AUTOMOBILES DATA ANALYSIS

Spring 2023

Ian Shaw, Richard Murad | STAT 3120 | 4/26/2023

Introduction

The automobile dataset consists of 398 vehicles manufactured from the 1970s to the early 1800s. This dataset includes the following specifications:

Variable Name	Description
mpg	fuel efficiency measured in miles per gallon
cylinders	number of cylinders in engine
displacement	total volume of engine cylinders (in cubic inches)
horsepower	engine power
weight	vehicle weight in pounds
acceleration	time (in seconds) to accelerate from 0 to 60 mph
model_year	model year
origin	country origin of vehicle
car_name	car name

The overall goal of this analysis is to determine whether certain relationships between variables are significant. More specifically, we would like to find out if the weight and fuel efficiency of a vehicle (mpg) have a relationship with the acceleration of a vehicle. Car manufacturers would find this data useful, since they often need to know what the performance of a vehicle looks like before production begins. [1] Computer modeling has merits in this field, a similar model made in 1961 showed extreme promise in predicting car performance.[2] As for our categorical variables, another research question that will be examined is if the country origin and number of cylinders in an engine correlate to one another.

[1] Setz, Henry L. "Computer Predicts Car Acceleration." *SAE Transactions*, vol. 69, 1961, pp. 351–60. *JSTOR*, http://www.jstor.org/stable/44553930.

[2] Ordorica, M. A. "Vehicle Performance Prediction." *SAE Transactions*, vol. 74, 1966, pp. 168–76. *JSTOR*, http://www.jstor.org/stable/44554201.

Methods

Variable Description and Choice Motive

Multiple Linear Regression Model

For our multiple linear regression test, we chose acceleration as our dependent variable. To explain acceleration, we chose fuel efficiency (mpg) and car weight (pounds). Both fuel efficiency and weight are commonly tracked metrics for manufactured cars, so using them to predict a less measured variable could prove useful. However, the model that these variables comprise were chosen primarily due to massive problems with other variables. Collinearity of independent variables, clustering by number of cylinders, and non-linear relationships plagued most other models we considered. With our linear regression analysis, we hope to answer the question: Does the weight and fuel efficiency of a vehicle (mpg) have a relationship with the acceleration of a vehicle?

First, we examine the continuous response variable, acceleration, in our multiple linear regression model. Acceleration was measured in terms of the time (in seconds) that a car took to reach 60 mph from 0 mph. The fastest car accelerated to 60 mph in 8 seconds, whereas the slowest took 24.8 seconds. The mean and median were 15.5 and 15.57 seconds respectively, with a standard deviation of 2.76. One continuous explanatory variable used was fuel efficiency, measured in miles per gallon. A larger value indicates a more fuel-efficient vehicle, and a smaller value indicates a less efficient vehicle. Miles per gallon peaked at 46.6 and observed its lowest value at 9. The mean and median mpg were 23.51 and 23 respectively, with a standard deviation of 7.82. Finally,

the last continuous explanatory variable was weight, measured in pounds, ranging from 1613 pounds to 5140 pounds. The mean and median weight were 2970.42 and 2803.5 pounds respectively.

Chi-Squared Test

Our Chi-Squared test of independence tested the country of origin of a car and how many cylinders its engine had. These variables were chosen to examine the effect of cultural differences on the propensity to produce certain types of cars. Origin and engine cylinders were also the most appropriate variables to use in a Chi-Squared test, so long as we removed the highly infrequent 3-cylinder and 5-cylinder cars. Through this test, we hope to find out how the country of origin for a car relates to the number of cylinders its engine has.

In our Chi-Squared test of independence, one categorical variable we used was country of origin (American, European, and Japanese). American cars comprised almost 250 of the total observations, where Europe and Japan both produced in the 70s range. We tested origin for independence with our other categorical variable, the number of cylinders. 3-cylinder and 5-cylinder cars comprised a touch over 2% of the total combined, and 4-cylinder cars were by far the most popular, more than doubling the second place 8-cylinder engine. 6-cylinder engine cars weren't far behind, comprising twenty fewer cars than 8-cylinder.

Descriptive Analysis Reasoning

To look at our quantitative statistics used in our linear regression model, we produced a table of descriptive statistics that described the distributions of each variable.

Furthermore, to examine the bivariate relationships between our variables, we produced a correlation matrix and a scatterplot matrix.

Finally, we produced frequency tables and bar plots of the origin and cylinder variables to analyze their univariate distribution. We then produced a contingency table to examine how the two variables interacted with each other.

Procedure Choice Rationale

We chose to perform a multiple linear regression analysis since our quantitative variables were best suited for predicting and forecasting and we did not possess a binary variable to make a different procedure more accessible. Some of the variables are uncommonly measured, so creating a model to predict an elusive variable seemed like a good idea. Of course, we assessed a numerous number of assumptions: assumption of linearity, normality of residuals, homogeneity of residual variance, residual independence, and collinearity. The evidence of the plots/tests will be shown below.

A Chi-Squared test was chosen since it suited the dataset the best. When looking at cars, the odds and risk of traits are less interpretable and useful than in other contexts. With a Chi-Squared test, the results have more meaning since it uncovers a pattern in car manufacturing and points towards a relationship existing if significant. To start, we checked 4 conditions that need to be met: simple random sample, no overlap in categories, two categorical variables, and expected frequency count cells greater than 5. A data cleaning step was implemented to help with expected frequencies satisfying the requirement; removing all vehicles with 3 and 5 cylinders was a great way to do so

because of low frequency. A contingency table was used to verify the last assumption, and the rest of the assumptions were intuitive, so no graphics were needed.

Both alpha thresholds for both inferential procedures are 0.05

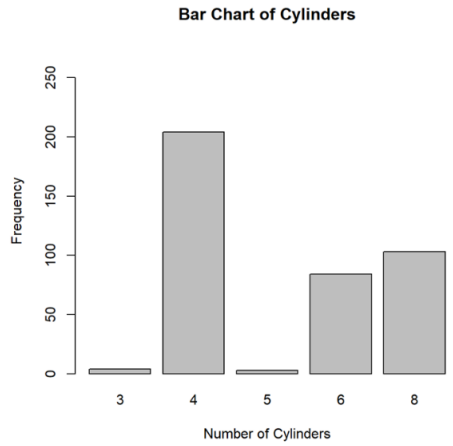
Results

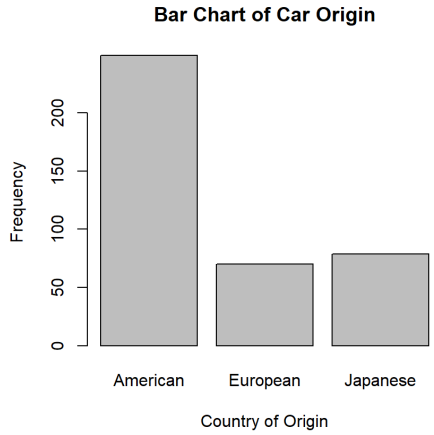
Descriptive Graphics

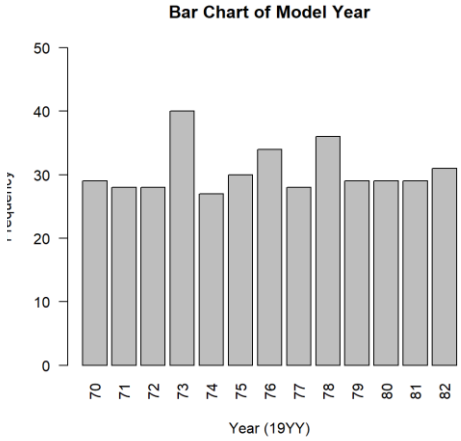
Descriptive Statistics of Quanitative Variables

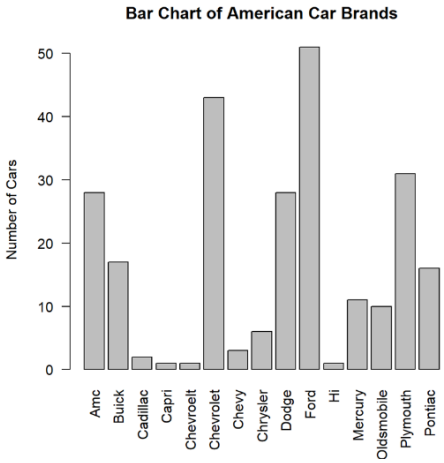
1.5% of Horsepower is missing.

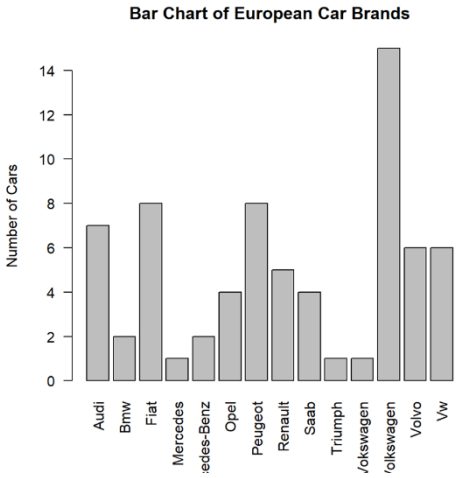
Variable	N	Mean	Standard Deviation	Median	Min	Max	Skew	Kurtosis
MPG	398	23.51	7.82	23.0	9	46.6	0.45	-0.53
DISPLACEMENT	398	193.43	104.27	148.5	68	455.0	0.71	-0.76
HORSEPOWER	392	104.47	38.49	93.5	46	230.0	1.08	0.65
WEIGHT	398	2970.42	846.84	2803.5	1613	5140.0	0.53	-0.80
ACCELERATION	398	15.57	2.76	15.5	8	24.8	0.28	0.38

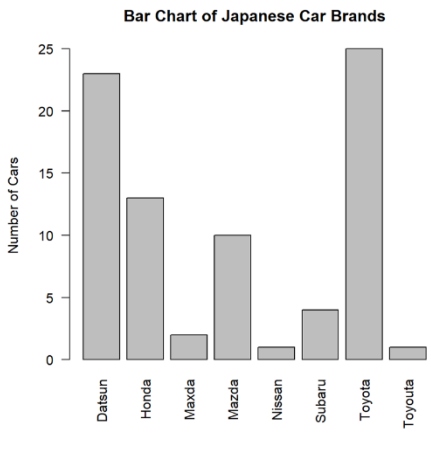












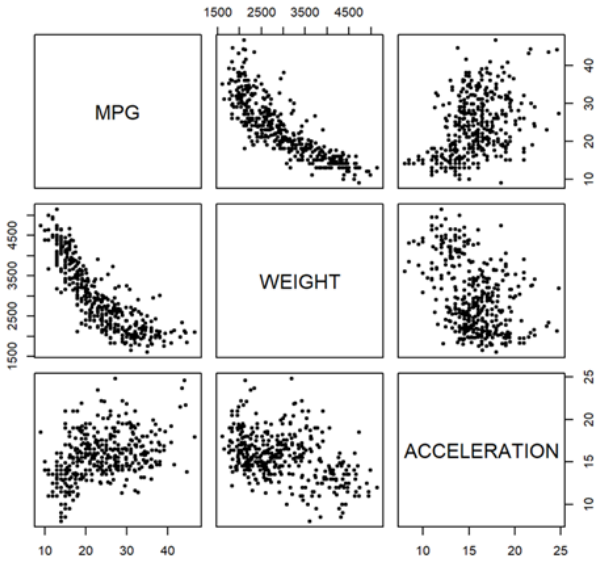
Assumptions for MLR

1) Assumption of Linearity

Correlation Matrix

Values are correlation coefficient (r)

	MPG	WEIGHT	ACCELERATION
MPG	1.000	-0.832	0.420
WEIGHT	-0.832	1.000	-0.417
ACCELERATION	0.420	-0.417	1.000



In comparison to our response variable (acceleration), we can see that our two predictor variables (mpg and weight) have a positive linear relationship and a negative linear relationship respectively. The r value of each predictor variable is between ± 0.25 to ± 0.50 which is considered a weak association (at least for this class). Nonetheless, both r values are on the high end of this interval, closer to the moderate threshold at ± 0.50. Intuitively, it makes sense for the acceleration variable to decrease as the weight of a vehicle increases since there will be more force opposing motion. As for our other independent variable, it does make some sense for acceleration time to be longer as fuel efficiency increases. Anytime you accelerate a vehicle, the engine will work harder to burn more fuel. Of course, fuel efficient cars will oppose that type of fuel burning, otherwise it wouldn't be efficient. As a result, the vehicle will be forced to accelerate at a slower time. Overall, the assumption of linearity is supported.

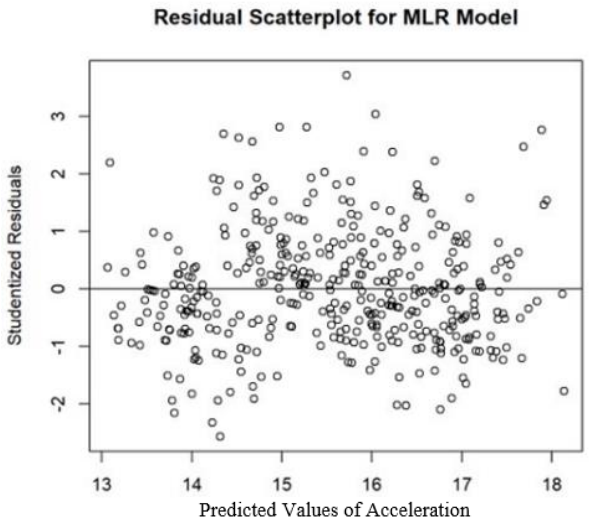
2) Normality of Residuals

2a. Skewness/Kurtosis

Skewness	0.53515574	Kurtosis	0.49537732

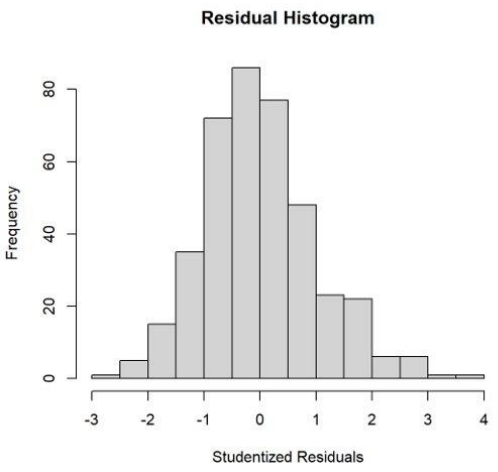
As you can see, the skewness value is 0.535 which is inside the interval of [-2,2] and the kurtosis value is 0.495 which is well within the range of [-3,3]. As a result, kurtosis and skewness support the assumption of normality.

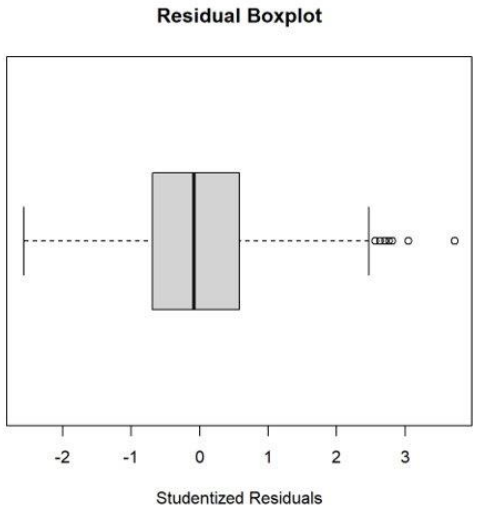
2b. Residual Plot



We observe that the residuals are approximately distributed evenly across the regression line (horizontal line at 0). There is no obvious sign of the residuals being dominated on either side for this residual plot. The distribution of these residuals seems to be random, and that is exactly what we are looking for to conclude that this evidence supports normality.

2c. Histogram/Boxplot/Outliers/Sample Size

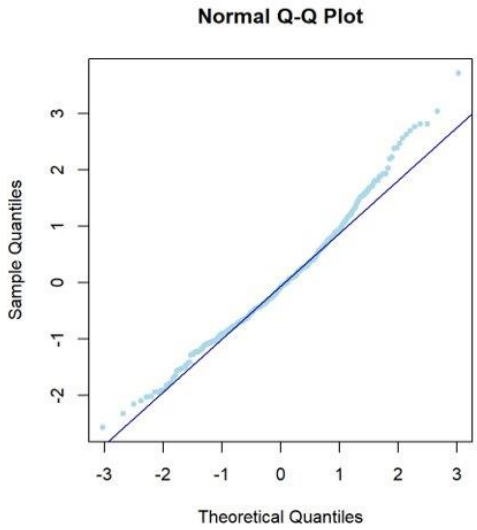




Obs	resids
60	3.01539
300	3.66528

The histogram seems to be a bit skewed on the right side and there is no obvious sign of the spread being too flat or peaky. For the boxplot, the whiskers are nearly the same length, but we have outliers that will extend the right whisker, causing it to be slightly skewed right like the histogram. Also, the mean and median are both very close to each other (the mean isn't shown in the boxplot, but its value is extremely close to 0). Here we only have two real outliers (unlike the boxplot) that are more than 3 standard deviations away from the mean. Since our sample size is huge with 398 observations, these two outliers don't have much effect on our data. Overall, the boxplot and histogram support the assumption of normality since the slight right skewness for both graphs are very minimal. Of course, our sample size of 398 is well above 30 which also supports normality.

2d. Q-Q Plot



Looking at the Q-Q plot, there is about a 5% deviation in the left tail and about a 15%-20% deviation in the right tail. The data points become a bit messy as we approach the end of both tails, especially for the right tail. These data points aren't severe enough to violate normality. So, the q-q plot supports normality.

2e. Tests of Normality

Tests for Normality

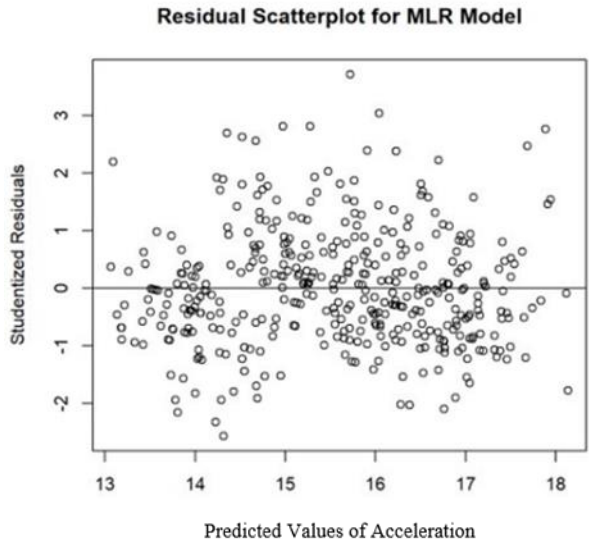
Test	Statistic	P-value
Shapiro-Wilk (W)	0.98134	0.00005
Kolmogorov-Smirnov (D)	0.05764	0.00292
Cramer-von Mises (W-Sq)	0.34791	0.00010
Anderson-Darling (A-Sq)	2.09826	0.00002

These normality tests are significant since the p-values are all less than our alpha (.05). Correspondingly, these tests violate our assumption of normality. It is important to

note that these tests are sensitive to sample size and again, our sample size of 398 is very large, so this specific piece of evidence will be taken with a grain of salt.

Now that we have covered all pieces of evidence for the assumption of normality, it is safe to say that this assumption meets the requirements and is now checked off our list.

3) Homogeneity of Residual Variance



For this assumption, we are analyzing the variance of the residuals. After looking at this residual plot for a bit, there is no increase or decrease in the studentized residuals (‘funnel shape’) along the regression line. We concluded earlier that the residuals seem to be randomly distributed across the line, which further disproves that there is a ‘funnel shape’ look in our residual plot. This assumption is supported.

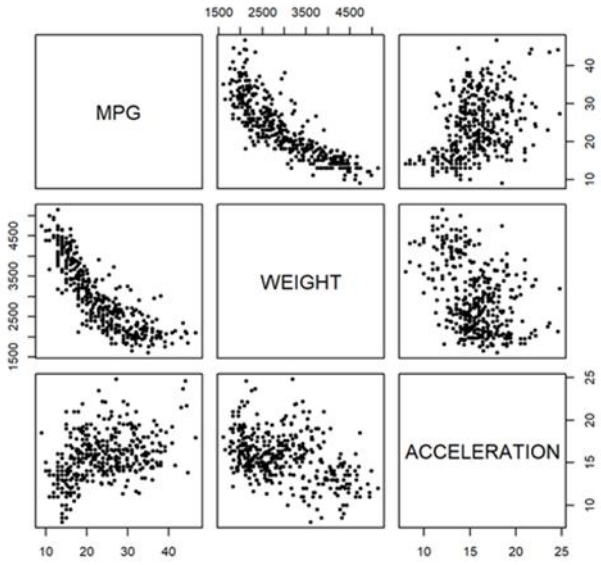
4) Independence of Residuals

This assumption comes down to study design. This data was collected at a single point in time, meaning that it is cross sectional data, which is assumed to be independent for residuals. Therefore, this assumption is supported.

5) No Collinearity Between Explanatory Variables

Variance Inflation Factor

WEIGHT	MPG
3.244572	3.244572



Correlation Matrix

Values are correlation coefficient (r)

	MPG	WEIGHT	ACCELERATION
MPG	1.000	-0.832	0.420
WEIGHT	-0.832	1.000	-0.417
ACCELERATION	0.420	-0.417	1.000

Our Variance Inflation Factor (VIF) values are between 1 and 5, indicating moderate collinearity. Furthermore, the strong correlation between fuel efficiency and weight points towards significant collinearity. Given these two factors, we conclude that collinearity is violated.

Assumption Results for χ^2 Test of Independence

Chi-Squared Table for Car Origin by Number of Cylinders

3 and 5 cylinder cars omitted due to infrequency

# of Cylinders	Origin		
	American	European	Japanese
Observed			
4	72.00	63.00	69.00
6	74.00	4.00	6.00
8	103.00	0.00	0.00
Expected			
4	129.91	34.96	39.13
6	53.49	14.39	16.11
8	65.59	17.65	19.76
Chi-Squared			
4	25.82	22.50	22.80
6	7.86	7.51	6.35
8	21.33	17.65	19.76
X-squared = 151.57, df = 4, p-value < 2.2e-16			

Our conditions for this inferential procedure are straightforward. We assume that the sample collected is a simple random sample. From the frequency table above, there are two categorical variables with 3 categories each and there cannot be any overlap in these categories. It doesn't make sense to have a car manufactured in two different countries or a car having both a 4-cylinder and 6-cylinder engine simultaneously. This leads us with one condition, and thanks to our data cleaning step, all expected values are greater than 5 as you can see. In conclusion, all 4 assumptions are met, and our new sample size is 391 after we removed 3-cylinder and 5-cylinder cars.

Inferential Results for MLR

Regression Output (f = 46.81, p < 0.001)

	ACCELERATION			
Predictors	*Estimates*	*CI*	*Statistic*	*p*
(Intercept)	15.73164	12.96498 – 18.49831	11.17889	**<0.001**
WEIGHT	-0.00072	-0.00124 – -0.00020	-2.70297	**0.007**
MPG	0.08365	0.02713 – 0.14018	2.90949	**0.004**
Observations	398			
R^2 / R^2 adjusted	0.192 / 0.188			

Our overall model is significant, and individually our predictors are as well. Individually, both vehicle weight (t=-2.70, p=.007) and fuel efficiency (t=2.91, p=.004) are significant predictors of acceleration. The R^2 value tells us that 19.2% of the variation in acceleration is explained by vehicle weight and fuel efficiency. Furthermore, for every one-pound increase in vehicle weight, there is on average a .00072 decrease in time to accelerate to 60 mph while all other variables are held constant (t=-2.70, p=.007), and for every mile per gallon increase in fuel efficiency, there is on average a .084 increase in time to accelerate to 60 mph while all other variables are held constant (t=2.91, p=.004).

Inferential Results for χ^2 Test of Independence

Chi-Squared Table for Car Origin by Number of Cylinders

3 and 5 cylinder cars omitted due to infrequency

# of Cylinders	Origin		
	American	European	Japanese
Observed			
4	72.00	63.00	69.00
6	74.00	4.00	6.00
8	103.00	0.00	0.00
Expected			
4	129.91	34.96	39.13
6	53.49	14.39	16.11
8	65.59	17.65	19.76
Chi-Squared			
4	25.82	22.50	22.80
6	7.86	7.51	6.35
8	21.33	17.65	19.76
X-squared = 151.57, df = 4, p-value < 2.2e-16			

Our p-value is less than 0.05, meaning we have significant findings. There is a significant relationship between country of origin and the number of cylinders in an engine. (X^2 = 151.57, df = 4, n=391, p<.001)

Code Appendix

SAS Code Used for Graphics:

```
/* Normality tests */
proc univariate data=stdres normal;
var resids;
run;


/* print outliers over 3 std away from mean */
PROC PRINT data= stdres;
var resids;
where abs(resids)>=3;
RUN;
```

R Code Used for Graphics:

```
library(readxl)
library(MASS)
library(e1071)
library(car)
library(gt)
library(psych)
library(stringr)
library(tidyverse)
library(dplyr)
library(sjPlot)
library(sjmisc)
library(sjlabelled)
library(mice)
library(tibble)

data <- read_xlsx("AutomobilesDataset.xlsx")

#### QUANITATIVE VARIABLES ####
data$HORSEPOWER <- as.numeric(data$HORSEPOWER)
data_quan = data[-c(2,7:9)]

q_table <- describe(data_quan, na.rm = TRUE)
q_table <- q_table[-c(1,6,7,10,13)]
q_table <- q_table |> as.data.frame()
q_table <- rownames_to_column(q_table, "VAR")
q_table <- q_table %>% mutate_if(is.numeric, round, digits = 2)
print(q_table)

cnames <- str_to_title(names(q_table))
cnames[4] <- "Standard Deviation"
names(q_table) <- cnames
```

```r
# Table of descripitve stats
gt_quan <- gt(q_table, rowname_col = "Var")

gt_quan <- gt_quan |> tab_stubhead(label = "Variable")
gt_quan <- gt_quan |> tab_header(
  title = "Descriptive Statistics of Quanitative Variables",
  subtitle = "1.5% of Horsepower is missing.")

gt_quan <- gt_quan |> opt_stylize(style = 2)

gt_quan |> gtsave("descriptive.png")

### QUALITATIVE ####

# Converting Cylinders & Model Year to char value
data$CYLINDERS <- as.character(data$CYLINDERS)
data$MODEL_YEAR <- as.character(data$MODEL_YEAR)

# Recoding the origin variable into what they represent
data$ORIGIN[data$ORIGIN == 1] <- "American"
data$ORIGIN[data$ORIGIN == 2] <- "European"
data$ORIGIN[data$ORIGIN == 3] <- "Japanese"

# Splitting Car Name into just the first word to make it more useful
data$BRAND <- str_to_title(str_split_i(data$`CAR NAME`, " ", 1))

# Dot product to check if brands produced in multiple locations (they
didn't)
brand_origin <- table(data$BRAND, data$ORIGIN)
c(
  brand_origin[,1] %*% brand_origin[,2],
  brand_origin[,1] %*% brand_origin[,3],
  brand_origin[,2] %*% brand_origin[,3]
)

list_bo <- list(brand_origin[,1], brand_origin[,2], brand_origin[,3])
|>
  lapply(function(x) x[x!=0])

df_a <- list_bo[1] |> unlist() |> as.data.frame.vector() |>
  rownames_to_column("C")
names(df_a) <- c("C","Frequency")
df_e <- list_bo[2] |> unlist() |> as.data.frame.vector() |>
  rownames_to_column("C")
names(df_e) <- c("C","Frequency")
df_j <- list_bo[3] |> unlist() |> as.data.frame.vector() |>
  rownames_to_column("C")
names(df_j) <- c("C","Frequency")


gt_a <- gt(df_a, rowname_col = "C") |> tab_stubhead("Brand") |>
  tab_header("Frequency Table of American Car Makers",
             subtitle = "15 of the 37 Brands were American") |>
  opt_stylize(style = 2)
```

```r
gt_e <- gt(df_e, rowname_col = "C") |> tab_stubhead("Brand") |>
  tab_header("Frequency Table of European Car Makers",
             subtitle = "14 of the 37 Brands were European") |>
  opt_stylize(style = 2)

gt_j <- gt(df_j, rowname_col = "C") |> tab_stubhead("Brand") |>
  tab_header("Frequency Table of Japanese Car Makers",
             subtitle = "8 of the 37 Brands were American") |>
  opt_stylize(style = 2)

gt_a |> gtsave("brandorigin_amer.png")
gt_e |> gtsave("brandorigin_euro.png")
gt_j |> gtsave("brandorigin_japn.png")

barplot(
  unlist(list_bo[1]),
  xlab = "",
  ylab = "Number of Cars",
  main = "Bar Chart of American Car Brands",
  las = 2
)
barplot(
  unlist(list_bo[2]),
  xlab = "",
  ylab = "Number of Cars",
  main = "Bar Chart of European Car Brands",
  las = 2
)
barplot(
  unlist(list_bo[3]),
  xlab = "",
  ylab = "Number of Cars",
  main = "Bar Chart of Japanese Car Brands",
  las = 2
)
# Model Year
year_table <- table(data$MODEL_YEAR)
barplot(
  year_table,
  xlab = "Year (19YY)",
  ylab = "Frequency",
  main = "Bar Chart of Model Year",
  ylim = c(0,50),
  las = 2
)
year_df <- as.data.frame.vector(year_table) |> rownames_to_column(var =
"YR")
year_gt <- gt(year_df, rowname_col = "YR") |> tab_header(
  title = "Frequency Table of Model Year") |>  tab_stubhead("Year") |>
  cols_label(year_table = "Frequency") |>  opt_stylize(style = 2)

year_gt |> gtsave("year.png")

# Cylinders
cyl_table <- table(data$CYLINDERS)
```

```r
barplot(
  cyl_table,
  xlab = "Number of Cylinders",
  ylab = "Frequency",
  main = "Bar Chart of Cylinders",
  ylim = c(0,275)
)
cyl_df <- as.data.frame.vector(cyl_table) |> rownames_to_column(var =
"CYL")
cyl_gt <- gt(cyl_df, rowname_col = "CYL") |> tab_header(
  title = "Frequency Table of Engine Cylinders") |>
tab_stubhead("Cylinders") |>
  cols_label(cyl_table = "Frequency") |>  opt_stylize(style = 2)

cyl_gt |> gtsave("cyl.png")

org_table <- table(data$ORIGIN)
barplot(
  org_table,
  xlab = "Country of Origin",
  ylab = "Frequency",
  main = "Bar Chart of Car Origin",
)


#### MULTIPLE LINEAR REGRESSION MODEL ####
# DEPENDENT VARAIBLE: ACCELERATION
# INDEPENDENT VARAIBLES: WEIGHT, MPG

# Creating model
accel_model <- lm(ACCELERATION ~ WEIGHT + MPG, data=data)

# Trimming down dataframe
data_acc <- data[c(1,5,6)]

# Getting mlr output (t-values, f-values, parameter est., etc.)
summary(accel_model)

model_table <- tab_model(accel_model, show.ci = .95, show.stat = TRUE,
                         show.fstat  = TRUE, digits = 5,
                         title = "Regression Output (f = 46.81, p <
0.001)")

# PLOTS FOR ASSUMPTIONS #

# Plotting studentized residual scatterplot
accel_res <- studres(accel_model)
plot(
  fitted(accel_model),
  accel_res,
  xlab = "Time to accelerate from 0 to 60 mph (s)",
  ylab = "Studentized Residuals",
  main = "Residual Scatterplot for MLR Model"
) +
  abline(0,0)
```

```r
median(studres(accel_model))

length(accel_res[abs(accel_res) < 2])

# Correlation matrix w/ scatterplots
cor_matrix <- cor(data_acc, use = "complete.obs")
pairs(data_acc, pch = 20)

cor_df <- as.data.frame.matrix(cor_matrix) |> rownames_to_column(var =
"Var")
cor_df <- cor_df %>% mutate_if(is.numeric, round, digits = 3)

cor_gt <- cor_df |> gt(
  rowname_col = "Var") |> tab_header(
    "Correlation Matrix", subtitle = "Values are correlation
coefficient (r)") |>
  opt_stylize(style=2)

cor_gt |> gtsave("cormatrix.png")

par(pty="m")

# Histogram & box plot
hist(
  accel_res,
  main = "Residual Histogram",
  xlab = "Studentized Residuals",
)
plot(density(accel_res))
boxplot(accel_res, horizontal = TRUE, xlab = "Studentized Residuals",
        main = "Residual Boxplot")

# Q-Q Plot
par(pty="s")
qqnorm(accel_res, pch=20, frame=TRUE, col = "lightblue")
qqline(accel_res, col="darkblue",lwd = 1)

# Kurtosis and Skewness
kurtosis(accel_res)
skewness(accel_res)

# Sample Size
length(data_acc$ACCELERATION)

# Variance inflation factor
vif(accel_model)


# Tests for Normality
sw <- shapiro.test(accel_res)
ks <- lillie.test(accel_res)
cvm <- cvm.test(accel_res)
ad <- ad.test(accel_res)
```

```r
normtest <- data.frame(
Test = c("Shapiro-Wilk (W)", "Kolmogorov-Smirnov (D)", "Cramer-von
Mises (W-Sq)", "Anderson-Darling (A-Sq)"),
Statistic = c(sw$statistic, ks$statistic, cvm$statistic, ad$statistic),
P_Value = c(sw$p.value, ks$p.value, cvm$p.value, ad$p.value)
)
normtest <- normtest %>% mutate_if(is.numeric, round, digits = 5)

norm_gt <- gt(normtest) |> tab_header("Tests for Normality",) |>
cols_label(P_Value = "P-value") |> opt_stylize(style=2)

norm_gt |> gtsave("normtest.png")

#### CHI-SQUARED TEST ####
# VARAIBLES: CYLINDERS, ORIGIN
data_chi <- data[c(2,8)]
# Recoding the origin variable into what they represent
data_chi$ORIGIN[data_chi$ORIGIN == 1] <- "American"
data_chi$ORIGIN[data_chi$ORIGIN == 2] <- "European"
data_chi$ORIGIN[data_chi$ORIGIN == 3] <- "Japanese"


# Note that we drop 3 and 5 cylinder cars since there are so few and
they
# invalidate the chi-squared assumption of expected count >= 5 for all
cells
data_chi <- data_chi[data_chi$CYLINDERS %in% c(4,6,8),]

# Calling the test function
chi <- chisq.test(data_chi$CYLINDERS, data_chi$ORIGIN)
chi

# Making observations contingency table
obs_table <- chi$observed
obs_df <- as.data.frame.matrix(obs_table) |> rownames_to_column("CYL")

# Getting expected values
exp_table <- chi$expected
exp_df <- as.data.frame.matrix(exp_table) |> rownames_to_column("CYL")

# Getting Chi-Squared values for each cell
chi_table <- ((obs_table-exp_table)^2)/(exp_table)
chi_df <- as.data.frame.matrix(chi_table) |> rownames_to_column("CYL")

# Merging the tables
chitest_df <- merge_df(obs_df, exp_df, chi_df)
chitest_df <- chitest_df %>% mutate_if(is.numeric, round, digits = 2)

# Creating and exporting the final table
chitest_gt <- gt(chitest_df, rowname_col = "CYL") |>
tab_header(
title = "Chi-Squared Table for Car Origin by Number of Cylinders",
subtitle = "3 and 5 cylinder cars omitted due to infrequency") |>
tab_stubhead(label = "# of Cylinders") |>
tab_spanner(label = "Origin", columns = 1:4)
```

```
chitest_gt <- chitest_gt |> tab_row_group(
      label = "Chi-Squared", rows = 7:9) |> tab_row_group(
      label = "Expected", rows = 4:6) |> tab_row_group(
      label = "Observed", rows = 1:3)

chitest_gt <- chitest_gt |> opt_stylize(style = 2) |>
      tab_footnote("X-squared = 151.57, df = 4, p-value < 2.2e-16")

chitest_gt |> gtsave("chitest.png")
```