DS210 Final Project - Part I

Bracha Stein

2024-12-20

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

```
# Load in the data set
path_to_file <- "C:\\Users\\brach\\Downloads\\auto-mpg(1).csv"
auto_data = read.csv(path_to_file)

# Check the data's structure
str(auto_data)
```

```
## 'data.frame':    398 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinder    : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : chr  "130" "165" "150" "150" ...
##  $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ model.year  : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ car.name    : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel
```

```
# Change horsepower to numeric
auto_data$horsepower <- as.numeric(as.character(auto_data$horsepower))
```

```
## Warning: NAs introduced by coercion
```

```
# Check for missing values
any(is.na(auto_data))
```

```
## [1] TRUE
```

```
#remove rows with missing data
auto_data <- na.omit(auto_data)

# Splitting the data into first 300 rows
auto_data1 <- auto_data[1:300, ]
```

```r
# Simple Linear Regression
# weight as the independent variable
simple_model <- lm(mpg ~ weight, data = auto_data1)
summary(simple_model)
```

```
##
## Call:
## lm(formula = mpg ~ weight, data = auto_data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2011 -1.9157 -0.0812  1.7341 15.0246
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.5619792  0.6461532   62.77   <2e-16 ***
## weight      -0.0062905  0.0001984  -31.71   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.032 on 298 degrees of freedom
## Multiple R-squared:  0.7714, Adjusted R-squared:  0.7706
## F-statistic:  1005 on 1 and 298 DF,  p-value: < 2.2e-16
```

```r
b0_1 = simple_model$coefficients[1]
b1_1 = simple_model$coefficients[2]

# Multiple R-Squared = 0.7714
# Adjusted R-Squared = 0.7706
# Linear Regression Equation: y = 40.6 + -0.00629 * weight

# Multiple linear regression
# Using horsepower, weight and displacement as independent variables
multiple_model <- lm(mpg ~ horsepower + weight + displacement, data = auto_data1)
summary(multiple_model)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower + weight + displacement, data = auto_data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9508 -1.8780 -0.0657  1.6311 14.6386
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.5540648  0.9286288  42.594   <2e-16 ***
## horsepower   -0.0225670  0.0097080  -2.325   0.0208 *
## weight       -0.0048045  0.0005383  -8.925   <2e-16 ***
## displacement -0.0052516  0.0050006  -1.050   0.2945
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## Residual standard error: 2.982 on 296 degrees of freedom
## Multiple R-squared:  0.7804, Adjusted R-squared:  0.7782
## F-statistic: 350.6 on 3 and 296 DF,  p-value: < 2.2e-16
```

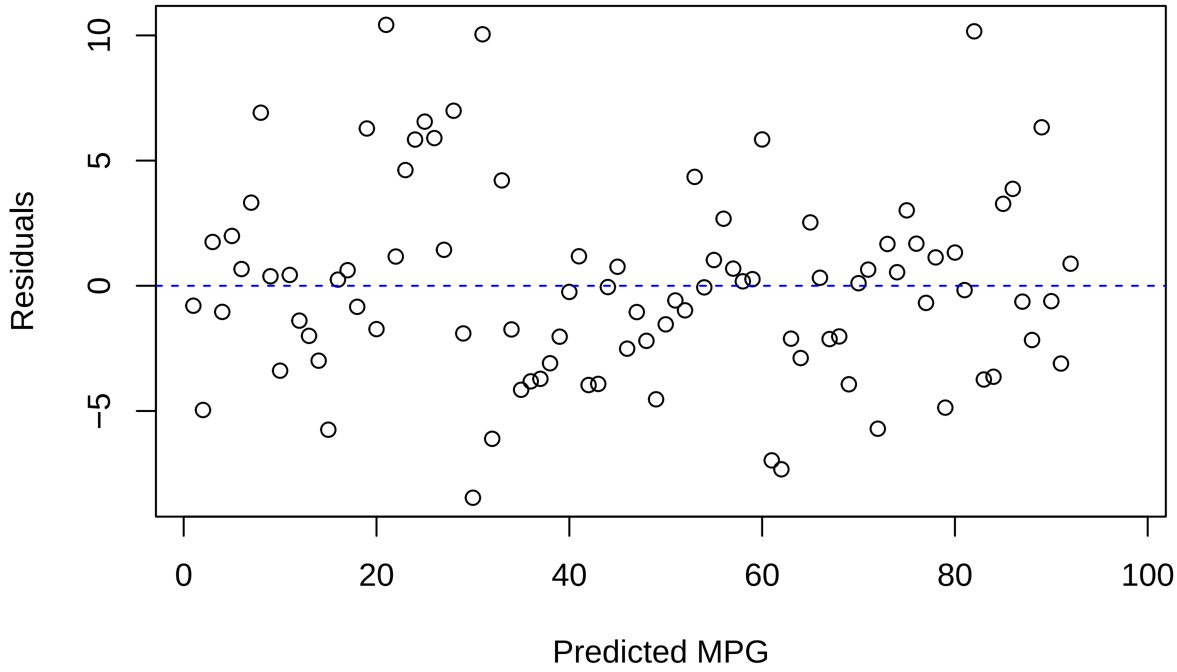
```
b0 <- multiple_model$coefficients[1]
b1 <- multiple_model$coefficients[2]
b2 <- multiple_model$coefficients[3]
b3 <- multiple_model$coefficients[4]

# Multiple R-Squared = 0.7804
# Adjusted R-Squared = 0.7782
# Multiple Linear Regression Equation: y = 39.6 + -0.0226 * horsepower + -0.0048 * weight + -0.00525 *

# Removing displacement as it is not statistically significant
modified_multiple_model <- lm(mpg ~ horsepower + weight, data = auto_data1)
summary(modified_multiple_model)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower + weight, data = auto_data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7069 -1.8380  0.0207  1.6877 14.5038
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.2587429  0.6420610  62.702  < 2e-16 ***
## horsepower  -0.0277594  0.0083560  -3.322  0.00101 **
## weight      -0.0052041  0.0003808 -13.666  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.982 on 297 degrees of freedom
## Multiple R-squared:  0.7796, Adjusted R-squared:  0.7781
## F-statistic: 525.2 on 2 and 297 DF,  p-value: < 2.2e-16
```

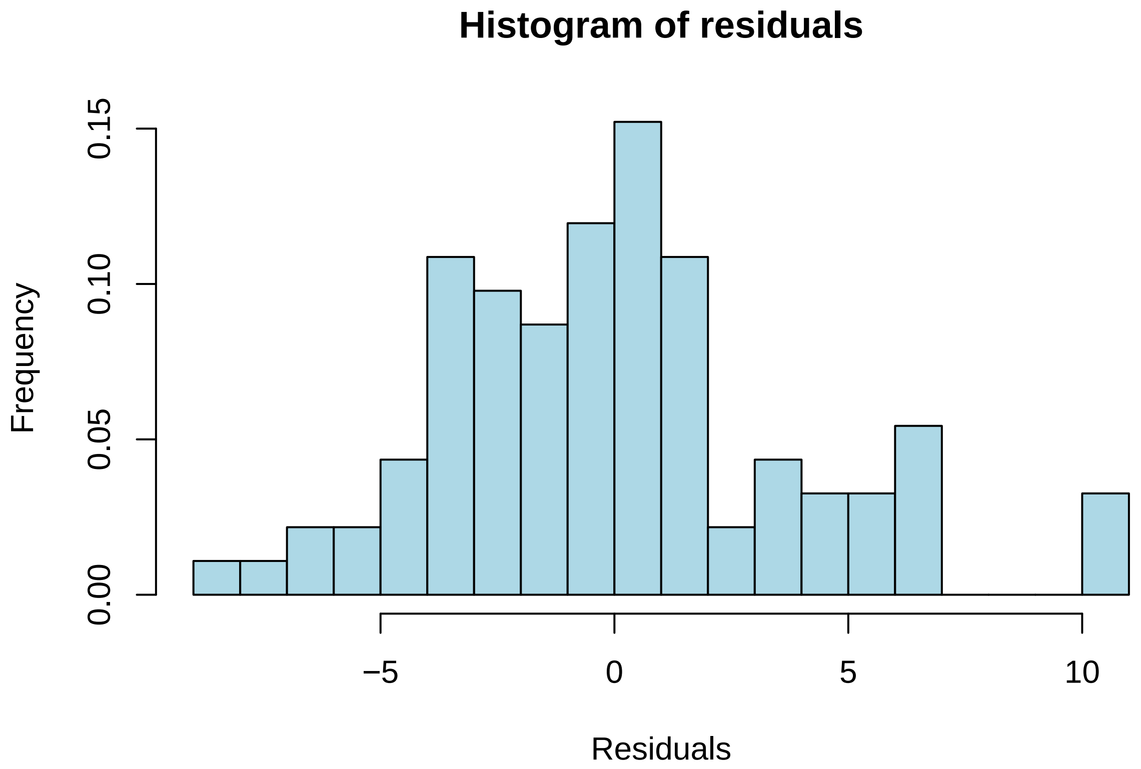
```
B0 <- modified_multiple_model$coefficients[1]
B1 <- modified_multiple_model$coefficients[2]
B2 <- modified_multiple_model$coefficients[3]

# Multiple R-Squared = 0.7796
# Adjusted R-Squared = 0.7781
# Multiple Linear Regression Equation: y = 40.3 + -0.0278 * horsepower + -0.0052 * weight

# Getting the last 98 samples from the dataset
auto_data2 <- auto_data[301:398,]

# Creating linear regression model
second_model <- lm(mpg ~ horsepower + weight, data = auto_data2)
summary(second_model)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower + weight, data = auto_data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4647 -2.2767 -0.1176  1.6726 10.4218
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 57.917423   2.438843  23.748  < 2e-16 ***
## horsepower  -0.124685   0.032347  -3.855 0.000219 ***
## weight      -0.006462   0.001274  -5.072 2.13e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.892 on 89 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.5721, Adjusted R-squared:  0.5625
## F-statistic: 59.49 on 2 and 89 DF,  p-value: < 2.2e-16
```

```r
BB0 <- second_model$coefficients[1]
BB1 <- second_model$coefficients[2]
BB2 <- second_model$coefficients[3]

# Multiple R-Squared = 0.5721
# Adjusted R-Squared = 0.5625
# Linear Regression Equation: y = 57.9 + -0.125 * horsepower + -0.00646 * weight

# Predicting mpg using second_model
mpg_predict <- predict(second_model, auto_data2)

# Comparing to real mpg
real_mpg <- auto_data2$mpg

residuals <- real_mpg - mpg_predict

# Residual Plot
plot(residuals, xlab="Predicted MPG", ylab="Residuals")
abline(0,0 ,col='blue', lty = 2)
```



```
# Histogram
hist(residuals, prob=T, breaks=20, xlab="Residuals", ylab="Frequency", col="light blue")
```


Histogram of residuals
Frequency
0.00
0.05
0.10
0.15
−5
0
5
10
Residuals