ST 495 Final Project

Bryant Willoughby

April 29, 2024

Department of Statistics

North Carolina State University

1 Introduction

This project is based on a dataset that originates from the StatLib library at Carnegie Mellon University about city-cycle field consumption. I developed a model fitting procedure to predict city-cycle fuel consumption based on provided population features. To assess the performance of this estimation procedure, I constructed a simulation study. This involves simulating the estimation procedure numerous times from synthetically generated data. Then, I reconstructed the simulation study with a bootstrapping approach. This involves subsampling from the observed data numerous times before conducting the same estimation procedure.

2 Data Description

This analysis uses a dataset on city-cycle fuel consumption, available at `https://archive.ics.uci.edu/dataset/9/auto+mpg`. The modified dataset, excluding eight records with unknown values, is selected. Originating from the StatLib library at Carnegie Mellon University, the dataset aims to predict city-cycle fuel consumption (mpg) based on population features. Table 1 details the variable types in the dataset.

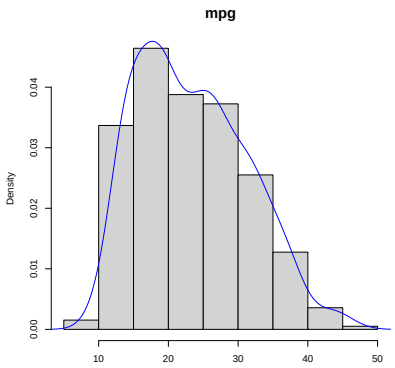
Variable	**Data Type**
mpg, displacement, horsepower, weight, acceleration	continuous
cylinders, modelyear, origin	multi-valued discrete
car name	string (unique for each instance)

Table 1: Variables in the Dataset

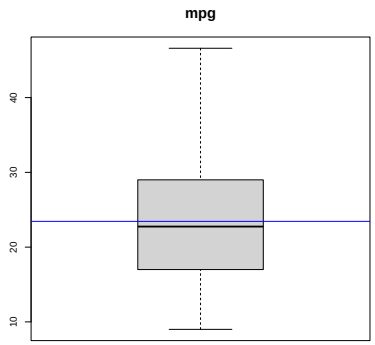
2.1 Data Cleaning and Exploratory Analysis

The dataset contains 398 records and nine variables. In cleaning the data, variables were renamed, and a thorough check for missing data was conducted. Six records with missing 'horsepower' values were removed. Due to the limited number of missing values relative to the dataset size, the decision was made to remove these records to avoid imputation techniques. Exploratory plots and

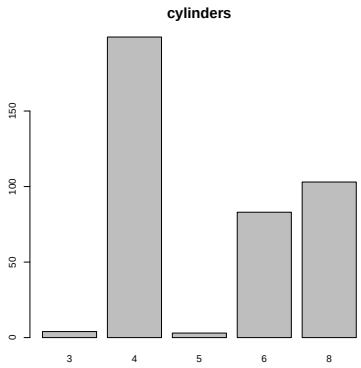
summary statistics are found in Figure (1). A histogram with an overlaid kernel density line shows the unimodal and slightly right-skewed nature of the 'mpg' variable. Boxplots provided another visual assessment of continuous variables, and contingency tables and bar plots offered insights into multi-valued discrete variables.



(a) Histogram for `mpg`



(b) Boxplot for `mpg`



(c) Barplot for `cylinders`

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
`mpg`	9.00	17.00	22.75	23.45	29.00	46.60
`displacement`	68.0	105.0	151.0	194.4	275.8	455.0
`horsepower`	46.0	75.0	93.5	104.5	126.0	230.0
`weight`	1613	2225	2804	2978	3615	5140
`acceleration`	8.00	13.78	15.50	15.54	17.02	24.80

(d) Numeric Summaries

`Cylinders`	Count
3	4
4	199
5	3
6	83
8	103

(e) Contingency Table for `cylinders`

Figure 1: Graphical and Numeric Summaries

3 Methods

In this section, the linear regression approach employed for predicting the `mpg` response variable is discussed. Sections 3.1 and 3.2 describe the more mathematically and statistically rigorous framework for linear regression, respectively. Section 3.3 details the variables of interest and their relationships being studied. Following this consideration, the model fitting process (Section 3.4) is explained. Section 3.5 describes diagnostic measures and a subsequent transformation strategy used to improve the overall model fit. The discussion of the fitted model on the real data, broader implications, and possible limitations of the procedure are found in Section 3.6.

3.1 Linear Regression: Mathematical Considerations

For linear regression, the linear system can be expressed as:

$$Y = X\beta, \quad \text{where } X \in \mathbb{R}^{n\times p}, Y \in \mathbb{R}^n, \beta \in \mathbb{R}^p.$$

The dimensions of X lead to different settings for study. We assume the *classical setting*. More formally, this implies that $n > p$ and X is of full rank (rank(X) = p). This means that all columns are linearly independent. In this setting, a unique solution can exist when $Y \in \text{col}(X)$, but it is more likely that $Y \notin \text{col}(X)$. Recall that the data contains $n = 392$ rows and $p = 9$ columns, placing the analysis in the classical setting. Assuming $Y \notin \text{col}(X)$, the goal is to find the least squares solution:

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \|Y - X\beta\|_2^2.$$

Denoting $Q(\beta) = \|Y - X\beta\|_2^2$, two conditions for optimality are required: (1) existence of derivatives and (2) convexity. Under these conditions, the least squares solution is given by:

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\text{argmin}}\, Q(\beta) = (X'X)^{-1}X'Y \text{ if } \text{rank}(X) = p.$$

3.2 Linear Regression: Statistical Considerations

Simple linear regression models the relationship between a dependent (response) variable and predictor (independent) variable, extending to multiple covariates in multiple linear regression. Linear regression captures linear relationships by fitting a line to approximate the observed data. The line of best fit is obtained by finding the least squares solution for these regression coefficients (see 3.1). Diagnostic assessments of assumptions regarding the error term ϵ_i follow model fitting.

Notation: Response variable observations Y_i, design matrix $X = [1_i, X_{i1}, \ldots, X_{ip-1}]$ for $(i = 1, \ldots, n)$ with each column weighted by a regression coefficient $\beta_j (j = 0, \ldots, p-1)$. Note that $X_{i1}, \ldots, X_{ip-1}$ denote the associated covariates.

Table 2: Linear Regression Models and Assumptions

Model	**Assumptions**
Simple Linear Regression	$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
Multiple Linear Regression	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i(p-1)} + \epsilon_i$
Assumptions for ϵ_i	
Linearity	$E(\epsilon_i) = 0 \Rightarrow Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i(p-1)}$
Constant Variance	$V(\epsilon_i) = \sigma^2 \Rightarrow V(Y_i) = \sigma^2$
Independence	$Cov(\epsilon_i, \epsilon_j) = 0 \Rightarrow Cov(Y_i, Y_j) = 0$
Normality	$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \Rightarrow Y_i \sim \mathcal{N}(\mu, \sigma^2)$

3.3 Variable Relationships

Miles per gallon (mpg) is the continuous response variable predicted in the regression model (see 3.2). To identify suitable covariates for model fitting, we first examine multi-valued discrete variables—number of cylinders, model year, and origin. Given their categorical nature and the absence of ordered values, these variables are excluded from the current model framework. The remaining continuous variables—displacement, horsepower, weight, and acceleration—are examined for their relationships with the response variable. Numerical (3) and graphical summaries (2) of their pairwise correlations are obtained.

Table 3: Correlation Matrix for Selected Variables

	mpg	**weight**	**horsepower**	**displacement**	**acceleration**
mpg	1.00	-0.83	-0.78	-0.81	0.42
weight		1.00	0.86	0.93	-0.42
horsepower			1.00	0.90	-0.69
displacement				1.00	-0.54
acceleration					1.00

The correlation matrix reveals significant relationships between selected vehicle attributes: strong negative correlations exist between mpg and weight, horsepower, and displacement, while a moderate positive correlation links mpg and acceleration. Additionally, vehicle weight is strongly positively correlated with horsepower and displacement, indicating substantial interdependence among these characteristics.

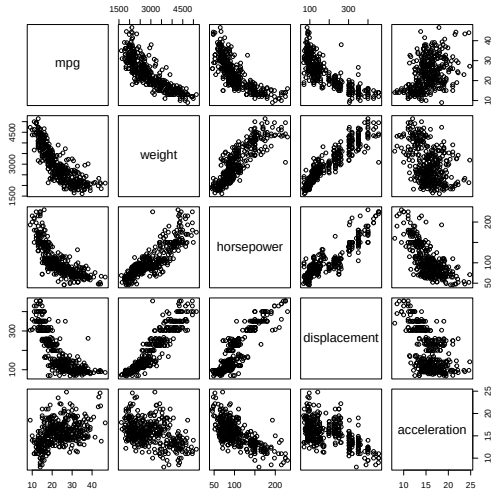


Figure 2: Correlation Matrix Plot

3.4 Model Fitting and Selection

The model fitting process involves making informed choices on model complexity, often favoring simplicity when comparable performance metrics are observed. In this analysis, the general linear test method compared a larger full model (FM) to a reduced model (RM). The three-step process includes defining the FM with all potential predictors, specifying the RM through the null hypothesis $H_0 : \beta_i = 0$, and using an F-statistic to decide whether to reject the RM in favor of the FM.

1. **Define Full Model (FM):** Includes all potential predictors.
2. **Define Reduced Model (RM):** Characterized by $H_0 : \beta_i = 0$ for i associated with the regression coefficients not present in the RM.
3. **Use F-Statistic for Comparison:** Calculated as $F^* = \frac{\frac{SSE(R)-SSE(F)}{df_R-df_F}}{\frac{SSE(F)}{df_F}}$, comparing the sum of squares for RM to FM. A rejection of H_0 suggests at least one significant relationship between β_i and Y.

We initiated model fitting with all continuous variables (displacement, horsepower, weight, and acceleration) for mpg prediction. Notably, only horsepower and weight proved significant, leading

to a refined model. A general linear test excluded displacement and acceleration, solidifying the final model ($Y = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{weight} + \epsilon$). The choice was backed by minimal differences in adjusted R-squared (0.7064 vs. 0.707) and RMSE (4.24 vs. 4.247), favoring simplicity without compromising performance.

3.5 Model Checking and Improvement

The selected model undergoes diagnostic tests to assess its appropriateness. Residual analysis, including plots of residuals vs fitted values, time sequence plots, and a normal QQ-plot, reveals problems with non-constant variance, lack of independence, and non-normality. To address these issues, a log transformation, $\ln(Y)$, is applied to the response variable. The resulting diagnostic plots are seen in Figure 3. This refined model offers a better fit to the data, mitigating the previously identified diagnostic violations. It can be succinctly expressed as:

$$\ln(\text{mpg}) = \beta_0 + \beta_1 \times \text{weight} + \beta_2 \times \text{horsepower} + \epsilon. \tag{1}$$

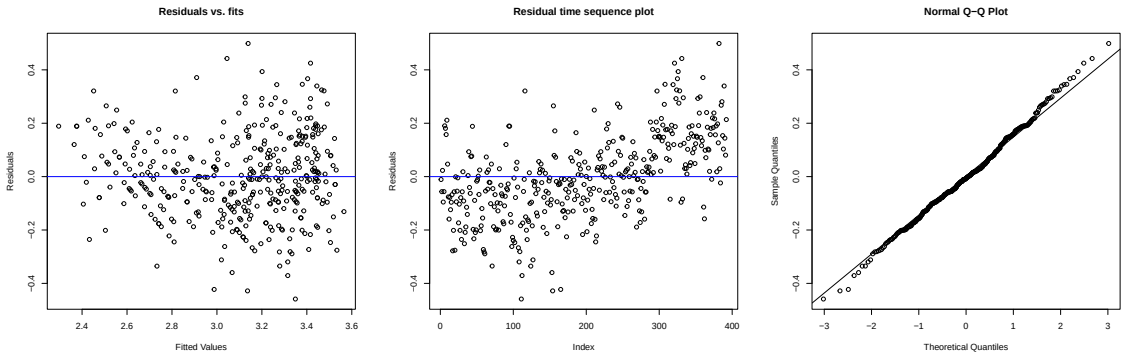


Figure 3: Diagnostic Plots - Residuals vs Fitted, Index, QQ (After Transformation)

3.6 Implications for Model Fit on Real Data

We assess the final model's fit with the transformed Y data (see 4), necessitating back-transformation for original-scale interpretations. The adjusted R-squared (0.7869) indicates a substantial reduc-

tion in total variation in mpg associated with horsepower and weight. The residual standard error (RMSE) is approximately $\exp(0.157) = 1.17$, representing the average deviation between predicted and observed mpg values. The F-test $H_0 : \beta_1 = \beta_2 = 0$ assesses whether horsepower or weight is significant to the model prediction. $P(F_{0.05,2,389} > 722.7) << 0.05$ provides evidence to reject H_0, i.e., at least one of the covariates is a significant linear predictor of mpg. Exponentiated coefficients ($\hat{\beta}_1 \approx 0.997, \hat{\beta}_2 \approx 0.9997$) suggest a 0.997 and 0.9997 increase in mpg for each unit rise in horsepower and weight respectively, holding the other constant. The 95% simultaneous confidence intervals for β_1 and β_2 are (0.997, 0.998) and (0.9997, 0.9998), respectively.

Table 4: Coefficients and Model Fit Statistics (with transformed data)

Variable	**Estimate**	**Std. Error**	**t value**	**Simultaneous CI**
`Intercept`	4.1109	0.0294	139.983	-
`horsepower`	-0.0026	0.0004	-6.231	(0.997, 0.998)
`weight`	-0.0003	0.00002	-13.462	(0.9997, 0.9998)
Residual SE	0.157	on 389 DF		
Adjusted R-squared	0.7869			
F-statistic	722.7	for 2 and 389 DF, p-value $<$ 2.2e-16		

In the real world, relationships between mpg and weight and mpg and horsepower are considered inverse. However, these confidence intervals suggest otherwise, partly due to exponentiation constraints. A consequence of necessary back-transformation implies positive coefficient estimates. So, the scale of increase is more important than the sign. We know mpg takes on values between (9, 46). In fact, most of the mpg data falls between 17-29 (see 1a). So, it is plausible that weight and mpg and horsepower and mpg are positively related for this range of mpg. Future work should explore car data with different distributions of mpg. A major limitation is the multicollinearity between the predictors horsepower and weight ($\approx$ 0.86). Computationally, this can lead to concerns about linear dependence among these columns in the design matrix X (see 3.1). One more general linear test compared the existing model (1) to reduced models with only horsepower and weight as the respective individual covariates. There was no evidence for removing either of these predictors. Future work should explore alternative approaches concerning this limitation.

4 Simulation Study

In this section, we conduct a simulation study to assess the performance of the estimation procedure applied to the real data. The generation of synthetic data from the proposed statistical model in addition to simulating numerous instances of this synthetic data is detailed in Section 4.1. Section 4.2 describes the sampling distribution for the least squares estimates in addition to graphical summaries. Section 4.3 describes how to numerically construct a confidence interval for the true regression coefficients. Finally, we consider the implications of the simulation study results for the estimator used in analyzing the real dataset in Section 4.4.

4.1 Generating Synthetic Data & Simulation Study

We generate synthetic data from the proposed statistical model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. This implies that $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, with Y being the only random variable. The design matrix (X) is fixed, and we assume that the true regression coefficients (β) are known. To obtain the true regression coefficients from our proposed model, we use the least squares solution (3.1), i.e., $\beta = (X'X)^{-1}X' \ln(Y)$. Computationally, this involves inverting the matrix $X'X$ and matrix multiplication. Having obtained the true regression coefficients (β), $X\beta$ represents the mean of our proposed statistical model.

We also introduce a random error component with mean 0 and standard deviation $\sigma = \frac{1}{n-p} \| \ln(Y) - XB \|_2^2$. Before constructing the random error term, set a random seed to ensure reproducibility. To generate the synthetic Y values, we add the mean and the random error terms together. This results in properly generated synthetic linear regression model data for a single dataset, denoted by $\ln(\text{mpg}_{\text{synth.}}) = \beta_0 + \beta_1 \times \text{weight} + \beta_2 \times \text{horsepower} + \epsilon$.

Our interest lies in maintaining the same structure as the real dataset, with X and $\ln(Y)$ representing the design matrix and the transformed *mpg* response variable, respectively. There are $n = 392$ records and $p = 3$ columns in X, corresponding to an intercept term and two covariates. The objective of the simulation study is to assess properties of the estimator used for the

real dataset on synthetic data that exactly follows the assumed distribution. We estimate the coefficients using the same procedure, i.e. $\hat{\beta} = (X'X)^{-1}X' \ln(Y_{\text{synth.}})$. The computational details are equivalent. Note that base R functions can also be used for fitting a (multiple) linear regression model. Increasing the sample size for the synthetic data will show the convergence of the estimators' sampling distribution to the true mean value. Given the relatively large size ($n = 392$) and the preference for this dimension to match the real dataset, we keep this value fixed.

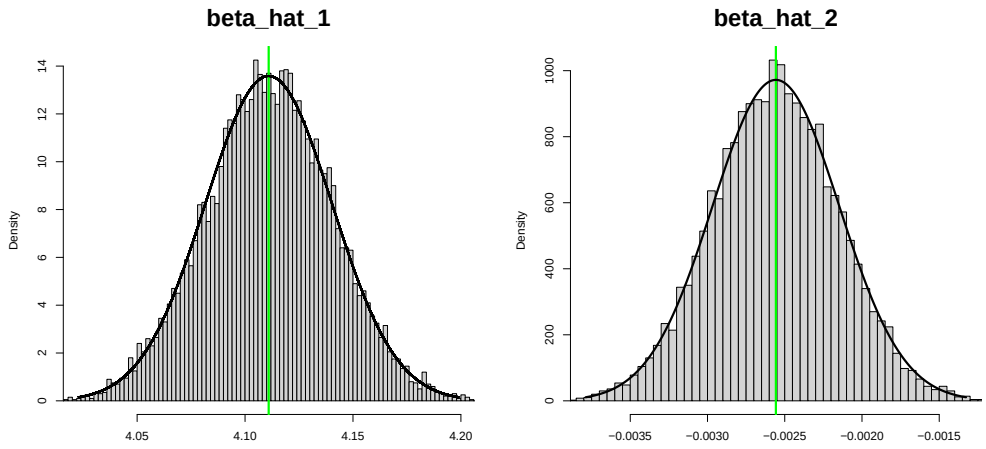
4.2 Sampling Distribution of Estimator

The number of simulations is another consideration that we can change. That is, we can simulate the estimation procedure from this synthetic data set numerous times. In doing so, we can assess the sampling distribution of these estimates ($\hat{\beta}$). For a fixed sample size, increasing the number of simulations will more closely resemble the estimators' overall sampling distribution. A histogram with an overlaid density plot is used to visually assess the estimators' sampling distribution for different simulation sizes. Starting at 1000 and ascending by sequences of 1000, we can choose a value for which the sampling distribution was most closely approximated by the density curve. $N = 10,000$ is the number of simulations for which this approximation was appropriately met.

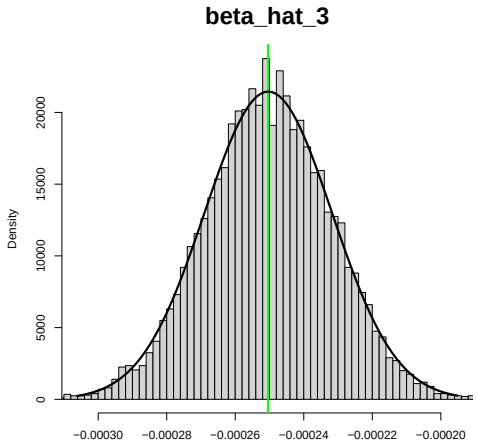
Simulation studies enable the numerical generation of plots depicting the sampling distributions of the estimator. Even without the ability or knowledge about how to derive an analytical sampling distribution, these plots can be informative. In this case, the sampling distribution for the least squares solution is known. It can be shown that

$$\text{For rank}(X) = p, \hat{\beta} = (X'X)^{-1}X'Y \sim \mathcal{N}(\beta, (X'X)^{-1}\sigma^2). \tag{2}$$

Figure 5 shows the sampling distributions of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$. The density curve is overlaid with the functional form described in Equation (2). This provides visual evidence that the sampling distribution of the least squares estimator converges to normality with mean β. This suggests that, on average, the true population features can be appropriately estimated from the synthetic data.



(a) $\hat{\beta}_0$ Sampling Distribution (b) $\hat{\beta}_1$ Sampling Distribution



(c) $\hat{\beta}_2$ Sampling Distribution

Figure 4: Sampling Distributions of Least Squares Estimates From Simulation Study

4.3 Confidence Interval for Regression Coefficients

The estimated regression coefficients are point estimators for the true coefficients. In the simulation study, we can verify the coverage of the confidence intervals for these least squares point estimates. Specifically, we check if

$$1 - \alpha = P(\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}}(\sigma\sqrt{(X'X)^{-1}_{jj}})),$$

where α is the coverage and $j = 0, 1, 2$. In our context,

$$P(\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}} \cdot SE(\hat{\beta}_j) \ni \beta) \approx \frac{1}{k} \sum_{i=1}^{k} \mathbf{1}(\hat{\beta}_j \pm Z_{1-\frac{\alpha}{2}} \cdot SE(\hat{\beta}_j) \ni \beta),$$

where k is the number of simulations and β is the known regression coefficeint.

For this case, a 95% confidence interval for β was numerically obtained from the simulation study. The observed coverage values for $\hat{\beta}_i (i = 0, 1, 2)$ are 0.9477, 0.9524, and 0.9509, respectively. These results indicate that the correct coverage is approximately achieved for a large number of simulations.

4.4 Connection to Real Dataset

The outcomes of our simulation study instill confidence in the validity of the estimator applied to the real dataset. The study provides evidence that the least squares estimates exhibit normality, centered around the true regression coefficients. Confidence intervals for the true regression coefficients are obtained from the simulation study. The observed coverage values show convergence to the predetermined true coverage values.

These promising findings imply that the least squares estimator is a suitable approach for the real data, particularly when it adheres closely to the linear regression model. Consequently, diagnostic methods for assessing assumptions in the real data become crucial. If these assumptions can reasonably be satisfied, the simulation study suggests that the linear regression model will aptly capture and fit the real data.

5 Bootstrappping

In this section we conduct a bootstrap study. Section 5.1 describes the procedure for generating one boostrapped data set from the real data set and statistical model. Section 5.2 presents the sampling distribution of the regression coefficient estimates from the boostrapped subsample. Section 5.3

compares the inference procedures drawn from the real data set to the bootstrapped data.

5.1 Bootstrapping Procedure

The basic idea of bootstrapping techniques is to subsample from an observed data set to approximate the sampling distribution of a statistic. The reason for why this is a reasonable idea comes from theory which establishes that under usual conditions, the empirical distribution function (EDF) will approximate the CDF. Note that the EDF is defined for a sample $x_1, \ldots, x_n$ as

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i \leq x\}.$$

A bootstrapped sample of size $k < n$ is defined as $x_1^*, \ldots, x_k^* \overset{i.i.d.}{\sim}$ Uniform$(\{x_1, \ldots, x_n\})$. Note that boostrapped sampling is without replacement.

Next, we consider how to bootstrap the sampling distributions of the coefficients in the simple linear regression model. Suppose $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$ for $i \in \{1, \ldots, n\}$. Here, since the $\epsilon_i, \ldots, \epsilon_n$ are unobservable, we could instead consider the pairs $(x_1, y_1), \ldots, (x_n, y_n) \overset{i.i.d.}{\sim} f_{x,y}$, where $f_{x,y}$ is a joint distribution. Then, we define the bootstrap samples as

$$(x_1^*, y_1^*), \ldots, (x_m^*, y_m^*) \overset{i.i.d.}{\sim} \text{Uniform}(\{(x_1, y_1), \ldots, (x_n, y_n)\}) \text{ for } m < n. \tag{3}$$

Note that this is still sampling with replacement. Next, we compute the bootstrapped coefficient estimates using the method of least squares.

Recall the final MLR model we used in the real data, i.e. $\ln(\text{mpg}) = \beta_0 + \beta_1 \times \text{weight} + \beta_2 \times \text{horsepower} + \epsilon$ for $i \in \{1, \ldots, 392\}$. We estimated the regression coefficients by method of least squares. That is, $\hat{\beta} = (X'X)^{-1}X' \ln(Y)$ where X_{i1}, X_{i2}, X_{i3} correspond to the columns for the intercept, weight, and horsepower explanatory variables respectively and $ln(Y)$ corresponds to the transformed mpg variable. Computationally, this involves matrix inversion and matrix multiplication.

Now consider the coefficient estimates obtained after generating one bootstrapped data set from the real data set and statistical model. Following the procedure developed in Equation (3), we generalize this approach to apply to our MLR model. First, let the size of each bootstrapped sample (m) be 70. Once again, we estimate the regression coefficients using least squares. That is, $\hat{\beta} = (X^{*\prime}X^*)^{-1}X^{*\prime}\ln(Y^*)$ where $X^*_{i1}, X^*_{i2}, X^*_{i3}$ correspond to the columns for the intercept, weight, and horsepower explanatory variables respectively for the subsampled data. Similarly, $\ln(Y^*)$ corresponds to the transformed mpg variable from the subsampled data. The remaining computational details stay the same.

5.2 Sampling Distribution of Bootstrapped Estimator

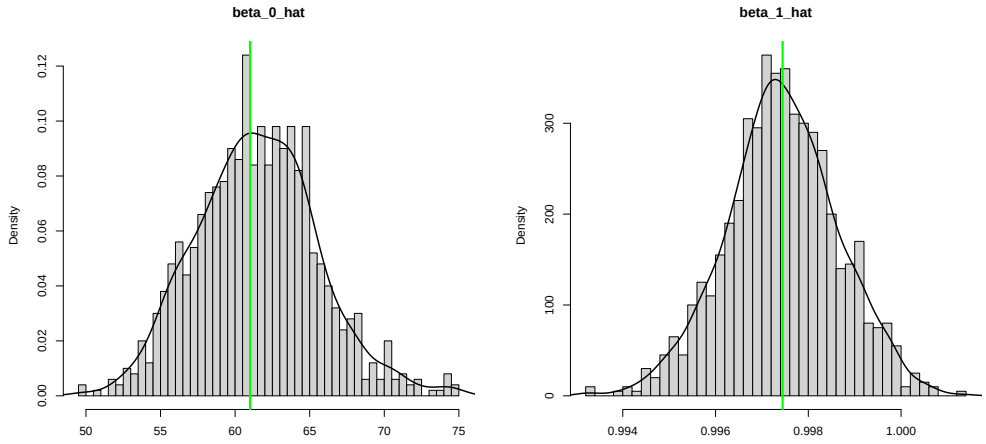
We aim to expand the bootstrap procedure to generate bootstrapped data for a larger sample size (N) to approximate the sampling distributions of the estimated regression coefficients. We are bootstrapping samples from a finite distribution, which may introduce bias when the tails of the distribution are not well-represented. This bias can lead to a narrower spread in the empirical distribution compared to the population distribution.

To address this, we increase the number of bootstrap samples from the real dataset to better reflect the overall sampling distribution of the estimators in the MLR model. Using random sampling with replacement, we set a random seed for reproducibility.

Table 5: Bootstrap Estimates of Regression Coefficients

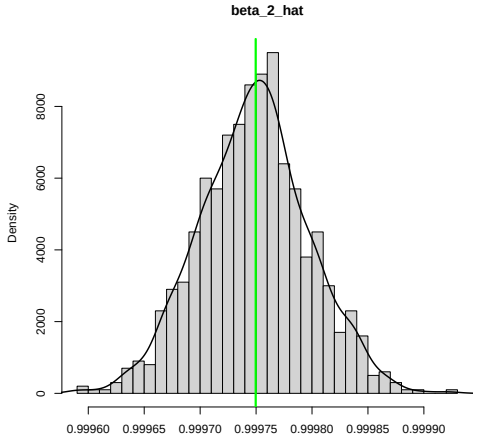
Parameter	Mean	Standard Deviation
$\hat{\beta}^*_0$	61.34	1.07
$\hat{\beta}^*_1$	0.997	1.001
$\hat{\beta}^*_2$	0.9997	1.000

We visually assess the sampling distribution of estimators across various bootstrap sample sizes using histograms with overlaid density plots. Starting at a sample size of 100 and increasing by increments of 100, we determine the sample size where the sampling distribution aligns closely with the previous increment size. We find that $N = 1000$ is an appropriate number of bootstrap



(a) $\hat{\beta}_0^*$ Sampling Distribution

(b) $\hat{\beta}_1^*$ Sampling Distribution



(c) $\hat{\beta}_2^*$ Sampling Distribution

Figure 5: Sampling Distributions of Least Squares Estimates From Bootstrapped Subsamples

samples for this approximation. Figure 5 illustrates unimodal, bell-shaped, and approximately symmetric curves for each respective bootstrapped coefficient estimate. We quantify the sampling distribution using mean and standard deviation measures, available in Table 5.

5.3 Inference Procedure Comparisons

In Section 3.6, we evaluated the MLR model's fit using the real data and determined the 95% confidence intervals for β_1 and β_2 to be (0.9966, 0.9983) and (0.9997, 0.9998) respectively. These intervals test $H_o : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ and $H_o : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$. Since neither interval

contains zero, weight and horsepower are deemed significant (linear) predictors of mpg.

While real data inference used t-distributed critical values, bootstrapped subsamples made no assumptions about the population distribution. We computed percentile bootstrap confidence intervals using empirical $\alpha/2$ and $1 - \frac{\alpha}{2}$ quantiles. The 95% confidence intervals from bootstrapped subsamples for β_1 and β_2 were (0.9950, 0.9998) and (0.9997, 0.9998) respectively. The absence of zero again suggests weight and horsepower are respectively significant predictors of mpg.

Table 6 summarizes these findings, indicating similar effect sizes between real data estimation and bootstrapped subsamples. Consequently, the conclusions remain unchanged regarding weight and horsepower as predictors of mpg.

Table 6: Comparison of Confidence Intervals

Parameter	**Real Data**	**Bootstrapped Subsample**
β_1	(0.9966, 0.9983)	(0.9950, 0.9998)
β_2	(0.9997, 0.9998)	(0.9997, 0.9998)

Appendix: Code Used

The code used in this project has the following files and methods:

1. **RealDataAnalysis.r**: Preliminary exploration, cleaning and model fitting procedure
2. **SimStudy.r**: Generating synthetic data and running the estimation procedure
3. **SimStudyResults.r**: Produces plots/tables for simulation study
4. **Bootstrap.r**: Conducts bootstrapping procedure
5. **BootstrapResults.r**: Produces plots/tables for bootstrap procedure
6. **workflow.sh**: Provides steps for reproducing all analysis and results in appropriate order
7. *Complete repository available at:* `https://github.com/bryant-willoughby/WilloughbyST495.git`