HW5

2024-05-26

Problem 1

A. For categorical variable, we apply Kruskal-Wallis test. For numerical variable, we apply the correlation test.

```
dataset <- read.csv("cars.csv")

numerical <- c("Dimensions.Height",
               "Dimensions.Length",
               "Dimensions.Width",
               "Engine.Information.Number.of.Forward.Gears",
               "Fuel.Information.Highway.mpg",
               "Identification.Year",
               "Engine.Information.Engine.Statistics.Horsepower",
               "Engine.Information.Engine.Statistics.Torque")

categorical <- c("Engine.Information.Driveline",
                 "Engine.Information.Transmission",
                 "Fuel.Information.Fuel.Type",
                 "Identification.Make",
                 "Identification.Classification")

p_num_cate <- list()
for (x in numerical){
  numerical_predictor  <- dataset[[x]]
  mpg <- dataset[["Fuel.Information.City.mpg"]]
  test_result <- cor.test(mpg, numerical_predictor )
  p_num_cate[x] <- test_result$p.value
  cat("Notion of correlation between Fuel Information.City mpg and", x, "\n")
  print(test_result)
}
```

```
## Notion of correlation between Fuel Information.City mpg and Dimensions.Height
##
##  Pearson's product-moment correlation
##
## data:  mpg and numerical_predictor
## t = 18.344, df = 5074, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2234071 0.2750095
```

```
## sample estimates:
##       cor
## 0.2493853
##
## Notion of correlation between Fuel Information.City mpg and Dimensions.Length
##
##  Pearson's product-moment correlation
##
## data:  mpg and numerical_predictor
## t = -1.3433, df = 5074, p-value = 0.1792
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.046341079  0.008661337
## sample estimates:
##         cor
## -0.01885414
##
## Notion of correlation between Fuel Information.City mpg and Dimensions.Width
##
##  Pearson's product-moment correlation
##
## data:  mpg and numerical_predictor
## t = -10.072, df = 5074, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1668662 -0.1129219
## sample estimates:
##       cor
## -0.139998
##
## Notion of correlation between Fuel Information.City mpg and Engine.Information.Num
ber.of.Forward.Gears
##
##  Pearson's product-moment correlation
##
## data:  mpg and numerical_predictor
## t = -2.5908, df = 5074, p-value = 0.009602
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.063794771 -0.008845448
## sample estimates:
##         cor
## -0.03634758
##
## Notion of correlation between Fuel Information.City mpg and Fuel.Information.Highw
ay.mpg
##
```

```
##  Pearson's product-moment correlation
##
## data:  mpg and numerical_predictor
## t = 123.15, df = 5074, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8585519 0.8723541
## sample estimates:
##       cor
## 0.8656173
##
## Notion of correlation between Fuel Information.City mpg and Identification.Year
##
##  Pearson's product-moment correlation
##
## data:  mpg and numerical_predictor
## t = 6.5846, df = 5074, p-value = 5.022e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.06469912 0.11925526
## sample estimates:
##        cor
## 0.09204627
##
## Notion of correlation between Fuel Information.City mpg and Engine.Information.Eng
ine.Statistics.Horsepower
##
##  Pearson's product-moment correlation
##
## data:  mpg and numerical_predictor
## t = -70.123, df = 5074, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.7152438 -0.6872908
## sample estimates:
##       cor
## -0.701537
##
## Notion of correlation between Fuel Information.City mpg and Engine.Information.Eng
ine.Statistics.Torque
##
##  Pearson's product-moment correlation
##
## data:  mpg and numerical_predictor
## t = -81.932, df = 5074, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  -0.7662659 -0.7425697
## sample estimates:
##        cor
## -0.7546638
```

```
# For category data:
for (x in categorical){
  categorical_predictor  <- dataset[[x]]
  mpg <- dataset[["Fuel.Information.City.mpg"]]
  test_result <- kruskal.test(mpg ~ categorical_predictor)
  p_num_cate[x] <- test_result$p.value
  cat("Association between Fuel Information.City mpg and", x, "\n")
  print(test_result)
}
```

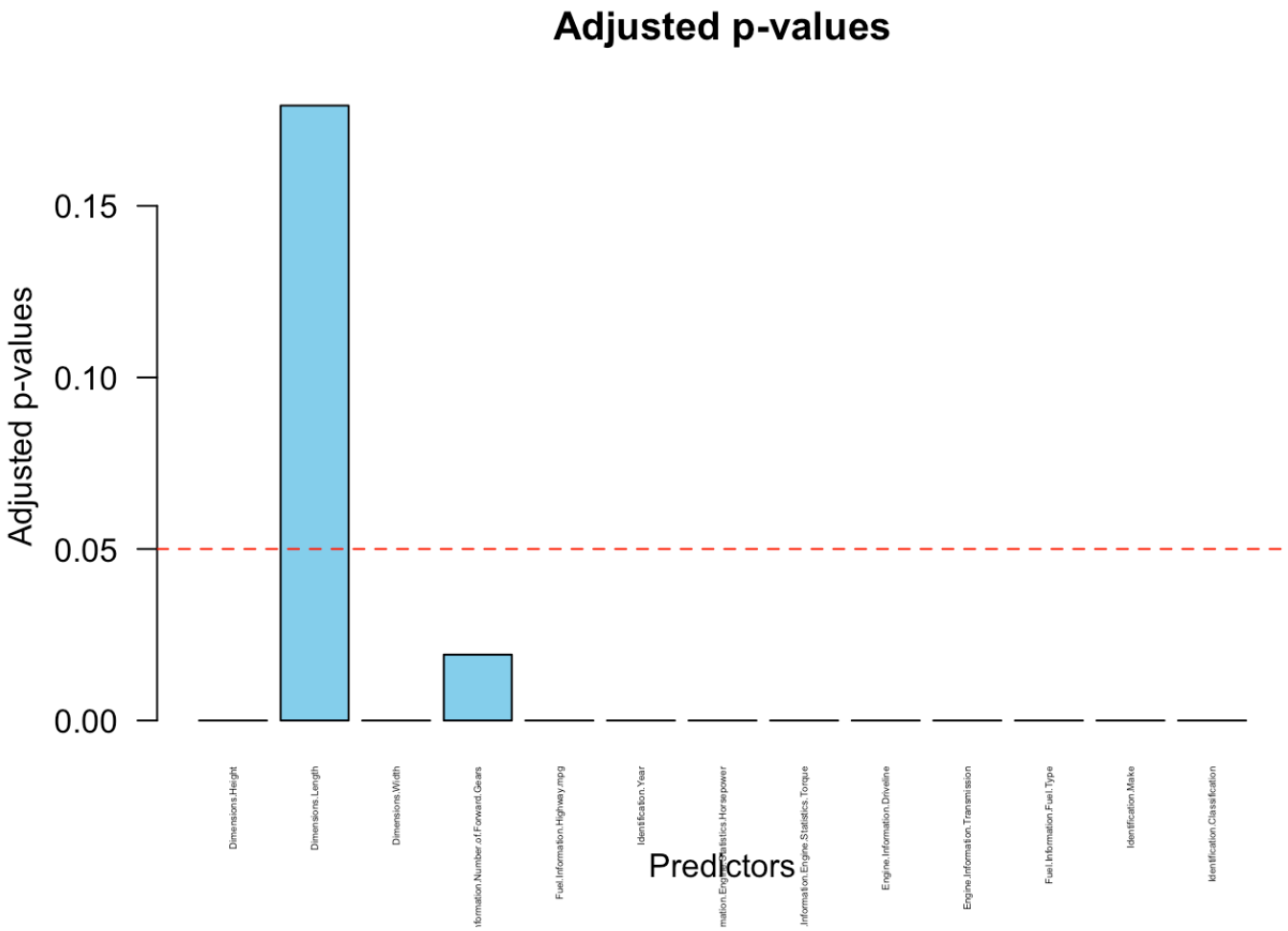
```
## Association between Fuel Information.City mpg and Engine.Information.Driveline
##
##  Kruskal-Wallis rank sum test
##
## data:  mpg by categorical_predictor
## Kruskal-Wallis chi-squared = 2216.4, df = 3, p-value < 2.2e-16
##
## Association between Fuel Information.City mpg and Engine.Information.Transmission
##
##  Kruskal-Wallis rank sum test
##
## data:  mpg by categorical_predictor
## Kruskal-Wallis chi-squared = 958.04, df = 10, p-value < 2.2e-16
##
## Association between Fuel Information.City mpg and Fuel.Information.Fuel.Type
##
##  Kruskal-Wallis rank sum test
##
## data:  mpg by categorical_predictor
## Kruskal-Wallis chi-squared = 1091.3, df = 3, p-value < 2.2e-16
##
## Association between Fuel Information.City mpg and Identification.Make
##
##  Kruskal-Wallis rank sum test
##
## data:  mpg by categorical_predictor
## Kruskal-Wallis chi-squared = 1952.9, df = 46, p-value < 2.2e-16
##
## Association between Fuel Information.City mpg and Identification.Classification
##
##  Kruskal-Wallis rank sum test
##
## data:  mpg by categorical_predictor
## Kruskal-Wallis chi-squared = 565.64, df = 1, p-value < 2.2e-16
```

Problem 1 B:

```
adjusted <- p.adjust(p_num_cate, method = "holm")
list_name <- list()
for (x in numerical){
  list_name[x] <- x
}
for(x in categorical){
  list_name[x] <- x
}
print(length(list_name))
```

```
## [1] 13
```

```
# Plot adjusted p-values
barplot(adjusted, names.arg = list_name,
        xlab = "Predictors", ylab = "Adjusted p-values",
        col = "skyblue", main = "Adjusted p-values",
        las = 2,
        cex.names = 0.3)
abline(h = 0.05, col = "red", lty = 2)
```

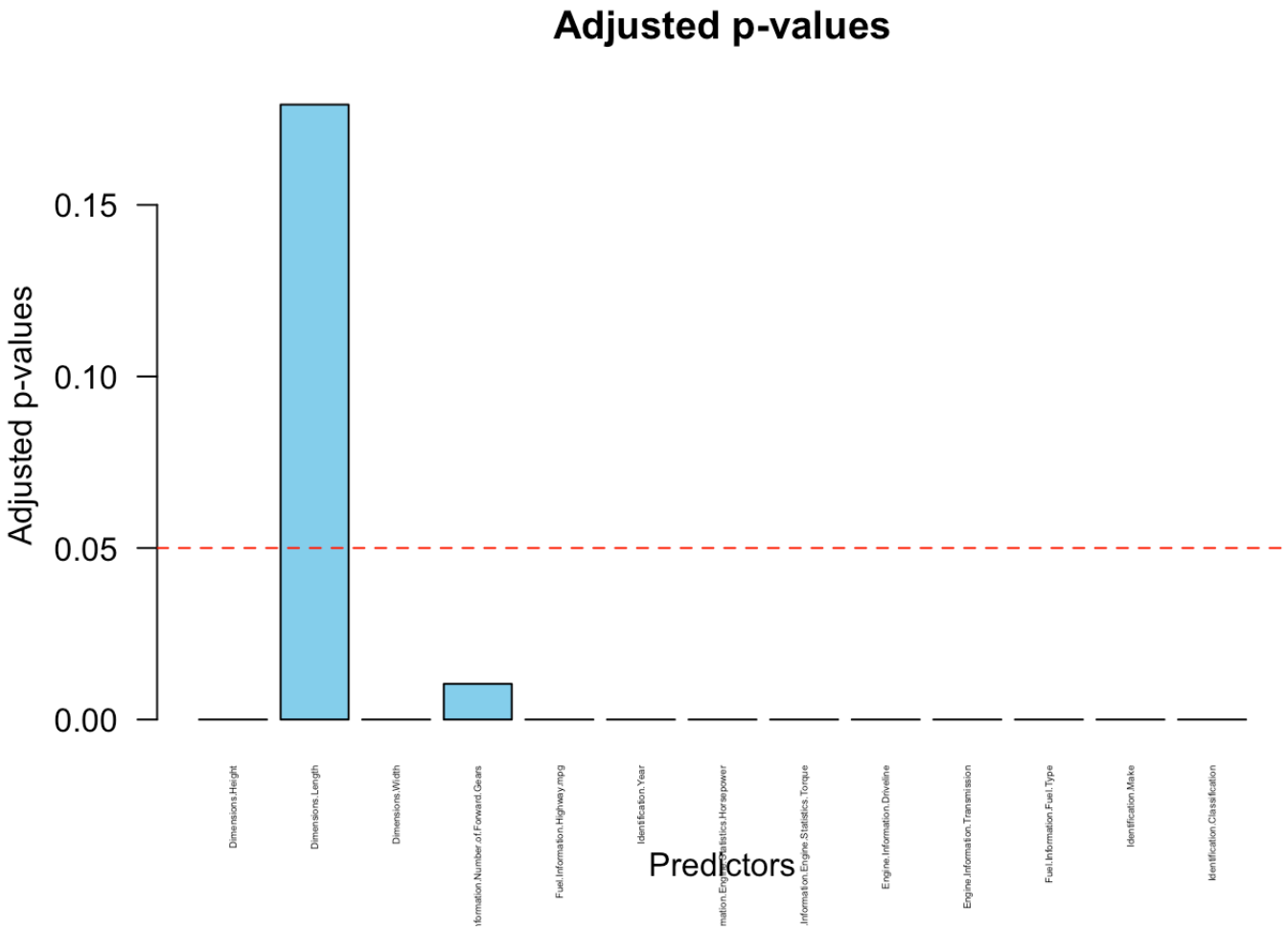


B. From above plot, we get to see that most of the predictor are likely associated with the fuel consumption in the city except for the Dimensions length as it has a much higher p value compare to other predictor.

```
# C
adjusted <- p.adjust(p_num_cate, method = "fdr")
list_name <- list()
for (x in numerical){
  list_name[x] <- x
}
for(x in categorical){
  list_name[x] <- x
}
print(length(list_name))
```

```
## [1] 13
```

```
# Plot adjusted p-values
barplot(adjusted, names.arg = list_name,
        xlab = "Predictors", ylab = "Adjusted p-values",
        col = "skyblue", main = "Adjusted p-values",
        las = 2,
        cex.names = 0.3)
abline(h = 0.05, col = "red", lty = 2)
```



C. From above, this adjust method also get the same result from part B. It show us that every predictor expect the Dimensions length are likely associated with the fuel consumption in the city.

Problem 2

A. We can apply Pearson's correlation test like we did in Q1

```
highway <- dataset[["Fuel.Information.Highway.mpg"]]
cor.test(mpg, highway)
```

```
##
##  Pearson's product-moment correlation
##
## data:  mpg and highway
## t = 123.15, df = 5074, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8585519 0.8723541
## sample estimates:
##       cor
## 0.8656173
```

From here, we can see that the there is a high correlation about 0.85, which indicate there is likely a linear relationship between the fuel consumption in the city and fuel consumption on the highway.

B

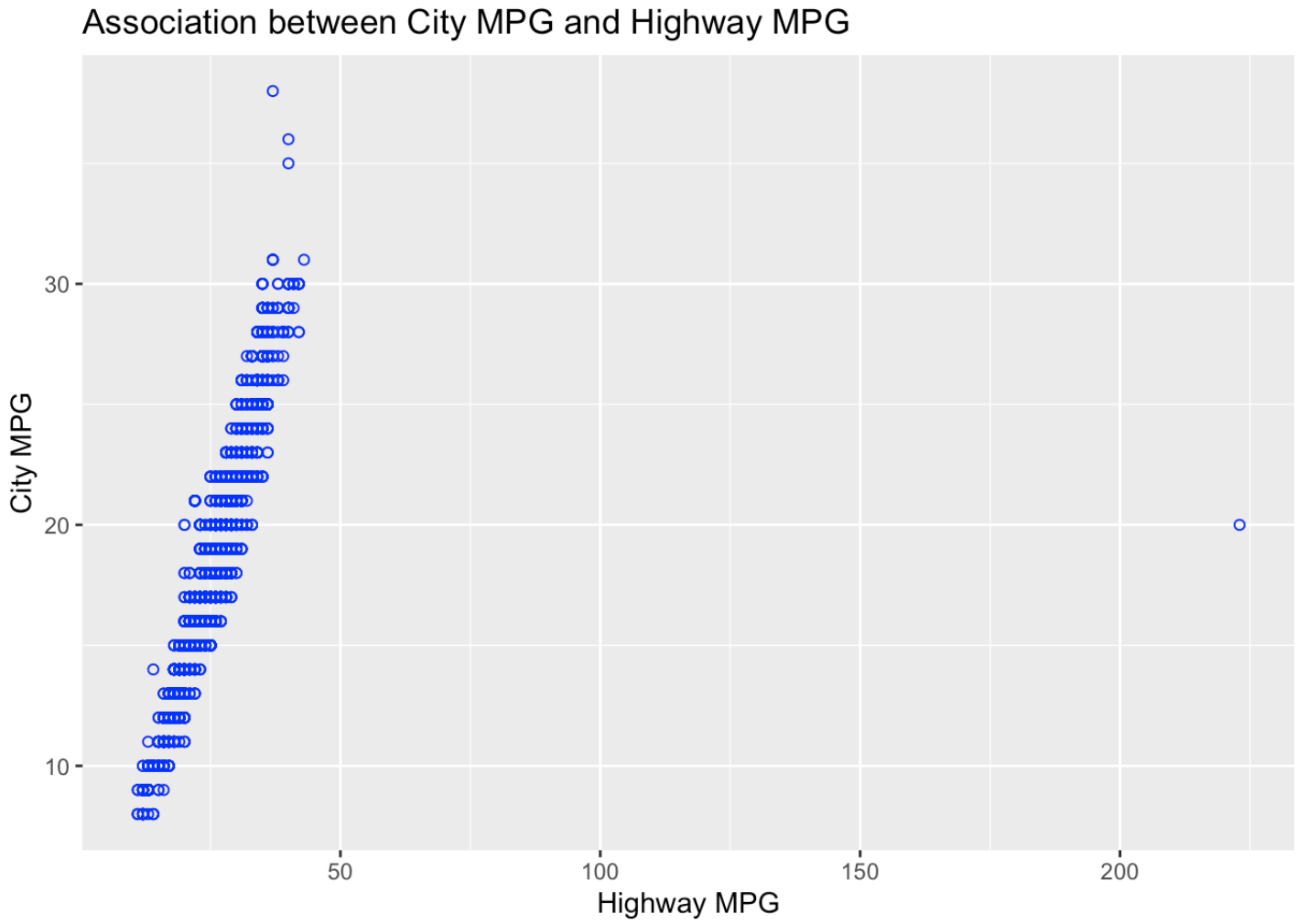
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked _by_ '.GlobalEnv':
##
##     mpg
```

```
ggplot(dataset, aes(x = highway, y = mpg)) +
  geom_point(color = "blue", shape = 1) +
  labs(x = "Highway MPG", y = "City MPG", title = "Association between City MPG and H
ighway MPG")
```



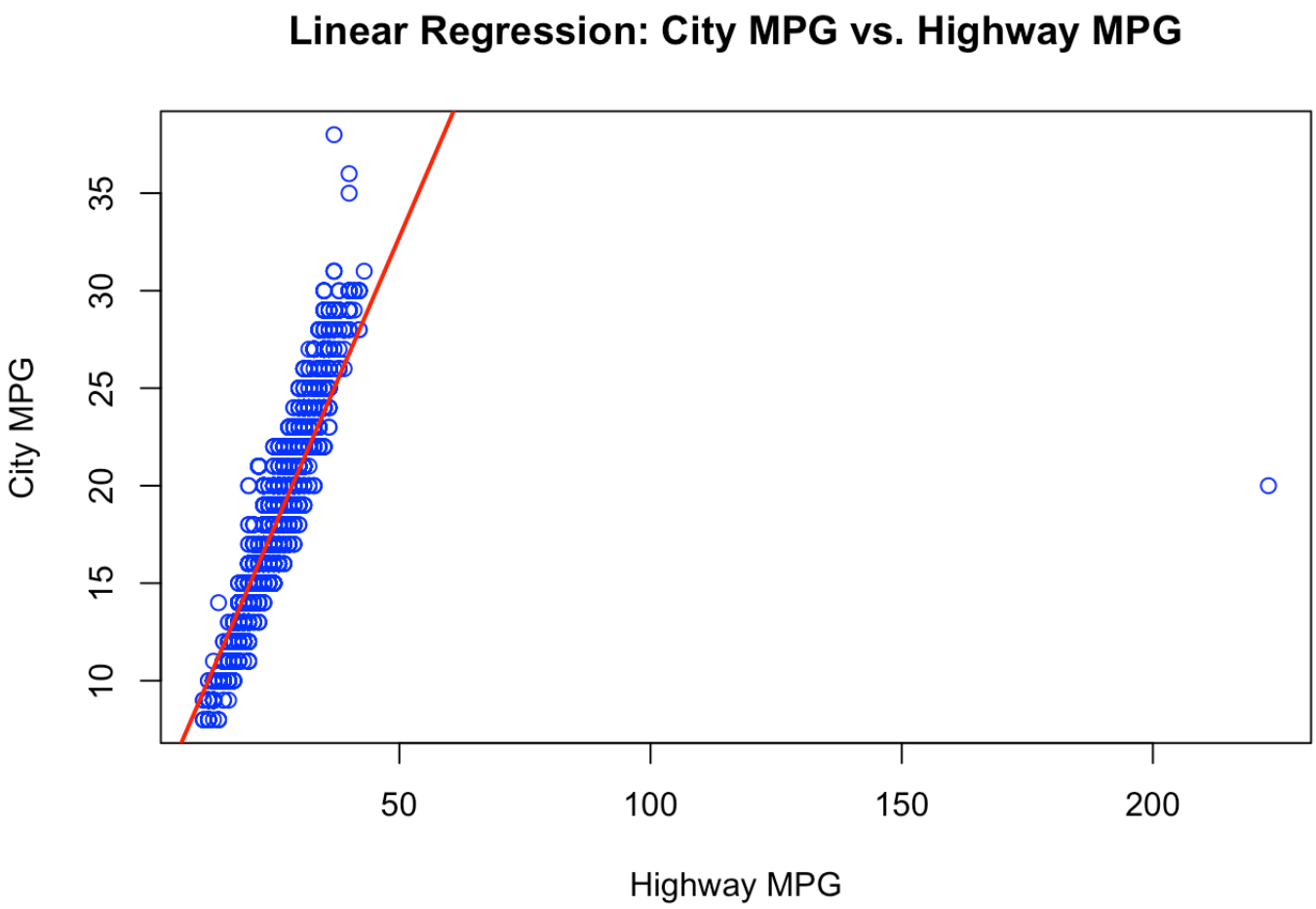
C

```
# We fit a line by least squares:
fit <- lm(mpg ~ highway, data = dataset)
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ highway, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -116.127   -0.994   -0.201    0.787   13.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.857981   0.121239   23.57   <2e-16 ***
## highway     0.597618   0.004853  123.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.243 on 5074 degrees of freedom
## Multiple R-squared:  0.7493, Adjusted R-squared:  0.7492
## F-statistic: 1.516e+04 on 1 and 5074 DF,  p-value: < 2.2e-16
```

D

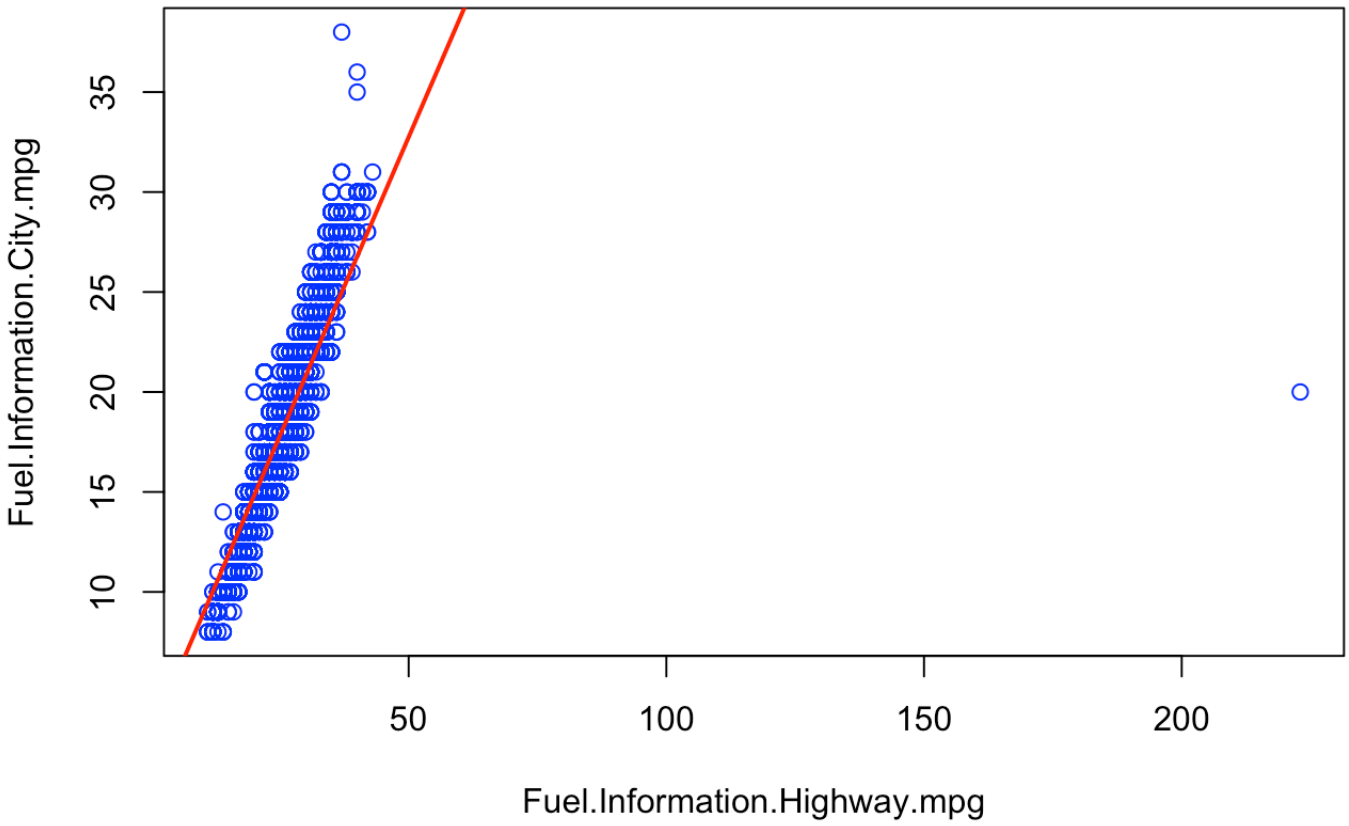
```
# D
plot(highway, mpg,
     xlab = "Highway MPG", ylab = "City MPG",
     main = "Linear Regression: City MPG vs. Highway MPG",
     col = "blue", pch = 1)
abline(fit, col = "red", lwd = 2)
```


Linear Regression: City MPG vs. Highway MPG
City MPG
10
15
20
25
30
35
50
100
150
200
Highway MPG

Problem 3

```
paired_data_associated_analysis <- function(data) {

  # get name
  variable_names <- colnames(data)

  # Perform correlation analysis
  correlation <- cor(data[, 1], data[, 2])
  cat("Pearson correlation result between: ", variable_names,
      correlation, "\n\n")

  # Fit function
  fit <- lm(data[, 2] ~ data[, 1], data = data)
  print(summary(fit))

  plot(data[,1], data[,2],
     xlab = variable_names[1], ylab = variable_names[2],
     col = "blue", pch = 1)
abline(fit, col = "red", lwd = 2)
}

paired_data_associated_analysis(dataset[, c("Fuel.Information.Highway.mpg",
                                  "Fuel.Information.City.mpg")])
```

```
## Pearson correlation result between:  Fuel.Information.Highway.mpg Fuel.Information
.City.mpg 0.8656173
##
##
## Call:
## lm(formula = data[, 2] ~ data[, 1], data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -116.127   -0.994   -0.201    0.787   13.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.857981   0.121239   23.57   <2e-16 ***
## data[, 1]   0.597618   0.004853  123.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.243 on 5074 degrees of freedom
## Multiple R-squared:  0.7493, Adjusted R-squared:  0.7492
## F-statistic: 1.516e+04 on 1 and 5074 DF,  p-value: < 2.2e-16
```


Fuel.Information.City.mpg
10
15
20
25
30
35
50
100
150
200
Fuel.Information.Highway.mpg