in the impact of X on predictions of Y. In fact, this leads into a technique called hypothesis testing, which can be broken down into:
Null Hypothesis: Denoted H_0, this is usually set up as the baseline and assumed to be true unless proven otherwise. In SLR, a null hypothesis that is commonly used is

$$H_0 : \beta_1 = 0.$$

In other words, there is no linear relationship between the predictor and response variable.

Alternative Hypothesis: Denoted H_a, this usually represents what we believe to be true and want to show. This tends to take the form of the parameter somehow not being equal to the value used in H_0. The alternative hypothesis to the null hypothesis stated above may look like:

$$H_a : \beta_1 \neq 0$$

In words, this would mean there is a non-zero linear relationship between the predictor variable and response variable. It could be a negative relationship or a positive relationship, since we did not specify whether $\beta_1 > 0$ or $\beta_1 < 0$, which are other viable alternative hypotheses.

Like most models, SLR comes with its own set of technical conditions that must be checked before proceeding with normal theory methods (i.e., the t-test) to ensure set Type I error rates (Kutner et al., 2005, pp. 9-10).

1.1.1 Technical Conditions of SLR

1. **Condition of linearity**: The relationship between the predictor variable and response variable should be linear; otherwise, a different model may make more sense to model a non-linear relationship. Thus, the data should follow the structure $Y = \beta_0 + \beta_1 X + \varepsilon$. In addition, we are assuming the mean of ε is 0. This is important because we want

   $$E[Y|X] = \mu_{Y|X} = \beta_0 + \beta_1 X,$$

   where μ is the mean. That is, the expected value of Y at some value of X is given by $\beta_0 + \beta_1 X$.

2. **Condition of independence**: Every observation should be independent from the rest. One observation should not carry any information that can be inferred by another observation in the dataset. A simple random sample tends to produce unbiased samples with independent observations that are representative of the population.

3. **Condition of constant variance**: The error terms from the model should not only have a mean of 0 but also have a constant variance of σ^2 for any value of X. If our data is showing signs of heteroskedasticity,

which is when variance is non-constant across different values of the predictor variable, it may be worthwhile to either transform the dataset (e.g., taking the log transformation) or use a different model.

4. **Condition of normal distribution**: The last assumption is that errors are distributed normally. Normal distributions are common and come with nice properties that are easier to work with.

1.1.2 Permutation Tests for SLR Models

In contrast to the t-test, which is used if all the technical conditions mentioned in Section 1.1.1 hold, we can apply a permutation test to assess the same hypotheses under less stringent conditions. Let's say we wanted to test whether the true slope of X, otherwise known as β_1, was 0 or positive. Then we would formulate the following hypotheses:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 > 0$$

First, fit a linear model on the data and record the original sample slope, b_1, and find the t-statistic associated with this value.

Then comes the permutation component. When doing a permutation test for an SLR model in the form of $E[Y] = \beta_0 + \beta_1 X$, note that permuting the predictor variable, X, has the same effect as permuting the response variable, Y, since these are the only variables in the model. By permuting either variable and reassigning different values of X to different values of Y, we are effectively breaking any possible relationship that may have existed between X and Y. In other words, the permutation forces the null hypothesis, H_0, to be true.

Without loss of generality, assume that we permute the response variable Y to get Y^*. From this permuted dataset, we fit a linear model of Y^* on X, find the sample slope b_1^*, and calculate the standardized sample slope statistic, $t^* = (b_1^* - 0)/\text{se}(b_1^*)$, where $\text{se}(b_1^*)$ is the standard error of b_1^* calculated from the permuted dataset. Repeated permutations of the data result in the construction of a null distribution of t^* values (since we are forcing H_0 to be true through permutations), from which a p-value can be obtained by counting how many of the simulated t^*-statistics were at least as extreme as t_{obs}. Note that although we use the letter t to describe the test statistic,

we do *not* use the t-distribution to calculate p-values mathematically for permutation tests. In this paper, we will use a significance level of $\alpha = 0.05$ and reject H_0 if $p < 0.05$.

What if the expected value of Y is a linear function of *multiple* predictor variables? Then, this would require a multiple linear regression model.

1.2 Multiple Linear Regression

Consider the following linear model involving multiple predictor variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon$$

where Y is the response variable; X_1, X_2, ..., X_{p-1} are the predictor variables; and ε is the random error term. In multiple linear regression (MLR), we have $p-1$ predictor variables thanks to the true intercept, β_0, which adds up to p parameters in total.

1.2.1 Types of Multiple Regression

Qualitative predictor variables: Predictor variables do not always have to be continuous. If you have categorical variables in your model, you would need to write it as one (or more) binary variables. For instance:

$$X_2 = \begin{cases} 0 & \text{if weekend} \\ 1 & \text{if weekday} \end{cases}$$

If a categorical variable has k levels, you generally need $k-1$ "dummy" variables to represent all possible values. The case where all dummy variables equal 0 represents the k-th level.

Interaction effects: In some cases, the predictor variables in your model may interact with each other. Perhaps X_2 has a different impact on Y depending on the level of X_1. A model with two predictor variables that takes into account their interaction may look like:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

where the interaction is multiplicative.

Polynomial regression: The best model may also be a function of a polynomial of a predictor variable:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

1.2.2 Permutation Tests for MLR Models

For the sake of simplicity, we will consider only X_1 and X_2 as the predictor variables in our model of interest. That is,

$$Y = \beta_0 + \beta_{1\cdot 2} X_1 + \beta_{2\cdot 1} X_2 + \varepsilon$$

where we borrow the following notation used by Anderson and Legendre (1999, p. 273): $\beta_{1\cdot 2}$ denotes the partial regression coefficient for the relationship between X_1 and Y while taking into consideration the effect of X_2, and similarly for $\beta_{2\cdot 1}$.

Now we have a new challenge. Suppose that we are trying to use X_1 and X_2 to predict the response variable Y. Somehow, we know that X_1 is essential to our model, but we want to check whether we actually also need X_2 in our model, or if simply regressing Y against X_1 is enough. Then, we would formulate the following hypotheses:

$$H_0 : \beta_{2\cdot 1} = 0$$
$$H_a : \beta_{2\cdot 1} \neq 0$$

H_0 says that the variable X_2 does not add any additional information to our model in predicting Y, whereas H_a proposes that X_2 adds information in predicting Y since $\beta_{2\cdot 1}$ is not equal to 0, even when X_1 is in the model.

However, we now have three variables to pick from for permutations. The crucial question is, which variable should be permuted to test our hypotheses? Or should we use a different permutation strategy?

In the next chapter, we will investigate one proposed method that involves permuting the Y values.

Chapter 2

Permutation of the Response Variable

To perform permutation tests for MLR models, Manly (1986, 1997) proposed randomly assigning values of the response variable, Y, to the sets of X values. The rationale behind permuting the response variable is that if Y is independent from all Xs, then Y values should be equally likely to occur with any sets of X values.

2.1 Procedure

The response variable permutation strategy as described by Anderson and Legendre (1999) is given by the following algorithm:

1. Regress the response variable Y on the predictor variables X_1 and X_2, which calculates the sample intercept b_0, sample slopes $b_{1\cdot2}$ and $b_{2\cdot1}$, and residuals $R_{Y|1,2}$:

   $$Y \sim X_1 + X_2 \quad \Longrightarrow \quad Y = b_0 + b_{1\cdot2}X_1 + b_{2\cdot1}X_2 + R_{Y|1,2}$$

   Use $b_{2\cdot1}$ to calculate the observed t-statistic, where $t_{\text{obs}} = b_{2\cdot1}/\text{se}(b_{2\cdot1})$. Save this t_{obs} for the last step.

2. Permute Y to get Y^*.

3. Regress Y^* on X_1 and X_2, which outputs b_0^*, $b_{1\cdot2}^*$, $b_{2\cdot1}^*$, and $R_{Y|1,2}^*$:

   $$Y^* \sim X_1 + X_2 \quad \Longrightarrow \quad Y^* = b_0^* + b_{1\cdot2}^*X_1 + b_{2\cdot1}^*X_2 + R_{Y|1,2}^*$$

The statistics and residuals from this equation have asterisks (*) to signify that they come from modified (permuted) data. Save $b^*_{2\cdot1}$ and use it to calculate t^*, where $t^* = b^*_{2\cdot1}/\text{se}(b^*_{2\cdot1})$.

4. Repeat Steps (2)-(3) many times to obtain a distribution of t^* values from many permutations.

5. Calculate the p-value of t_{obs}. Reject H_0 if $p < 0.05$.

2.2 Considerations of Manly's Method

The impact of this permutation strategy on the relationships between the variables can be summarized as follows:

Broken Relationships	*Preserved Relationships*
X_1 & Y	X_1 & X_2
X_2 & Y	

We infer that Manly's method may be appropriate if X_1 and X_2 are related, since it may be of interest to preserve the relationship between them in order to accurately make use of the data.

On the other hand, permuting Y gets in the way of isolating the effect of X_2 when X_1 is in the model. This was not a problem in SLR, which only had one predictor variable, X, and permuting X was algorithmically equivalent to permuting Y. Here, however, we are dealing with both X_1 and X_2. Since X_1 is already contributing to Y in the model, we are unable to isolate the effect of X_2 by simply permuting Y.

What if we wanted to do a better job of isolating the additional impact of X_2 on a model that already includes X_1? In the next chapter, we will discuss a method that proposes permuting the variable of interest, X_2.

Chapter 3

Permutation of the Predictor Variable

Draper and Stoneman (1966) proposed permuting values of the predictor variable of interest—in our case, X_2—to perform permutation tests. Draper and Stoneman's method may be appropriate in a situation where the treatment is randomly assigned and the driving question is whether this treatment variable had any impact on the response Y, after taking into consideration the effect of other covariates in the model (Manly, 1997). Since a randomly assigned treatment variable will inherently be independent from the pre-existing covariates, permuting the treatment variable will not violate any relationships that may have existed between the treatment and the other covariates, as there was none to begin with.

3.1 Procedure

The predictor variable permutation strategy is given by the following algorithm:

1. Regress the response variable Y on the predictor variables X_1 and X_2, which calculates the sample intercept b_0, sample slopes $b_{1\cdot2}$ and $b_{2\cdot1}$, and residuals $R_{Y|1,2}$:

   $$Y \sim X_1 + X_2 \quad \Longrightarrow \quad Y = b_0 + b_{1\cdot2}X_1 + b_{2\cdot1}X_2 + R_{Y|1,2}$$

   Use $b_{2\cdot1}$ to calculate the observed t-statistic, where $t_{\text{obs}} = b_{2\cdot1}/\text{se}(b_{2\cdot1})$. Save this t_{obs} for the last step.

2. Permute X_2 to get X_2^*.

3. Regress Y on X_1 and X_2^*, which outputs b_0^*, $b_{1\cdot 2}^*$, $b_{2\cdot 1}^*$, and $R_{Y|1,2}^*$:

$$Y \sim X_1 + X_2 \quad \Longrightarrow \quad Y = b_0^* + b_{1\cdot 2}^* X_1 + b_{2\cdot 1}^* X_2^* + R_{Y|1,2}^*$$

   The statistics and residuals from this equation have asterisks (*) to signify that they come from modified (permuted) data. Save $b_{2\cdot 1}^*$ and use it to calculate t^*, where $t^* = b_{2\cdot 1}^*/\text{se}(b_{2\cdot 1}^*)$.

4. Repeat Steps (2)-(3) many times to obtain a distribution of t^* values from many permutations.

5. Calculate the p-value of t_{obs}. Reject H_0 if $p < 0.05$.

3.2 Considerations of Draper and Stoneman's Method

The impact of this permutation strategy on the relationships between the variables can be summarized as follows:

Broken Relationships	*Preserved Relationships*
X_1 & X_2	X_1 & Y
X_2 & Y	

Note that this method of permutation preserves the relationship between X_1 and Y, which is nice since we want to isolate the additional impact of X_2 in our model.

Draper and Stoneman's method may be appropriate if X_2 is independent of X_1 (or however many other covariates there may be), such as in situations akin to a clinical trial. For instance, imagine covariates $X_1, X_2, \ldots, X_j$ to be a patient's demographic or biographic information, and X_{j+1} to be a randomly assigned treatment that has no relation to the covariates. So, $X_1, X_2, \ldots, X_j$ could be someone's age, fitness level, height, etc., while X_2 could be a randomly assigned drug. Since the treatment is random, a person's health background will have nothing to do with the treatment they get, so

breaking the relationship between X_1 and X_2 by permuting X_2 would not be problematic.

On the other hand, permuting X_2 may be unsuitable when X_1 and X_2 are indeed related. Rearranging values of X_2 would ignore any relationship between the predictor variables; therefore, the predictor variable permutation strategy will not be able to accurately measure the additional impact of having X_2 in the model.

What if we wanted a method that not only preserves the relationship among predictor variables X_1 and X_2 but *also* preserves the relationship between X_1 and Y? Is there such a method?

Chapter 4

Permutation Of Residuals

The third technique discussed in this paper involves permuting neither the response variable nor the predictor variable of interest. Instead, both Freedman and Lane (1983) and ter Braak (1992) have proposed algorithms that randomly shuffle the residuals from a linear model. While there are similarities between the two algorithms, slight differences may result in different power depending on the characteristics of the data.

4.1 Permutation of Residuals Under the Reduced Model

The permutation strategy by Freedman and Lane (1983) focuses on the observed residuals from a restricted model, where the coefficient(s) of interest are set to 0, as the permutable units for the test. Anderson and Legendre (1999) refer to this method as the permutation of residuals under the reduced model. This name comes from the part of the procedure that relies on forcing the null hypothesis ($\beta_{2 \cdot 1} = 0$) to be true, because $\beta_{2 \cdot 1} = 0$ reduces the model to:

$$Y = \beta_0' + \beta_1' X_1 + \varepsilon'$$

where β_0' is the intercept and β_1' is the regression coefficient for the relationship between X_1 and Y without controlling for the effect of X_2. The random error ε' is treated as the stochastic element and separated from each value of Y, as you will see in Section 4.1.1.

4.1.1 Procedure

The reduced model residual permutation strategy, as described by Anderson and Legendre (1999), is given by the following algorithm:

1. Regress the response variable Y on the predictor variables X_1 and X_2, which calculates the sample intercept b_0, sample slopes $b_{1\cdot2}$ and $b_{2\cdot1}$, and residuals $R_{Y|1,2}$:

$$Y \sim X_1 + X_2 \quad \Longrightarrow \quad Y = b_0 + b_{1\cdot2}X_1 + b_{2\cdot1}X_2 + R_{Y|1,2}$$

   Use $b_{2\cdot1}$ to calculate the observed t-statistic, where $t_{\text{obs}} = b_{2\cdot1}/\text{se}(b_{2\cdot1})$. Save this t_{obs} for the last step.

2. Regress Y on X_1 alone, which outputs b'_0, b'_1, and $R_{Y|1}$:

$$Y \sim X_1 \quad \Longrightarrow \quad Y = b'_0 + b'_1X_1 + R_{Y|1} \tag{4.1}$$

   Here, the statistics have prime symbols ($'$) to signify that they come from the reduced model. Keep b'_0, b'_1, and $R_{Y|1}$ for later.

3. Permute $R_{Y|1}$ to get $R^*_{Y|1}$.

4. Add the permuted residuals $R^*_{Y|1}$ to the fitted values, using b'_0 and b'_1 from Equation (4.1), and calculate synthetic values for the response variable, denoted Y^*:

$$Y^* = b'_0 + b'_1X_1 + R^*_{Y|1}$$

5. Regress Y^* on X_1 and X_2, which outputs b^*_0, $b^*_{1\cdot2}$, $b^*_{2\cdot1}$, and $R^*_{Y|1,2}$:

$$Y^* \sim X_1 + X_2 \quad \Longrightarrow \quad Y^* = b^*_0 + b^*_{1\cdot2}X_1 + b^*_{2\cdot1}X_2 + R^*_{Y|1,2}$$

   The statistics and residuals from this equation have asterisks (*) to signify that they come from modified (synthetic reponse variable) data. Save $b^*_{2\cdot1}$ and use it to calculate t^*, where $t^* = b^*_{2\cdot1}/\text{se}(b^*_{2\cdot1})$.

6. Repeat Steps (3)-(5) many times to obtain a distribution of t^* values from many permutations.

7. Calculate the p-value of t_{obs}. Reject H_0 if $p < 0.05$.

4.1.2 Considerations of Freedman and Lane's Method

The impact of this permutation strategy on the relationships between the variables can be summarized as follows:

Broken Relationships	*Preserved Relationships*
X_2 & Y (if X_1 & X_2	X_1 & X_2
are not highly collinear)	X_1 & Y

In order to use this procedure, Freedman and Lane (1983) emphasize that the data should not have any extreme outliers, and X_1 and X_2 should not be highly collinear. This is because the null hypothesis, $H_0 : \beta_{2\cdot1} = 0$, cannot be enforced when the correlation between X_1 and X_2 is too high.

To elucidate this concept, we use $\rho(X_1, X_2)$ to denote the correlation coefficient between the random variables X_1 and X_2. The correlation coefficient informs us to what extent the variables are linearly associated and ranges from -1 to 1. (Note that when $H_0 : \beta_{2\cdot1} = 0$ is true, the correlation coefficient of Y^* and X_2 should be near 0, i.e., $\rho(Y^*, X_2) = 0$.) We claim that the correlation of Y^* and X_2 is dependent on that of X_1 and X_2.

Claim. The correlation between Y^* and X_2 is dependent on the correlation between X_1 and X_2.

Proof.

$$
\begin{aligned}
\rho(Y^*, X_2) &= \frac{cov(Y^*, X_2)}{\sqrt{var(Y^*)var(X_2)}} \\
\text{where } cov(Y^*, X_2) &= cov(b_0' + b_1' X_1 + R^*_{Y|1}, X_2) \\
&= cov(b_0', X_2) + cov(b_1' X_1, X_2) + cov(R^*_{Y|1}, X_2) \\
&= cov(b_1' X_1, X_2) && (4.2) \\
&\approx b_1' \cdot cov(X_1, X_2) && (4.3) \\
&= b_1' \cdot \rho(X_1, X_2)\sqrt{var(X_1)var(X_2)} \\
var(Y^*) &= var(b_0' + b_1' X_1 + R^*_{Y|1}) \\
&\approx (b_1')^2 var(X_1) + var(R^*_{Y|1}) && (4.4) \\
var(R^*_{Y|1}) &= var(R_{Y|1})
\end{aligned}
$$

$$
\begin{aligned}
&= var(Y - b_0' - b_1' X_1) \\
&\approx var(Y) + (b_1')^2 var(X_1) - 2b_1' cov(Y, X_1) \qquad (4.5) \\
cov(Y, X_1) &= cov(\beta_0 + \beta_{1\cdot 2} X_1 + \beta_{2\cdot 1} X_2 + \varepsilon, X_1) \\
&= cov(\beta_0, X_1) + cov(\beta_{1\cdot 2} X_1, X_1) \\
&\quad + cov(\beta_{2\cdot 1} X_2, X_1) + cov(\varepsilon, X_1) \\
&= \beta_{1\cdot 2} var(X_1) + \beta_{2\cdot 1} cov(X_2, X_1) \qquad (4.6) \\
\text{which means } var(Y^*) &\approx (b_1')^2 var(X_1) + var(Y) + (b_1')^2 var(X_1) \\
&\quad - 2b_1'[\beta_{1\cdot 2} var(X_1) + \beta_{2\cdot 1} cov(X_2, X_1)] \\
&= 2(b_1')^2 var(X_1) + var(Y) \\
&\quad - 2b_1'[\beta_{1\cdot 2} var(X_1) + \beta_{2\cdot 1} cov(X_2, X_1)] \\
&= 2b_1'(b_1' - \beta_{1\cdot 2}) var(X_1) + var(Y) - 2b_1' \beta_{2\cdot 1} cov(X_2, X_1) \\
&= 2b_1'[(b_1' - \beta_{1\cdot 2}) var(X_1) - \beta_{2\cdot 1} cov(X_1, X_2)] + var(Y),
\end{aligned}
$$

so putting it all together gives us

$$
\begin{aligned}
\rho(Y^*, X_2) &= \frac{cov(Y^*, X_2)}{\sqrt{var(Y^*) var(X_2)}} \\
&\approx \frac{b_1' \cdot \rho(X_1, X_2) \sqrt{var(X_1) var(X_2)}}{\sqrt{var(Y^*) var(X_2)}} \\
&= \frac{b_1' \cdot \rho(X_1, X_2) \sqrt{var(X_1)}}{\sqrt{var(Y^*)}} \\
&= \frac{b_1' \cdot \rho(X_1, X_2) \sqrt{var(X_1)}}{\sqrt{2b_1'[(b_1' - \beta_{1\cdot 2}) var(X_1) - \beta_{2\cdot 1} cov(X_1, X_2)] + var(Y)}} \qquad (4.7)
\end{aligned}
$$

Note that b_1' is a random variable because it is a statistic, so Equations (4.3) to (4.5) are all approximate. In Equation (4.2), we assume that b_0' and $R^*_{Y|1}$ are both independent of X_2, so $cov(b_0', X_2) = cov(R^*_{Y|X_1}) = 0$. Similarly, we also assume independence in Equation (4.6), leading to $cov(\beta_{2\cdot 1} X_2, X_1) = cov(\varepsilon, X_1) = 0$. These approximations are validated empirically; see Figure 4.1. □

Thus, $\rho(Y^*, X_2)$ is dependent on $\rho(X_1, X_2)$. Due to this dependency, if X_1 and X_2 are too highly correlated, Freedman and Lane's algorithm will not produce Y^* values such that $\rho(Y^*, X_2) = 0$ in Step (4). This is something to

Simulation Values	
n	20, 40, 80
$\beta_{1 \cdot 2}$	0.5, 1.0, 2.5
$\beta_{2 \cdot 1}$	0.00, 0.25, 0.50, 0.75, 1.00
$\rho(X_1, X_2)$	0.0, 0.5, 0.9

Table 4.1: Different iterations of values used to generate the datasets for assessing the approximation of $\rho(Y^*, X_2)$. Every possible combination of these values was used, resulting in 135 different parameter configurations for the simulations.

keep in mind because the null hypothesis, $\beta_{2 \cdot 1} = 0$, cannot be enforced when $\rho(Y^*, X_2)$ is not driven to 0. Under this condition, the algorithm would not be able to produce a null sampling distribution of t^*.

We performed simulations in R to empirically assess Equation (4.7) with regard to the sample size (n), the size of the covariable's parameter ($\beta_{1 \cdot 2}$), the size of the coefficient of interest ($\beta_{2 \cdot 1}$), and the correlation between X_1 and X_2 ($\rho(X_1, X_2)$), as seen in Table 4.1. Simulations were run for all combinations of these factors, which summed up to 135 different parameter configurations. For each configuration, 1,000 datasets were generated using the `mvrnorm()` function from the MASS library. In each dataset, values of Y^* were calculated from a permuted version of the dataset 1,000 times for a total of 1,000 sample correlations between Y^* and X_2 per dataset. The average of these sample correlations was recorded for each configuration and denoted by $\widehat{corr}(Y^*, X_2)$ in Figure 4.1.

There was almost perfect overlap between the theoretical $\widehat{corr}(Y^*, X_2)$ (in red) from our formula and the empirical average $\widehat{corr}(Y^*, X_2)$ (in blue) in all of the subplots, which means we were able to successfully demonstrate that $\rho(Y^*, X_2)$ is highly dependent on the correlation between X_1 and X_2.

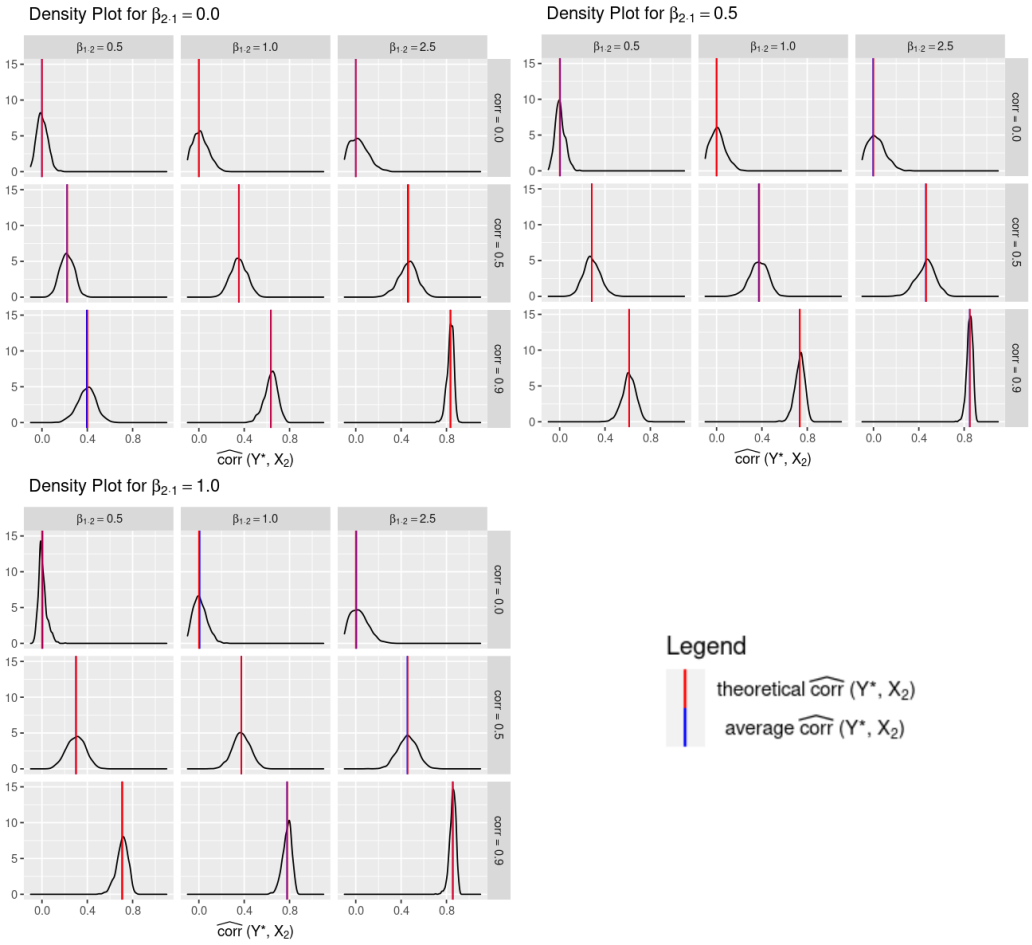


Figure 4.1: Density plots for datasets of $n = 80$ observations generated from $\beta_{2\cdot1} = 0.0$, $\beta_{2\cdot1} = 0.5$, and $\beta_{2\cdot1} = 1.0$. $\widehat{corr}(Y^*, X_2)$ indicates the average of the average correlation between Y^* and X_2 from permutations of each dataset. The theoretical $\widehat{corr}(Y^*, X_2)$ that is approximated from Equation (4.7) is plotted in red, while the average $\widehat{corr}(Y^*, X_2)$ from the simulations is plotted in blue. Due to the overwhelming extent to which the empirical results aligned with the theoretical value calculated using Equation (4.7), it is rather difficult to identify the blue line.

On the other end of the spectrum, if X_1 and X_2 are entirely uncorrelated, then the method works, but it might be uninteresting. If $\rho(X_1, X_2) = 0$, we cannot have both $\rho(X_1, Y)$ and $\rho(X_2, Y)$ be big in magnitude, since X_1 and X_2 supposedly have no correlation.

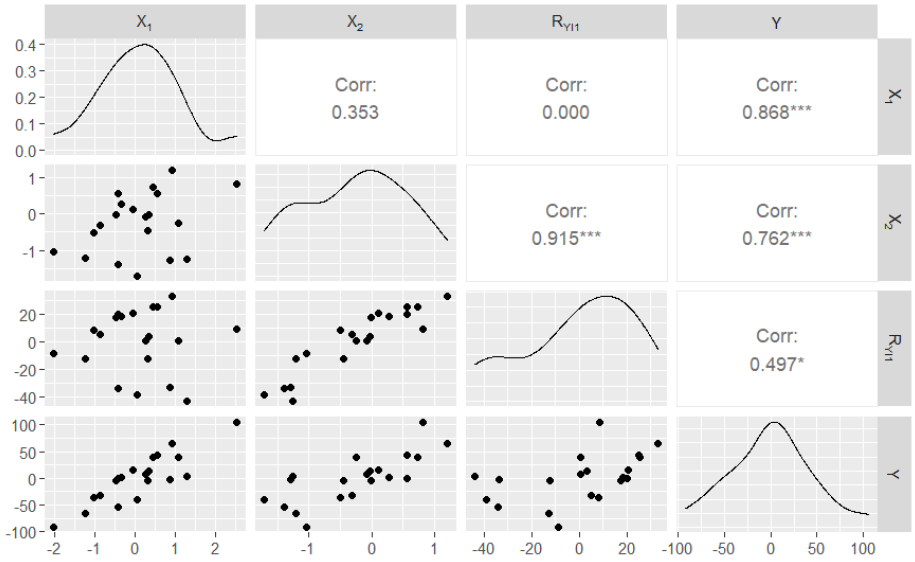
$$\begin{aligned} \rho(X_1, X_2) = 0 \ \& \ \rho(X_1, Y) = \text{high} \quad &\implies \quad \rho(X_2, Y) = \text{low} \\ \rho(X_1, X_2) = 0 \ \& \ \rho(X_2, Y) = \text{high} \quad &\implies \quad \rho(X_1, Y) = \text{low} \end{aligned}$$

Hence, as long as X_1 and X_2 are not highly collinear, we can demonstrate how the process of calculating new Y^* values will successfully break any possible relationship found between X_2 and Y in X_2 and Y^*. Moreover, this method preserves the relationships between Y and X_1, X_1 and X_2, and among all the predictor variables.

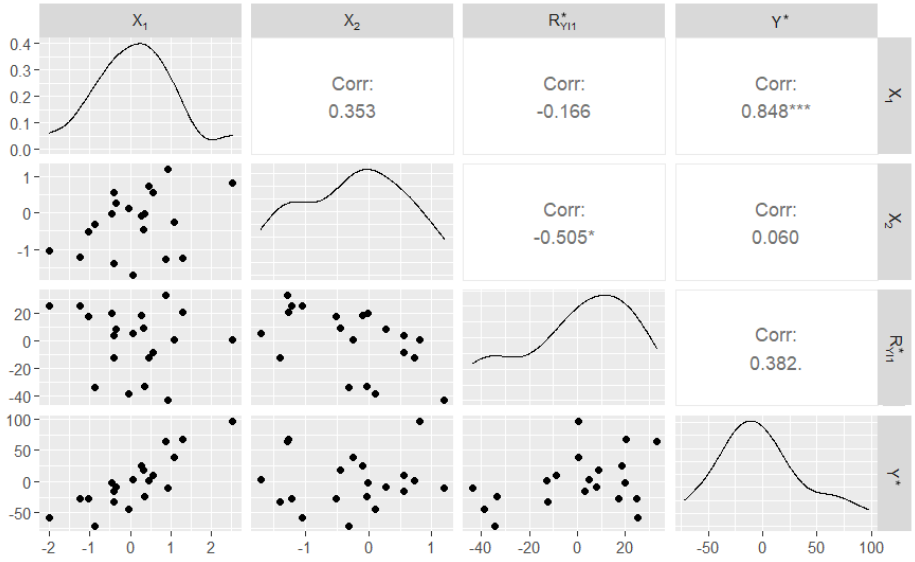
Y^* and X_1 maintain similar degrees of collinearity as Y and X_1 because Y and Y^* are regressed against X_1 in both the full and reduced models respectively. Since one of the assumptions of linear regression is that errors are normally distributed with constant variance, shuffling the residuals $R_{Y|1}$ and adding them back to calculate new Y^* values would emulate a similar relationship between that of Y and X_1.

To provide a visual demonstration of which relationships are preserved and which are broken through Freedman and Lane's permutation strategy, Figure 4.2 displays the linear relationships between X_1, X_2, residuals from the reduced model, and the response variable—either the original Y values, or Y^* calculated after one permutation of the residuals from the reduced model. For the purposes of this demonstration, this dataset was generated from a normal distribution with $\beta_0 = 5$, $\beta_{1\cdot2} = 30$, $\beta_{2\cdot1} = 30$, $\mu = 0$, $\sigma = 5$, and $\rho(X_1, X_2) = 0.4$ (which is considered a medium-low correlation). On the diagonal are the univariate density plots for each variable. By looking at the scatterplot between X_2 and Y in Figure 4.2a, there is evidence of a linear relationship between the two variables. However, after permuting the residuals from the reduced model to calculate the synthetic response variable Y^*, in Figure 4.2b, we find that that relationship seems to have disappeared in the scatterplot between X_2 and Y^*.

Figures 4.3 and 4.4 show different angles of a 3D scatterplot between X_1, X_2, and either Y or Y^*. These scatterplots display how the linear relationship between X_1 and Y is generally maintained between X_1 and Y^*, while the relationship between X_2 and Y is practically nullified between X_2 and Y^*. The color of the points indicate how positive (red) or negative (blue) the associated residual from the reduced model is.

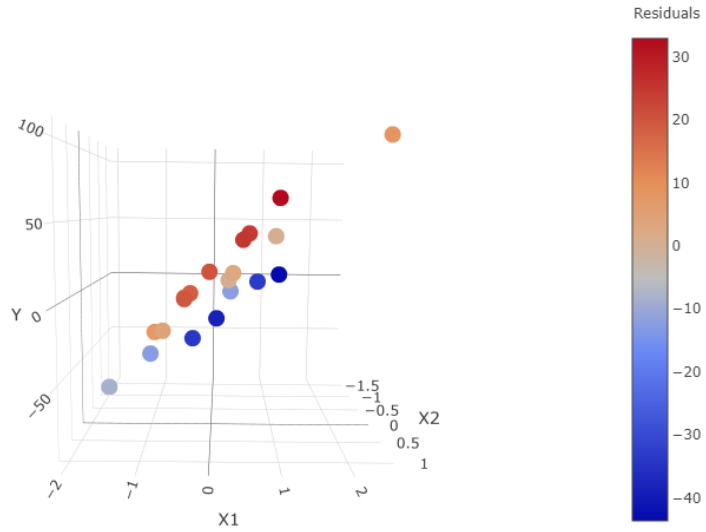


(a) Pairs plot of X_1, X_2, reduced model residuals, and Y

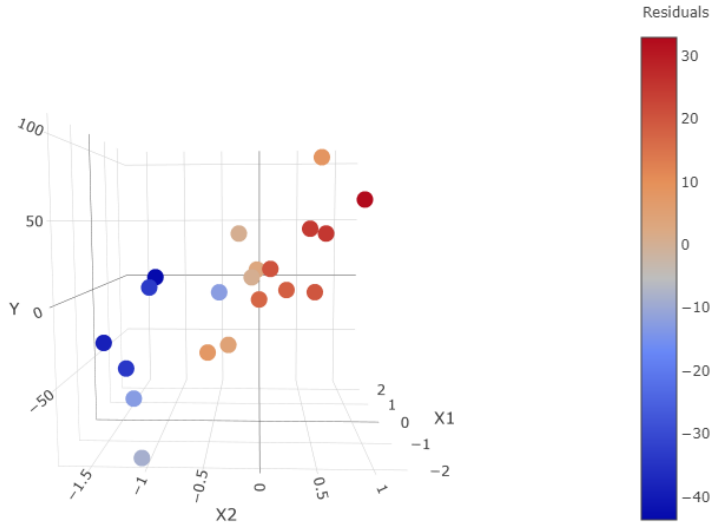


(b) Pairs plot of X_1, X_2, permuted reduced model residuals, and Y^*

Figure 4.2: (a) shows the scatterplot matrix of variables and unpermuted residuals from the reduced model. (b) shows the scatterplot matrix after permutation. The dataset was generated from a normal distribution with $\beta_0 = 5$, $\beta_{1\cdot 2} = 30$, $\beta_{2\cdot 1} = 30$, $\mu = 0$, $\sigma = 5$, and $\rho(X_1, X_2) = 0.4$. Note that $\rho(X_1, Y^*) = 0.848$ in (b) remains almost as high as $\rho(X_1, Y) = 0.868$ in (a), while $\rho(X_2, Y^*) = 0.060$ is almost entirely nullified in comparison to $\rho(X_2, Y) = 0.762$.

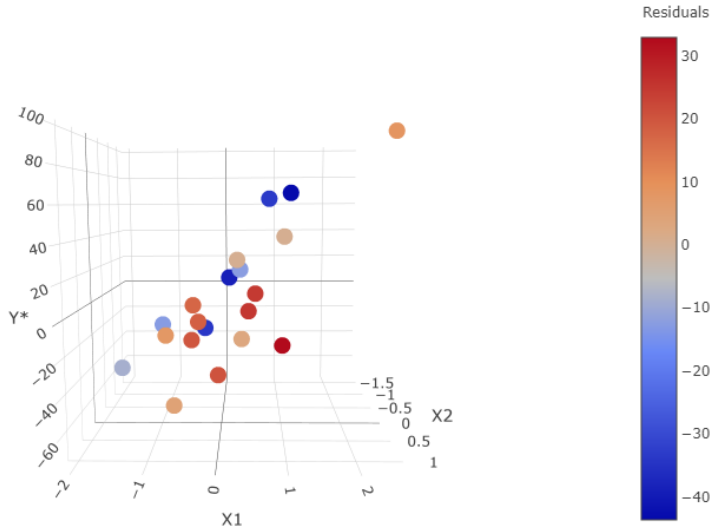


(a) 3D plot projected onto X_1 and Y

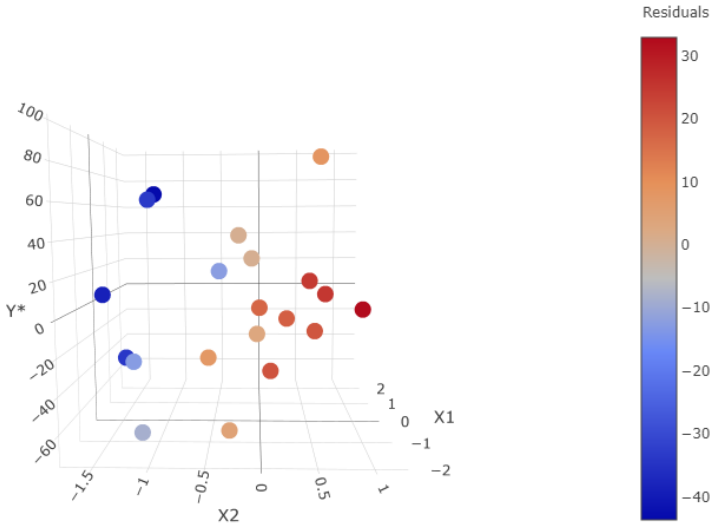


(b) 3D plot projected onto X_2 and Y

Figure 4.3: Two different views of a 3D scatterplot between X_1, X_2, and Y. The color of each point indicates how positive (red) or negative (blue) the original associated residual from the reduced model is. Both (a) and (b) display a strong correlation between X_1 and Y and between X_2 and Y, respectively.



(a) 3D plot projected onto X_1 and Y^*



(b) 3D plot projected onto X_2 and Y^*

Figure 4.4: Two different views of a 3D scatterplot between X_1, X_2, and Y^*. The color of the points indicates how positive (red) or negative (blue) the original associated residual from the reduced model is. (a) still shows a general positive linear relationship between X_1 and Y, imitating that found in Figure 4.3a, while the positive linear relationship between X_2 and Y in Figure 4.3b is no longer visible between X_2 and Y^* in (b).

4.2 Permutation of Residuals Under the Full Model

The following algorithm developed by ter Braak (1990, 1992) is the fourth and final strategy that we will discuss in this paper. Like Freedman and Lane's method in Section 4.1, ter Braak's strategy also permutes residuals, but this time, the residuals are obtained from the full model, where Y is regressed against X_1 and X_2:

$$Y = b_0 + b_{1\cdot2}X_1 + b_{2\cdot1}X_2 + R_{Y|1,2}$$

4.2.1 Procedure

The full model residual permutation strategy, as described by Anderson and Legendre (1999), is given by the following algorithm:

1. Regress the response variable Y on the predictor variables X_1 and X_2, which calculates the sample intercept b_0, sample slopes $b_{1\cdot2}$ and $b_{2\cdot1}$, and residuals $R_{Y|1,2}$:

   $$Y \sim X_1 + X_2 \quad \Longrightarrow \quad Y = b_0 + b_{1\cdot2}X_1 + b_{2\cdot1}X_2 + R_{Y|1,2} \tag{4.8}$$

   Keep b_0, $b_{1\cdot2}$, $b_{2\cdot1}$, and $R_{Y|1,2}$ for later. Use $b_{2\cdot1}$ to calculate the observed t-statistic, where $t_{\text{obs}} = b_{2\cdot1}/\text{se}(b_{2\cdot1})$. Save this t_{obs} for the last step.

2. Permute $R_{Y|1,2}$ to get $R^*_{Y|1,2}$.

3. Add permuted residuals $R^*_{Y|1,2}$ to the fitted values, using b_0, $b_{1\cdot2}$, and $b_{2\cdot1}$ from Equation (4.8), and calculate synthetic values for the response variable, denoted by Y^*:

   $$Y^* = b_0 + b_{1\cdot2}X_1 + b_{2\cdot1}X_2 + R^*_{Y|1,2}$$

4. Regress Y^* on X_1 and X_2, which outputs b^*_0, $b^*_{1\cdot2}$, $b^*_{2\cdot1}$, and $R^*_{Y|1,2}$:

   $$Y^* \sim X_1 + X_2 \quad \Longrightarrow \quad Y^* = b^*_0 + b^*_{1\cdot2}X_1 + b^*_{2\cdot1}X_2 + R^*_{Y|1,2}$$

   The statistics have asterisks (*) to signify that they come from modified (synthetic reponse variable) data. Save $b^*_{2\cdot1}$ and use it to calculate t^*, where $t^* = (b^*_{2\cdot1} - b_{2\cdot1})/\text{se}(b^*_{2\cdot1})$.

5. Repeat Steps (2)-(4) many times to obtain a distribution of t^* values from many permutations.

6. Calculate the p-value of t_{obs}. Reject H_0 if $p < 0.05$.

4.2.2 Considerations of ter Braak's Method

The impact of this permutation strategy on the relationships between the variables can be summarized as follows:

Broken Relationships	*Preserved Relationships*
None	X_1 & X_2
	X_1 & Y
	X_2 & Y

The permutation of residuals under the full model differs from the method described in Section 4.1, where permuted residuals came from a linear model fitted on the predictor variables *excluding* the one involved in the hypothesis test.

ter Braak (1992) initially introduced the bootstrap significance test before motivating its permutation analog, which is the algorithm described in this chapter. But what is bootstrapping?

Bootstrapping is a technique that involves sampling with replacement from the dataset, and it can be used to estimate the distribution of a statistic of interest (Efron and Tibshirani, 1994). For example, if we were interested in the true slope $\beta_{2\cdot1}$ in a linear model where Y is regressed on both X_1 and X_2, a possible bootstrap approach would be to sample n observations from the original dataset with replacement, calculate $b^*_{2\cdot1}$ (our bootstrap estimate of $\beta_{2\cdot1}$), repeat the former two steps for a total of 1,000 times to eventually generate 1,000 bootstrapped datasets and corresponding $b^*_{2\cdot1}$ estimates, and build a bootstrap sampling distribution using these estimates. The bootstrap approach is an effective way to estimate the variability associated with $b_{2\cdot1}$ without needing more observations, because the variability of $b^*_{2\cdot1}$ around $b_{2\cdot1}$ imitates the variability of $b_{2\cdot1}$ around $\beta_{2\cdot1}$.

The bootstrap model described by ter Braak (1992) involves calculating estimates of $\beta_{2\cdot1}$—denoted by $b^+_{2\cdot1}$—from synthetic datasets generated from bootstrapping the residuals of the full model, $Y \sim X_1 + X_2$. To highlight

the distinction, ter Braak's bootstrap test takes $b^+_{2\cdot1}$ values from synthetic datasets where the residuals have been resampled *with* replacement, whereas its permutation analog builds a distribution of $b^*_{2\cdot1}$ values from synthetic datasets where the residuals have been resampled *without* replacement. Both the bootstrap test and its permutation analog estimate the sampling distribution of t^* (i.e., the standardized sampling distribution of $b_{2\cdot1}$), which is used to estimate the true value of $\beta_{2\cdot1}$ in the following assumption:

$$F\left(\frac{b^*_{2\cdot1} - b_{2\cdot1}}{\text{se}(b^*_{2\cdot1})}\right) \approx F\left(\frac{b_{2\cdot1} - \beta_{2\cdot1}}{\text{se}(b_{2\cdot1})}\right)$$

The sampling distribution of t^* approximates the sampling distribution of t due to the theoretical underpinnings from bootstrapping. The variability of $b^*_{2\cdot1}$ around $b_{2\cdot1}$ mimics the variability of $b_{2\cdot1}$ around $\beta_{2\cdot1}$.

Under this assumption, we can use the t^* distribution constructed from many permutations of the same dataset to carry out a hypothesis test: if $H_0 : \beta_{2\cdot1} = 0$ is true, then $t_{\text{obs}} = (b_{2\cdot1} - 0)/\text{se}(b_{2\cdot1})$ would be a likely value in the t^* distribution, corresponding to an non-significant p-value that results in failing to reject H_0.

However, if $H_a : \beta_{2\cdot1} \neq 0$ is true and, say, $\beta_{2\cdot1} = 47$, then we would expect $\widehat{t} = (b_{2\cdot1} - 47)/\text{se}(b_{2\cdot1})$ to lie well within the t^* distribution, while $t_{\text{obs}} = (b_{2\cdot1} - 0)/\text{se}(b_{2\cdot1})$ would lie on the margins, leading to a small p-value that concludes the test by rejecting H_0.

Furthermore, ter Braak (1992) synthesizes equations from various authors—Efron (1982) for the bootstrap, along with Cox and Hinkley (1974) and Lehmann and D'Abrera (1975) for the permutation—to make the following statements regarding the expected value and variance of both types of estimated slope coefficients. Recall that $b^+_{2\cdot1}$ corresponds to the bootstrap, while $b^*_{2\cdot1}$ corresponds to the permutation of the residuals under the full model:

$$E^+(b^+_{2\cdot1}) = E^*(b^*_{2\cdot1}) = b_{2\cdot1} \tag{4.9}$$

$$var^+(b^+_{2\cdot1}) = (1 - 1/n)var^*(b^*_{2\cdot1}) \tag{4.10}$$

which means that the expected values of $b^+_{2\cdot1}$ and $b^*_{2\cdot1}$ are both $b_{2\cdot1}$, and that the variance of $b^+_{2\cdot1}$, the bootstrapped estimates, is smaller than the variance of $b^*_{2\cdot1}$. Hence, ter Braak uses Equations (4.9) and (4.10), along with the order property that $b^+_{2\cdot1}$ and $b^*_{2\cdot1}$ differ by $\text{O}(1/n)$ in second or higher order moments, to justify his proposal of the full model residual permutation strategy as an alternative for the bootstrap model (ter Braak, 1992, p. 6).

Permuting the residuals under the full model preserves any relationships among the predictor variables as well, which may not be the case as seen in Draper and Stoneman's method in Chapter 3. Anderson and Legendre (1999) also mention that ter Braak's procedure requires a pivotal statistic, such as the t-statistic.

Chapter 5

Simulations

We have discussed the following four permutation strategies:

1. Permutation of the response variable (Manly, 1986, 1997)
2. Permutation of the predictor variable of interest (Draper and Stoneman, 1966)
3. Permutation of the reduced model residuals (Freedman and Lane, 1983)
4. Permutation of the full model residuals (ter Braak, 1990, 1992)

We will assess these four methods of permutation on simulated datasets in R to determine the most appropriate permutation method under conditions of known changes to particular factors. Each permutation method is evaluated based on their Type I error and power.

Previous literature on permutation tests for MLR models found similar results with differences across methods only when simulating outliers and non-normal error structures (Anderson and Legendre, 1999). Following the structure of the simulations from their study, we borrow similar values of $\beta_{1\cdot 2}$ and $\beta_{2\cdot 1}$ in our setup.

5.1 Methodology

To evaluate how well the four permutation methods performed under different settings, we varied the sample size (n), the size of the covariable's parameter ($\beta_{1\cdot 2}$), the size of the coefficient of interest ($\beta_{2\cdot 1}$), and the correlation

Simulation Values	
n	20, 40, 80
$\beta_{1\cdot 2}$	0.5, 1.0, 2.5
$\beta_{2\cdot 1}$	0.00, 0.30, 0.40, 0.75
$\rho(X_1, X_2)$	0.0, 0.5, 0.9

Table 5.1: Different iterations of values used to generate the datasets for determining the Type I error and power of each of the four permutation methods. Every possible combination of these values was used, resulting in 108 different parameter configurations for the simulations.

between X_1 and X_2 ($\rho(X_1, X_2)$), as seen in in Table 5.1. All combinations of these factors were used in generating datasets. Simulations were run for all combinations of these factors, summing up to 108 different parameter configurations. For each configuration, 1,000 datasets were generated using the `mvnorm()` function from the MASS library to generate X_1 and X_2, from which Y was calculated as a linear combination of X_1 and X_2 plus Gaussian noise. There were 1,000 permutations done according to each permutation method per dataset. In the end, there were four p-values—one from each permutation method—associated with every dataset.

After categorizing datasets by their parameter configuration, Type I error was calculated by computing how often the test gave a p-value less than 0.05 when $\beta_{2\cdot 1} = 0$. Power was calculated in the same way for datasets generated from nonzero $\beta_{2\cdot 1}$ values.

5.2 Results

Type I error and power across datasets generated under the 108 parameter configurations are summarized in the following multi-panel plots, which are grouped by $\beta_{2\cdot 1}$ values, with the rejection rate plotted along the y-axis and $\beta_{1\cdot 2}$ values plotted along the x-axis.

5.2.1 Examination of Type I Error

Figure 5.1 displays the Type I error of different data configurations (when $\beta_{2\cdot1} = 0$). The four methods generally appear to have a Type I error rate around $\alpha = 0.05$, though the plot with $\rho(X_1, X_2) = 0.0$ and $n = 80$ and the plot with $\rho(X_1, X_2) = 0.9$ and $n = 40$ suggest inflated Type I error rates. Most plots show similar trends across methods, but there seems to be more variation in the plot with $\rho(X_1, X_2) = 0.9$ and $n = 40$: Freedman and Lane's method goes up at $\beta_{1\cdot2} = 1.0$ and goes down at $\beta_{1\cdot2} = 2.5$; Draper and Stoneman's method first dips and stays the same; ter Braak's method steadily increases; and Manly's method steadily decreases. There was no consistent trend in Type I error as $\rho(X_1, X_2)$ or n increased.

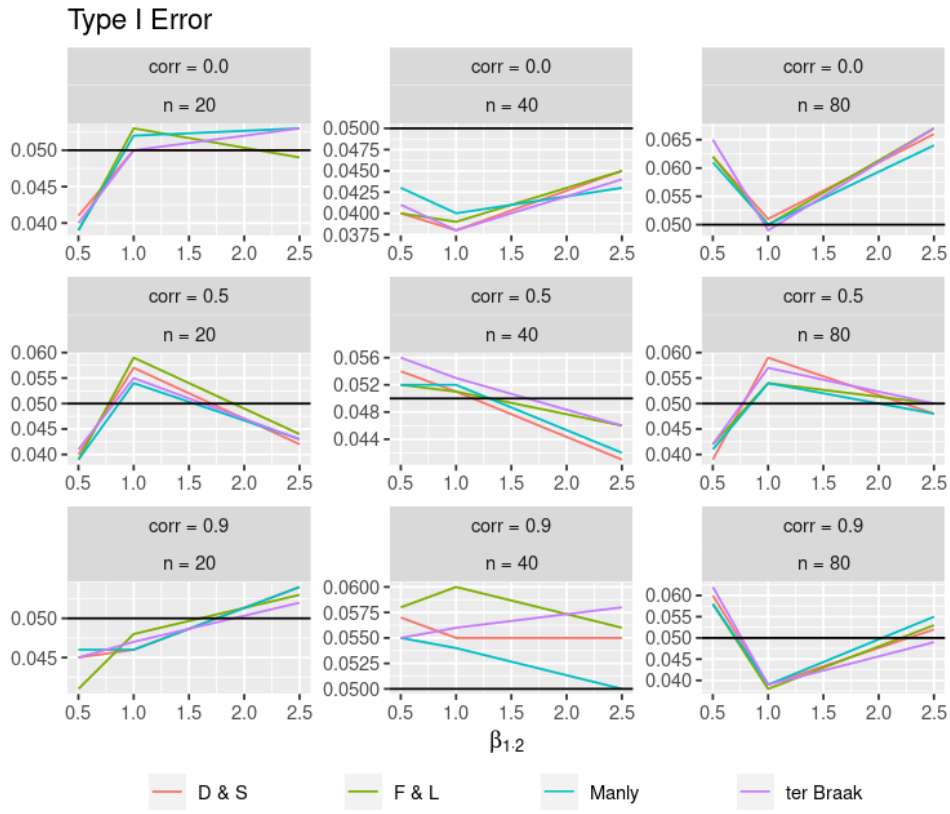


Figure 5.1: Type I error (when $\beta_{2\cdot1} = 0$) is plotted along the y-axis, while $\beta_{1\cdot2}$ is plotted along the x-axis. Different methods are denoted by their colors. A horizontal line has been graphed to indicate the significance level, $\alpha = 0.05$.

5.2.2 Examination of Power

Figure 5.2 displays the power of different data configurations with $\beta_{2\cdot1} = 0.3$. All four methods display a decrease in power with increasing $\rho(X_1, X_2)$ and an increase in power with increasing n. There appears to be more variation across the methods as $\beta_{1\cdot2}$ increases for the plots with $n = 20, 40$ aside for the plot with $\rho(X_1, X_2) = 0.9$ and $n = 40$.

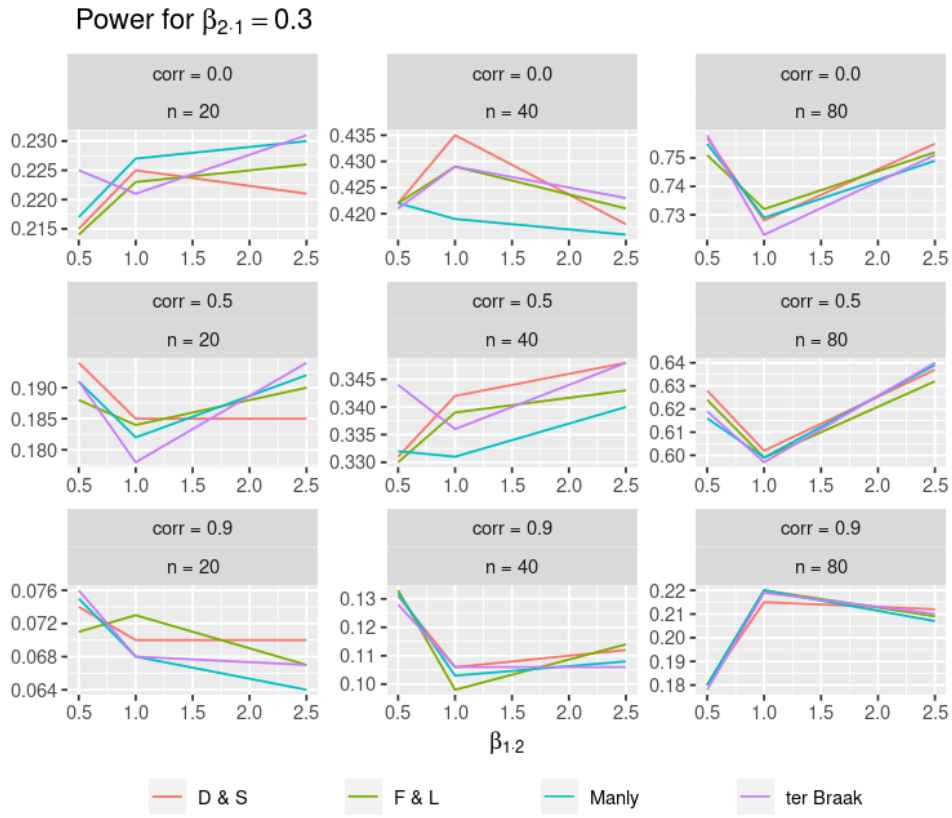


Figure 5.2: Power when $\beta_{2\cdot1} = 0.3$ is plotted along the y-axis, while $\beta_{1\cdot2}$ is plotted along the x-axis. Different methods are denoted by different colors.

Figure 5.3 displays the power of different data configurations with $\beta_{2\cdot1} = 0.4$. All plots look similar to those displayed in Figure 5.2 except for the plot with $\rho(X_1, X_2) = 0.0$ and $n = 80$ and the plot with $\rho(X_1, X_2) = 0.9$ and $n = 20$, both of which seem to have opposite patterns. There are fewer plots that show considerable deviations in patterns across the methods, most notably in the plot with $\rho(X_1, X_2) = 0.5$ and $n = 40$. Not surprisingly, power when $\beta_{2\cdot1} = 0.4$ is higher in all situations than when $\beta_{2\cdot1} = 0.3$. We observe some similar trends as those found in Figure 5.2, where power increases with bigger n and smaller $\rho(X_1, X_2)$.

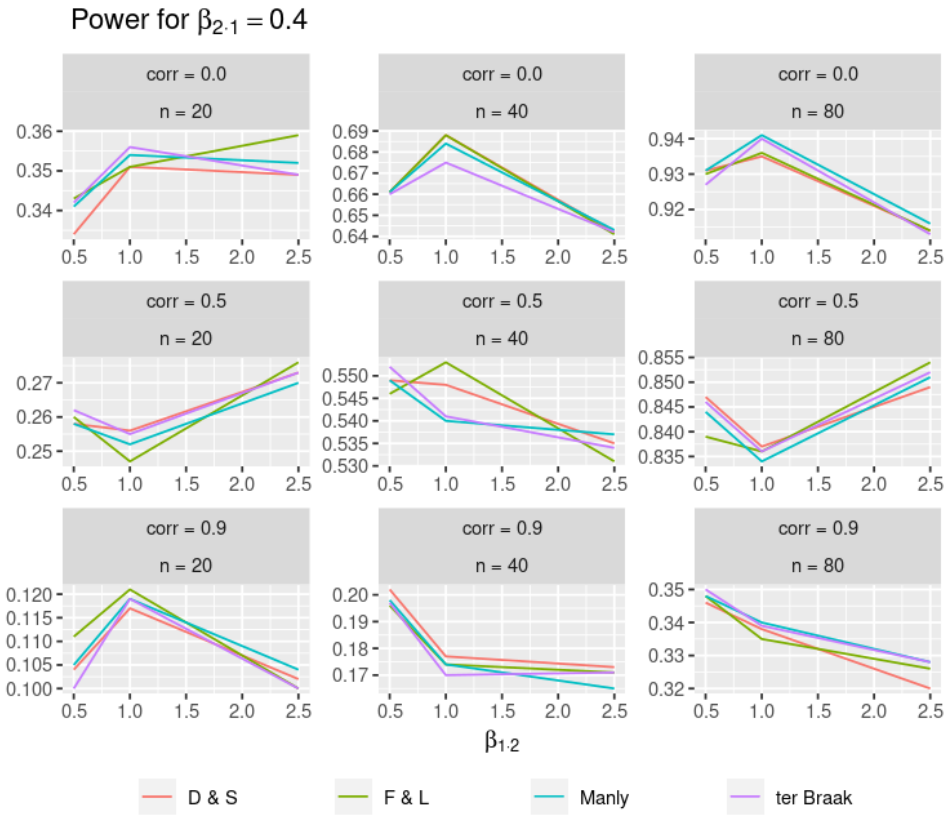


Figure 5.3: Power when $\beta_{2\cdot1} = 0.4$ is plotted along the y-axis, while $\beta_{1\cdot2}$ is plotted along the x-axis. Different methods are denoted by different colors.

Figure 5.4 displays the power of different data configurations with $\beta_{2\cdot1} = 0.75$. All plots seem to follow the same trends; the only line that just barely stands out is ter Braak's method in the plot with $\rho(X_1, X_2) = 0.0$ and $n = 20$, though it is also decreasing with increasing $\beta_{1\cdot2}$ like the other three methods. Not surprisingly, power when $\beta_{2\cdot1} = 0.75$ is higher in all situations than when $\beta_{2\cdot1} = 0.3, 0.4$. We observe some similar trends as those found in Figures 5.2 and 5.3, where power increases with bigger n and smaller $\rho(X_1, X_2)$. In fact, power maxes out at 1.0 for all methods when $n = 80$ and $\rho(X_1, X_2) = 0.0, 0.5$.

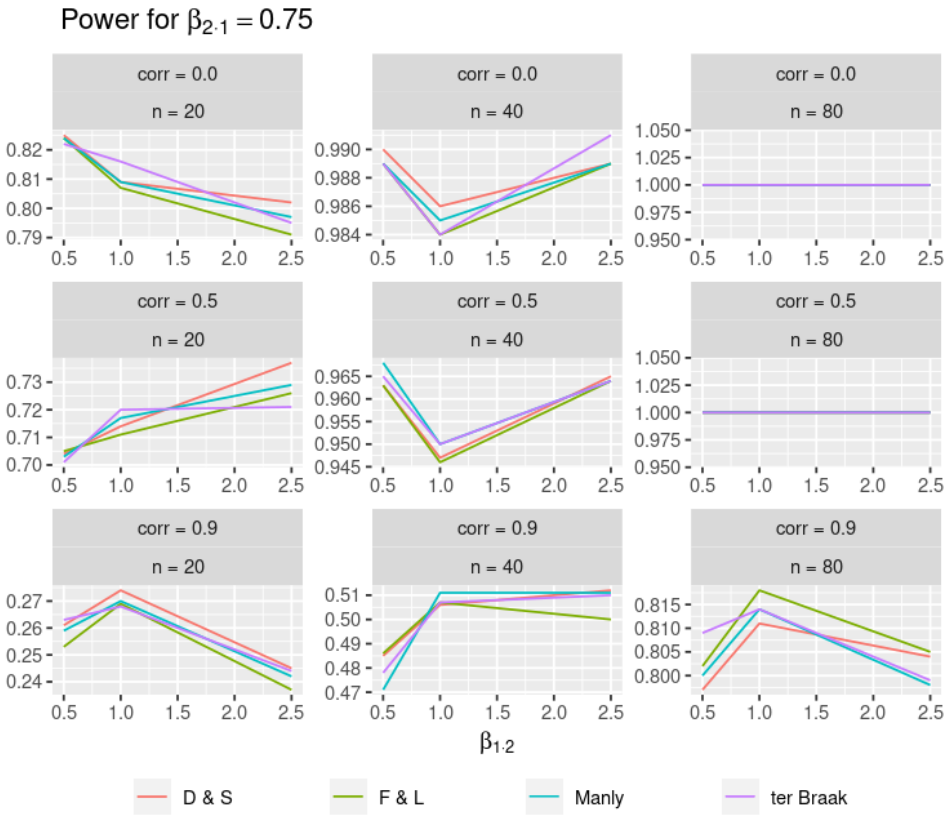


Figure 5.4: Power when $\beta_{2\cdot1} = 0.75$ is plotted along the y-axis, while $\beta_{1\cdot2}$ is plotted along the x-axis. Different methods are denoted by different colors.

Chapter 6

Discussion

Based on the results of the simulation study in Chapter 5, there are no notable differences in Type I error or in power across the four permutation methods. These results were a bit unexpected for the following considerations:

1. Manly's method of permuting the response variable would break the relationship between X_1 and Y, as mentioned in Section 2.2. This was expected to hinder an accurate isolation of the additional effect of adding X_2 to a model that already accounts for X_1.

2. Draper and Stoneman's method of permuting the predictor variable of interest would break the relationship between X_1 and X_2, as mentioned in Section 3.2. This seemed to be problematic in circumstances where X_1 and X_2 are indeed related in the data.

3. Freedman and Lane's method of permuting the reduced model residuals would fail to force H_0 to be true if $\rho(X_1, X_2)$ was too high, as mentioned in Section 4.1.2.

Thus, it is a little surprising that despite all these aforementioned caveats, the four permutation methods performed rather similarly across all 108 combinations of different factors. Our results are consistent with a previous study done by Anderson and Legendre (1999), who did not find any significant differences on datasets without outliers and with normal error structures. (Both of these conditions also describe the datasets used in our simulation.)

It is possible that the reason why all four methods were so correlated is in part due to only generating 1,000 datasets—as opposed to 10,000 datasets in

Anderson and Legendre (1999)—for each combination. It would be worthwhile to perform more analysis on the variability among different datasets (that were unfortunately not kept after the simulations were finished), which may be driving the observed trends in power moreso than increasing values of $\beta_{1\cdot 2}$ are. Thus, more simulations and analysis should be done before believing in the spikey patterns shown in Figures 5.1 to 5.4. Though these patterns do not align much with the corresponding plots in Anderson and Legendre (1999), the general range of Type I error rates and power are comparable under similar situations (roughly within ± 0.01 bounds).

Datasets with interaction between the predictor variables, extreme outliers, and non-normal error structures were not simulated in this paper due to time constraints, but it would be a viable future direction that would be helpful in expanding upon this study. Future studies in this field should also employ even more generated datasets (e.g., at least 10,000) for each parameter configuration. In addition, further analysis should be carried out to assess the variability among datasets to see what is actually driving patterns in Type I error and power.

Acknowledgements

I'm extremely grateful to Professor Hardin, my thesis and primary advisor, who patiently guided me through the process of creating this paper every week—you've always inspired me to keep asking questions and push beyond my limits. I'd also like to give a big shoutout to Lauren Quesada (2022) for all the work she had covered on permutation tests in her own undergraduate thesis; it really helped us get the ball rolling before I even had my first draft done. Special thanks to the rest of Mathematics and Statistics Department at Pomona College for not only enriching my education for the last four years but also fostering such a wonderful community of peers without whom I wouldn't have made it this far. Lastly, I'd like to extend my deepest gratitude to my family for their endless love and support.

Bibliography

Anderson, M. J. and 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(3):271–303.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman & Hall.

Draper, N. R. and Stoneman, D. M. (1966). Testing for the inclusion of variables in linear regression by a randomisation technique. *Technometrics*, 8(4):695–699.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman & Hall.

Freedman, D. and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1(4):292–298.

Kutner, M., Nachtsheim, C., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill, 5th edition.

Lehmann, E. L. and D'Abrera, H. J. M. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

Manly, B. F. J. (1986). Randomization and regression methods for testing for associations with geographical, environmental and biological distances between populations. *Population Ecology*, 28(2):201–218.

Manly, B. F. J. (1997). *Randomization, bootstrap and Monte Carlo methods in biology*. Texts in Statistical Science Series. Chapman & Hall, London, 2nd edition.

ter Braak, C. J. F. (1990). Update notes: Canoco, version 3.10. *Wageningen, The Netherlands: Agricultural Mathematics Group*, page 35.

ter Braak, C. J. F. (1992). Permutation Versus Bootstrap Significance Tests in Multiple Regression and Anova. *Austr. J. Statist.*, 29:79–85.