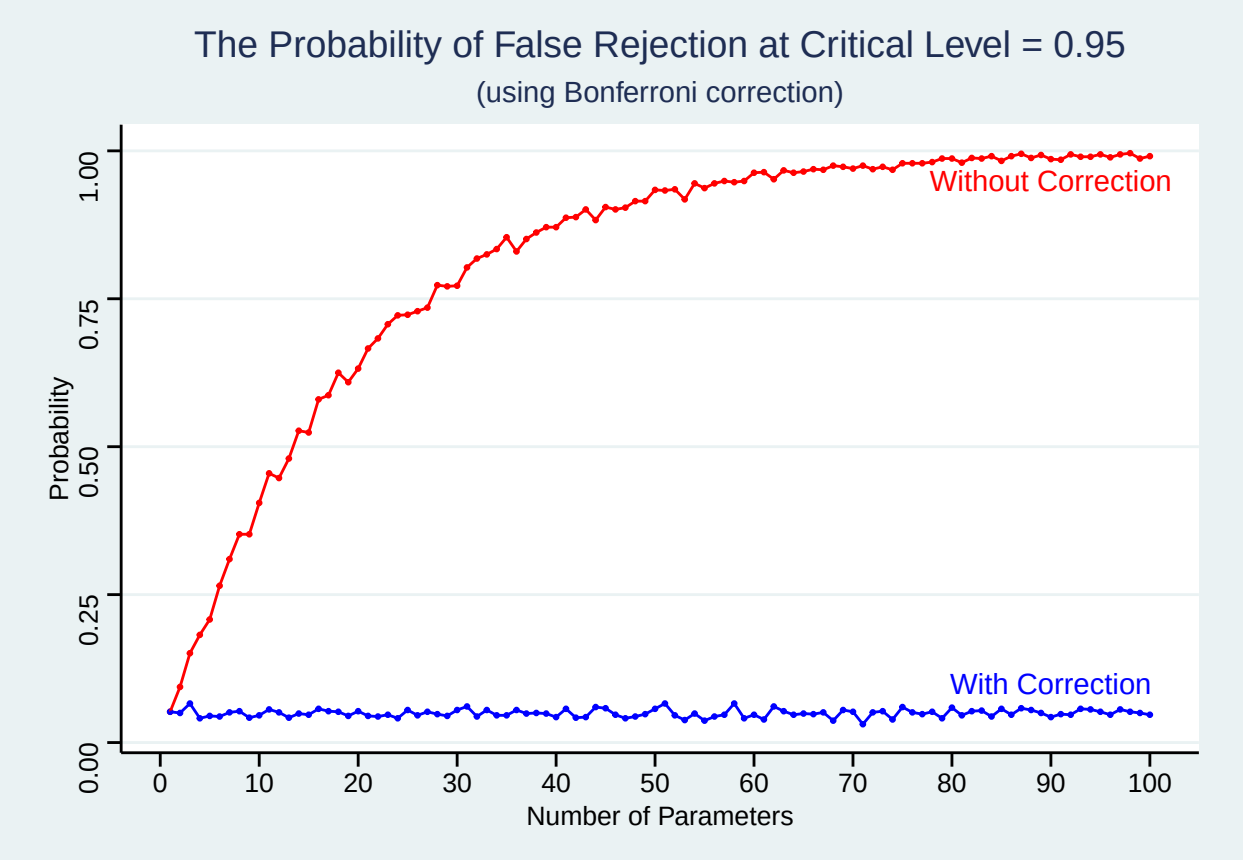
Boseong Yun - PSET 1

Boseong Yun

1/25/2021

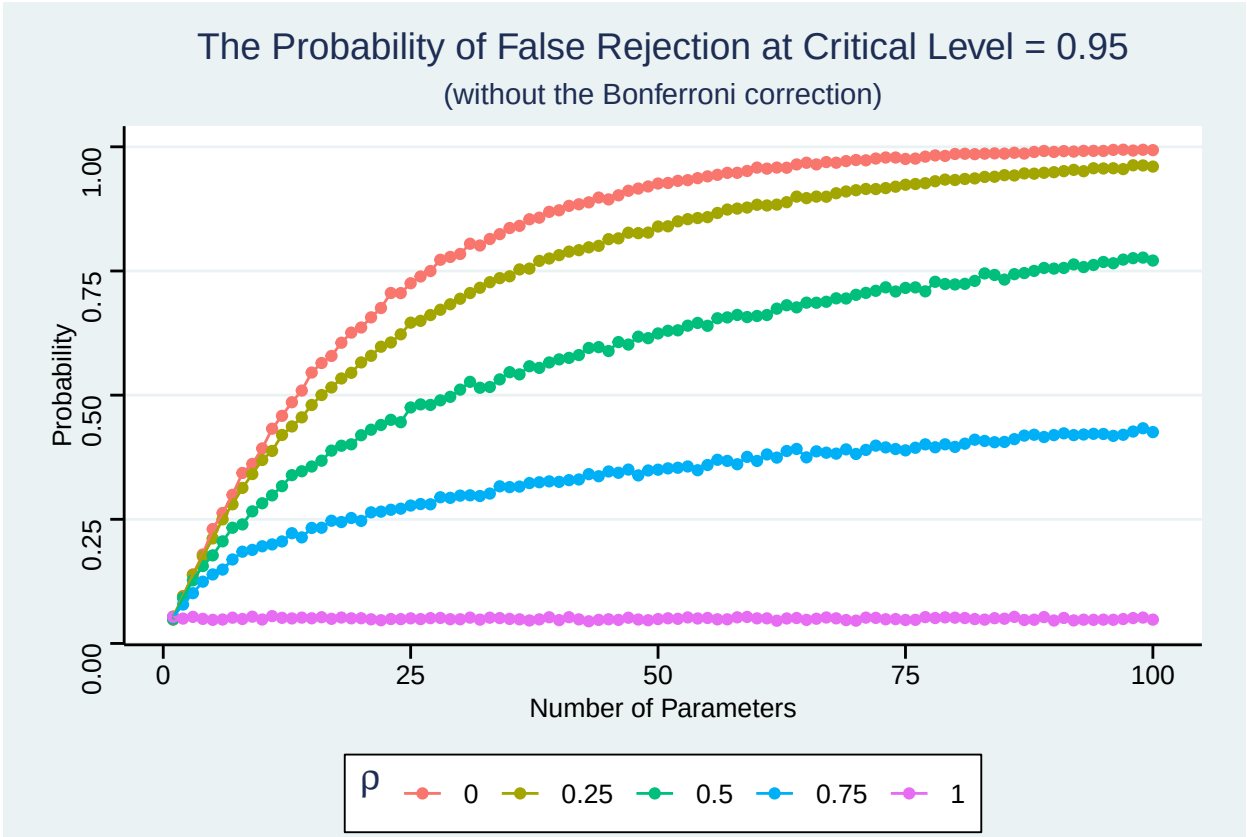
1.a. Generate the following figure: for test statistics with null distribution (2), plot the probability of false rejection of the joint null (that the p individual null distributions are the true distributions) at critical level $alpha = 0.95$ against the value of p, using the Bonferroni correction.

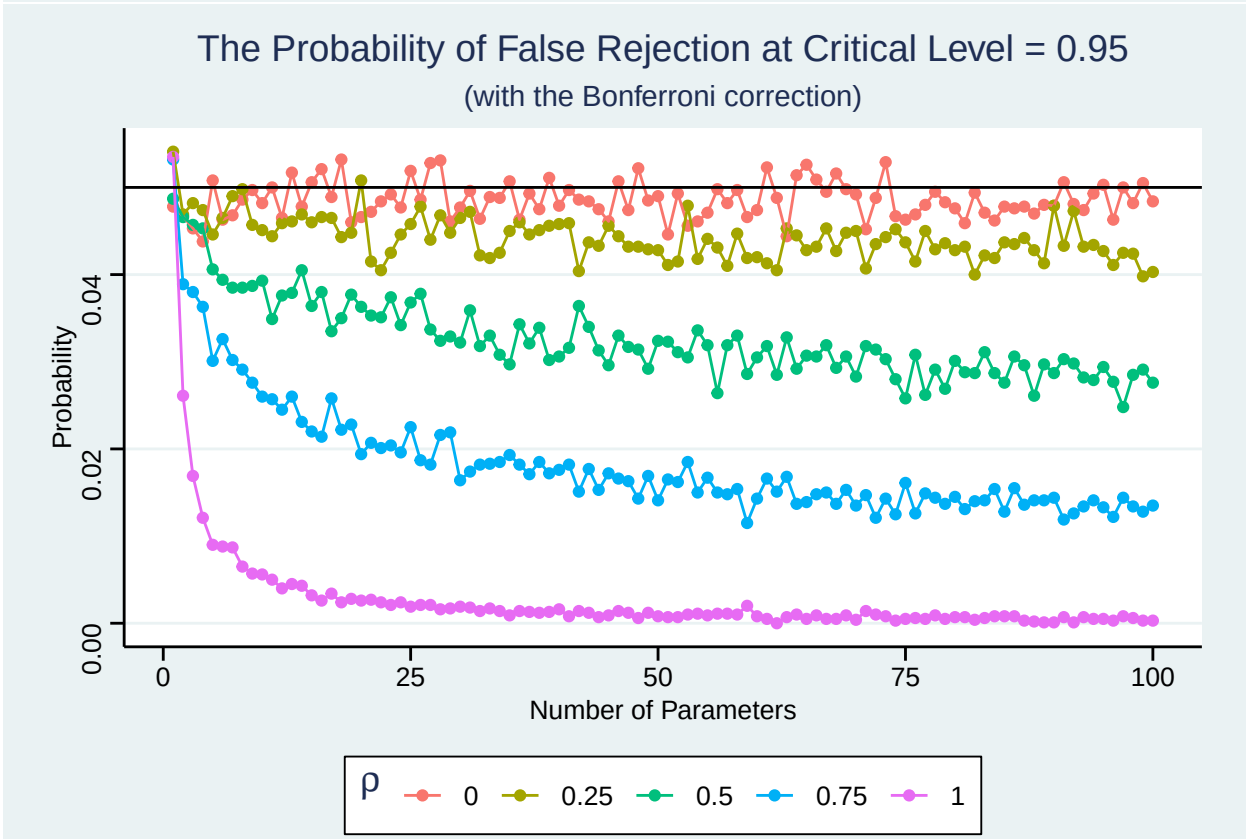
Answer: I have created test statistics that follow the null distribution given in (2). Afterwards, I plotted the probability of false rejection at critical level 0.95 against the value of p using the Bonferroni correction.



1.b **Answer: I have created two plots where the probability of false rejection at critical level 0.95 is uncorrected and corrected. I have done it by creating a nested for loop where for every value of ρ in 0, 0.25, 0.5, 0.75, and 1, I have plotted the probabilty of rejection against the number of parameters. I have used RowSums function to find out the number of rejections in each test and divided the sum by the number of simluations.**

(Thank you so much to all TAs who made this possible!)


The Probability of False Rejection at Critical Level = 0.95
(without the Bonferroni correction)
Probability
0.00
0.25
0.50
0.75
1.00
0
25
50
75
100
Number of Parameters
ρ
0
0.25
0.5
0.75
1


The Probability of False Rejection at Critical Level = 0.95
(with the Bonferroni correction)
Probability
0.00
0.02
0.04
0
25
50
75
100
Number of Parameters
ρ
0
0.25
0.5
0.75
1

1.c. The biggest value of ρ that one can use such that sigma remains a covariance matrix?

Answer: The biggest value of ρ that one can use such that sigma remains a covariance matrix is any positive value very close to 1 but not actually 1. If ρ takes the value of 1, the covariance matrix will no longer be a positive-definite matrix because one of the eigenvalues will be 0. Also, it means that the model will suffer from perfect multicolinearity.

2-a) Download the data set Hitters from the ISLR library –the R CRAN library complementing the course's textbook.

2-b) Divide the dataset into a training set of ntrain observations and a test set of ntest observations.

Answer: There is a tradeoff between having a big training & small test versus small training & big test. For instance, the parameter estimates will have higher variance with less training data and the performance statistics will have higher variance with less testing data. Although it depends on the model, more observations in the training data set causes less bias but high variance (overfitting) and more observations in the test data set causes high bias but less variance (underfitting). While there is no absolute method for determining the ideal ratio since the split is dependent on the given data, it is useful to think about the bias-variance tradeoff when making the decision.

According to the Pareto principle, 80/20 is the preferred ratio for dividing the training data and test data. Some people also use the 75/25 ratio. These ratios are reasonable because these ratios ensure that both the bias and variance are small. The optimal ratio must be calculated in consideration of the dataset given at hand.

2-c) Using the training data, select the model made of the 7 coefficients with the smallest p-values according to the regression fit of the full model.

Table 1: The Best Model Using the Smallest P-Value of the Full Model

term	estimate
Walks	5.5826781
PutOuts	0.2253253
DivisionW	-115.6191801
AtBat	-1.6894684
CRuns	1.7459174
Assists	0.4999659
CWalks	-0.7351627
(Intercept)	193.1511514

2-d) d. Using the training data, select the best model made of 7 coefficients according to the forward stepwise selection procedure (p. 247 for hints). You do not need to code the procedure yourself –use an R package!– but explain what this procedure does, and how it is different from the one in c.

Answer: (Source:Pg. 78-79 ISLR) The forward stepwise selection procedure starts with the null model that contains only the intercept. It then adds the coefficients that results in the lowest RSS until some stopping conditions. It has the risks of including variables in the early stage that later become redundant. This is different from c where the proceudre starts from the full model and remove variables that are the least significant.

Table 2: The Best Model Using the Forward Stepwise Selection Procedure

	Estimates
(Intercept)	159.5817121
AtBat	-0.8614299
Hits	4.4645620
CAtBat	-0.3883580
CHits	1.0803774
CRuns	1.3106797
DivisionW	-104.5143395
PutOuts	0.2503880

2-e) Using training data, select the best model made of 7 coefficients according to the best subset procedure (p. 244 for hints).

Table 3: The Best Model Using the Best Subset Procedure

	Estimates
(Intercept)	128.6685947
Walks	4.9184367
CAtBat	-0.4322119
CHits	1.4517383
CRuns	1.1280754
CWalks	-0.4040476
DivisionW	-121.5552554
PutOuts	0.2138940

2-f. Compute the sample mean squared error in the test set for each method fitted in c, d and e, and collect the results in a table.

Answer: In this given dataset, the sample mean squared error in the test for method c (selecting the smallest p-value) is the smallest. It is also important to notice that the difference in the sample mean squared error in the test for method d (forward selection) and method e (best subset) is relatively negligible. I present the change in the sample mean squared error in the test against the number of coefficients in 2-g to further investigate the differences.

Table 4: The Sample Mean Squared Error in the Test set for Each Method

Fitted.Models	Sample.Mean.Squared.Errors
c	118305.5
d	134865.3
e	136358.9

2-g. Repeat exercises c-f for different sizes of the subset of coefficients, and present your results in an extended table or plot.

Answer: The following table shows that there is some variability in performance when the number of coefficient is low. However, the differences in the sample mean squared errors become miniscule as the size of coefficients grows

Table 5: The Sample Mean Squared Errors for model c, d, e

Num of Parameters	Model C	Model D	Model E
1	159469.9	159469.9	159469.9
2	136483.5	133003.6	133003.6
3	137620.5	128183.3	154461.5
4	133740.7	112161.4	147843.6
5	117698.6	110795.6	136043.5
6	118243.2	116714.7	142235.5
7	118305.5	134865.3	136358.9
8	118305.5	128824.5	128824.5
9	118305.5	123836.0	123836.0
10	118305.5	108025.3	108025.3

2-h. bonus question: For selecting larger subsets with the best subset selection method, compare the performance of the package leaps with that of bestsubset.2 Consider adding interactions.

Answer: I have tried to downnload the package leaps using the link at the bottom of the homework document. Unfortunately, some of the dependent packages were not compatible with my latest version of R and hence I was not able to use them.

2-i. Can you suggest a more efficient way to split and use the data as training and testing sets?

Answer: Yes, we can use a K-fold cross-validation approach where we can divide the training and validation data sets into K subsets where we treat the kth subset and the rest subsets as validation and traning sets for every k to K. According to ISLR, using k = 5 or k = 10 have shown empirically to yield test error rates that suffer neither from excessivel high bias nor from very high variance (page. 184)

3-a. **Answer: The glm function does not return R^2. Since the lm function returns R^2 with the same coefficients, I use the information obtained from the lm function to answer this question. I also provide an additinal answer where I compute using the $Psuedo - R^2$**

```
## The in-sample sum of squared errors is 726316624
```

```
## The R-Squared is 0.734876
```

```
## The Adjusted R-Squred is 0.726251
```

```
## The Pseudo R-Squared is 0.734876
```

3-1b. The model *daylm* has been fitted using the Ordinary Least Squares Method where the method minimizes the sum of squared residuals. The model allows us to find the average change in total sales for year, month, and the interaction between month and year associated with each day (observation i refers to each day). The mathematical formula for daylm is defined as:

$$Total_i = \beta_0 + \beta_1 Year_i + \beta_2 Month_j + \beta_3 Year_i Month_j + \epsilon_i$$

and

$$E(Total|Year, Month) = \beta_0 + \beta_1 Year_i + \beta_2 Month_j + \beta_3 Year_i Month_j$$

where

$$Y \sim N(X\beta, \sigma^2 I_n)$$

and

$$f_y(y; \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{2}{n}}} exp[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)]$$

The implied probability model forms the basis of inference. I provide further information about the model by interpreting the regression outputs produced below.

Table 6: The Regression Outputs for the model daylm

Term	Estimate	Std.error	Statistic	P.value
(Intercept)	1231.90323	182.0423	6.76712	0.00000
yr	1888.87097	257.4468	7.33694	0.00000
mnth2	490.06106	264.2527	1.85452	0.06408
mnth3	834.06452	257.4468	3.23976	0.00125
mnth4	1930.43011	259.5833	7.43665	0.00000
mnth5	3149.41935	257.4468	12.23328	0.00000
mnth6	3551.83011	259.5833	13.68282	0.00000
mnth7	3327.48387	257.4468	12.92494	0.00000
mnth8	3177.48387	257.4468	12.34229	0.00000
mnth9	3015.36344	259.5833	11.61617	0.00000
mnth10	2752.32258	257.4468	10.69084	0.00000
mnth11	2173.66344	259.5833	8.37366	0.00000
mnth12	1584.96774	257.4468	6.15649	0.00000
yr:mnth2	-54.38698	372.0132	-0.14620	0.88381
yr:mnth3	1363.70968	364.0847	3.74558	0.00019
yr:mnth4	756.26237	367.1062	2.06006	0.03976
yr:mnth5	48.03226	364.0847	0.13193	0.89508
yr:mnth6	88.39570	367.1062	0.24079	0.80979
yr:mnth7	119.70968	364.0847	0.32880	0.74241
yr:mnth8	621.19355	364.0847	1.70618	0.08841
yr:mnth9	1149.62903	367.1062	3.13160	0.00181
yr:mnth10	541.12903	364.0847	1.48627	0.13765
yr:mnth11	-205.63763	367.1062	-0.56016	0.57555
yr:mnth12	-715.00000	364.0847	-1.96383	0.04994

The interpretation shows that the total sales are going to increase by **1888.87 + 490.0610 + (-54.38698)** dollars on average in Februrary, 2012 compared to January 2011 (the base timeline). In 2011 Februray, the total sales are going to rise by **490.06** dollars on average relative to January 2011. It is important to notice that the synergy effect between year and month must be carefully read in order to correctly find out the impact of year, month, and their synergy effects on total sales in this model.

Although it requires more information about the purpose and assumptions behind the model to evaluate its strength and weaknesses, the model *daylm* does not seem to maintain its model assumptions. Primarily, the model assumes that the variance is a constant. However, it is reasonable to suspect hetereoskedasticity because bike sales are more likely to be sold during June or July than December or January. This can break the constant variance assumption and weaken or invalidate our inferences.

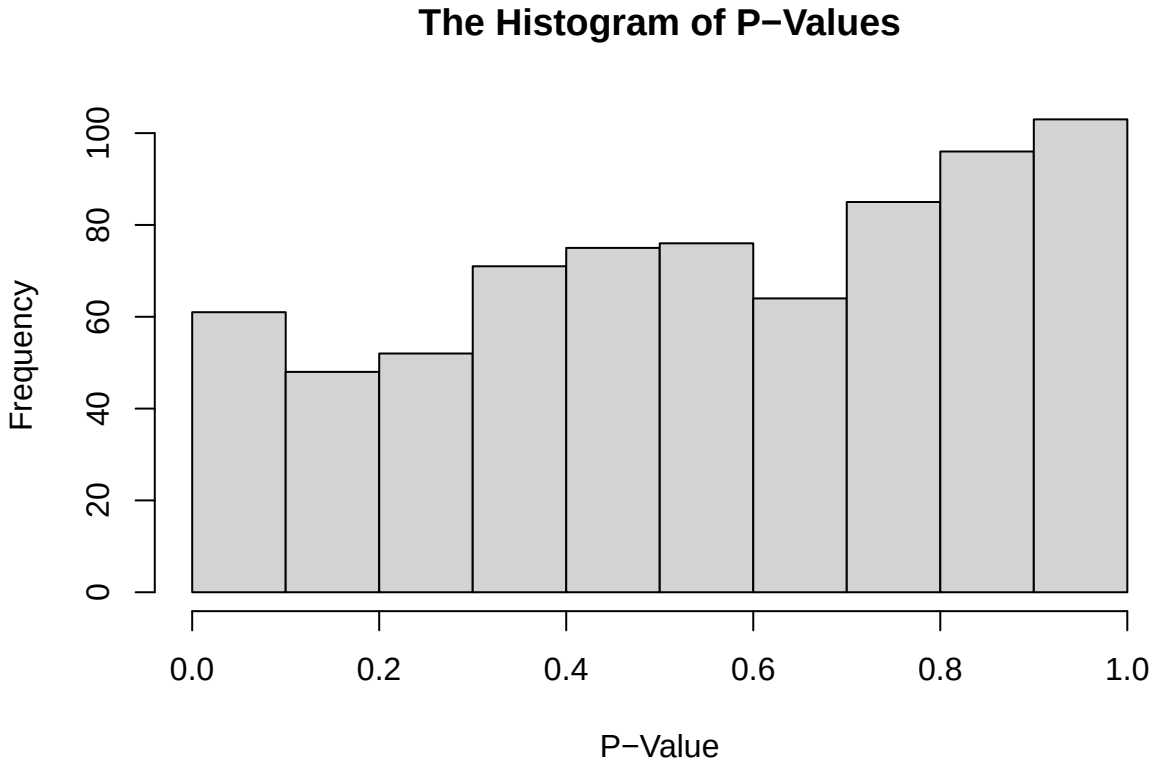
Additionally, I believe that the model should have more explanatory variables in the model. Specifically, there are many more important factors other than year and month that describe the total sales of bike. For instance, it is important to have information about the costs of public transportation, the costs of parking

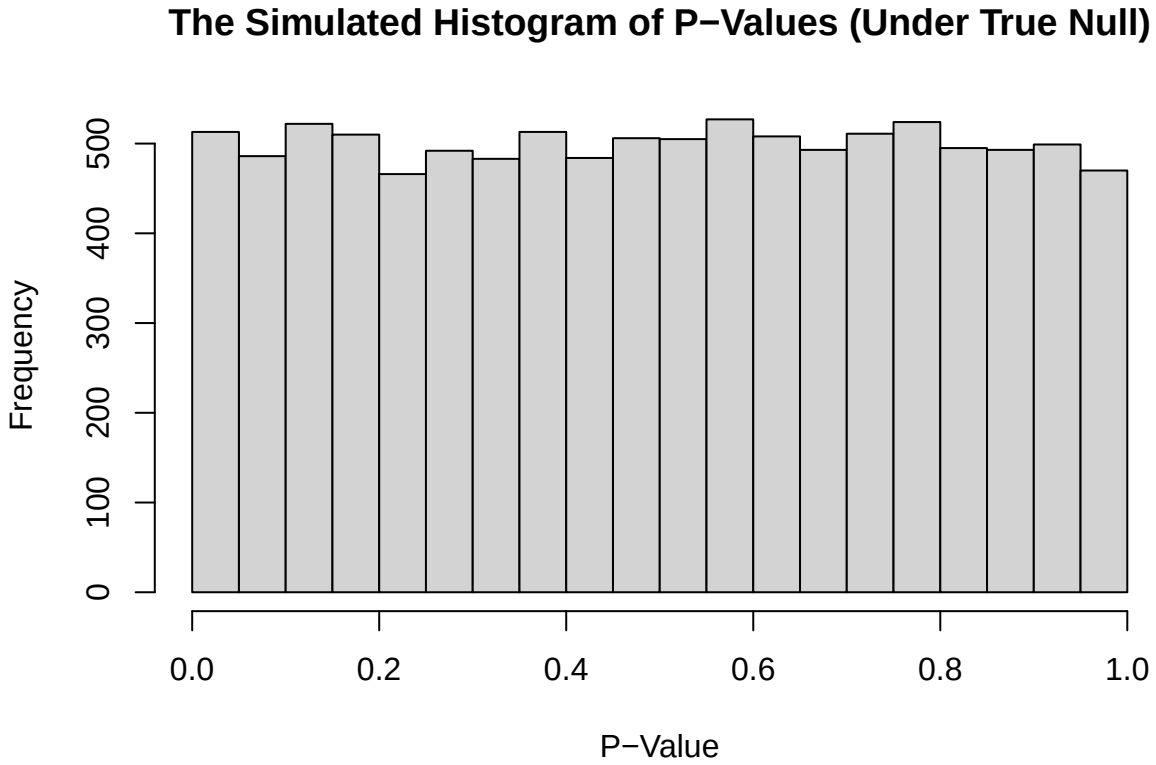
space, traffic load, GDP, and many others to find out what drives the bike sales. This hints at potential omitted variable bias problems where the omitted variables are corrleated with independent and dependent variables. On a policy side, this would allow those who are in the bike business to develop specific business strategies to boost bike sales. Thus, I think the model should have more explanatory variables in the model.

3-1c. **Answer: The p-values are generated and calculated from a normal distribution where Z ~ N(0, 1). That is, the p-values correspond to a set of p-values for standard normal random variables. The distributiuon of p-value is going to be uniform when the null hypothesis is true. Also, we expect the distribution to be skewed towards 0 (small) when the null hypothesis is not true and thus has some predictive signal. That is, small p-values indicate a possible outlier day because the probability of happening is extreme. Further analysis will be revealed in the next question.**

3-d. Plot the p-value distribution. What does it tell you about the assumptions of the probability model we used for our regression? Discuss.

Answer: The histogram shows that it is not uniformly distributed. Specifically, the histogram is skewed to the right. This suggests that some of the assumptions of the probablity model we used for our regression are not met. This issue can be problematic because it can invalidate our inferences. As the simulated histogram shows, the p-values has to be uniformly distributed under true null hypothesis. This suspcious result links back to the potential criticism we made in 3b about how the assumptions of the probability model could be broken.





3-e. Consider the drawing on the second to last slide of the deck of Lecture 2. Produce the equivalent drawing to illustrate the omitted variable bias phenomenon.

Answer: Professor Pouliot demonstrated this problem in his third lecture. As the drawings show, the estimates and error are well defined for the long regression on the column sapce (plane) spanned by both X1 and X2. On the other hand, the estimates and errors for the short regression are poorly defined because the column space is only a line. This illustrates omitted variable bias where the short regression fails to fully capture the extent of estimates and errors due to low dimensionality