HW 3

Jared Andreatta

2025-03-31

Problem 1

The null hypotheses for the p-values is that the leading coefficient for each variable is 0. Formally, we can write

$$H_0 : \beta_i = 0$$

for $i = 1, 2, 3, 4$. What we can inference off of this is how likely that the coefficient of each variable is statistically significant (most likely not equal to 0). The **sales, TV**, and **radio** all have negligibly small p-values, meaning that there is a very low likelihood of their corresponding β is 0, meaning that we can reject the null hypothesis with >99% confidence. The **newspaper** variable, however, has a p-value of 0.8599, meaning that the β corresponding to **newspaper** is likely 0, therefore we fail to reject the null hypothesis that the coefficient estimate is 0.

Problem 7

NOTE: I will use common linear algebra notation (as seen in MATH531) as it is easier for me to understand and compute this way.

Proof. It is given that this regression is centered around a mean of 0, i.e. $\bar{x} = \bar{y} = 0$. First, we define the squared correlation coefficient between x and y. In the case of simple linear regression, we have

$$\text{Cor}(x, y)^2 = \left(\frac{x^T y}{\sqrt{x^T x}\sqrt{y^T y}} \right)^2 = \frac{(x^T y)^2}{x^T x y^T y}$$

Now, we can algebraically simplify the expression for R^2 to show that it is simply the squared correlation coefficient. We know that the expression for R^2 is

$$R^2 = 1 - \frac{RSS}{TSS}$$

which, in linear algebra notation, can be simply written as

$$R^2 = 1 - \left(\frac{(y - x\hat{\beta})^T (y - x\hat{\beta})}{y^T y} \right)$$

We can expand the inner product term in the numerator.

$$
\begin{aligned}
R^2 &= 1 - \left(\frac{y^t y - y^T x\hat{\beta} - (x\hat{\beta})^T y + (x\hat{\beta})^T x\hat{\beta}}{y^T y} \right) \\
&= 1 - \left(\frac{y^T y}{y^T y} - \frac{-y^T x\hat{\beta} - (x\hat{\beta})^T y + \hat{\beta}^T x^T x\hat{\beta}}{y^T y} \right) \\
&= \frac{y^T x\hat{\beta} + (x\hat{\beta})^T y - \hat{\beta}^T x^T x\hat{\beta}}{y^T y}
\end{aligned}
$$

Note that $y^T x\hat{\beta} = (x\hat{\beta})^T y$. They are scalar values, so they are equal to their own transpose. Additionally, we note that since this is the case of simple linear regression, $\hat{\beta}$ is simply a scalar value. Therefore, we have

$$
R^2 = \frac{2\hat{\beta} x^T y - \hat{\beta}^2 x^T x}{y^T y}
$$

Recall the definition of the OLS estimator (i.e. $\hat{\beta}$):

$$
\hat{\beta} = \frac{x^T y}{x^T x}
$$

Using this definition, we can substitute this back into the expression for R^2.

$$
\begin{aligned}
R^2 &= \frac{2\left(\frac{x^T y}{x^T x}\right) x^T y - \left(\frac{x^T y}{x^T x}\right)^2 x^T x}{y^T y} \\
&= \frac{2\frac{(x^T y)^2}{x^T x} - \frac{(x^T y)^2}{x^T x}}{y^T y} \\
&= \frac{\frac{(x^T y)^2}{x^T x}}{y^T y} \\
&= \frac{(x^T y)^2}{x^T x y^T y}
\end{aligned}
$$

Hence, we have shown that

$$
R^2 = \frac{(x^T y)^2}{x^T x y^T y} = \mathrm{Cor}(x, y)^2
$$

□.

Problem 8

a.

i.

There is a statistically significant relationship between the predictor and the response.

ii.

The horsepower coefficient estimate is -0.157845, meaning that, on average, for every increase in 1 HP, we expect a .157845 decrease in MPG.

iii.

The relationship is negative; as horsepower increases, we expect decreases in MPG.

iv.

The estimated MPG of a car with 98 horsepower is 24.4670. The 95% confidence interval is [24.46708 23.97308 24.96108] and the 95% prediction interval is [24.46708 14.8094 34.12476].

```
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.4.3
```

```
data <- Auto

fit <- lm(mpg ~ horsepower, data=data)

summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
#iv.
est <- coef(fit)[1] + 98 * coef(fit)[2] # Estimating for 98 HP
print(est)
```

```
## (Intercept)
##    24.46708
```

```r
predict(fit, newdata = data.frame(horsepower = 98), interval="prediction") # Prediction intervals
```

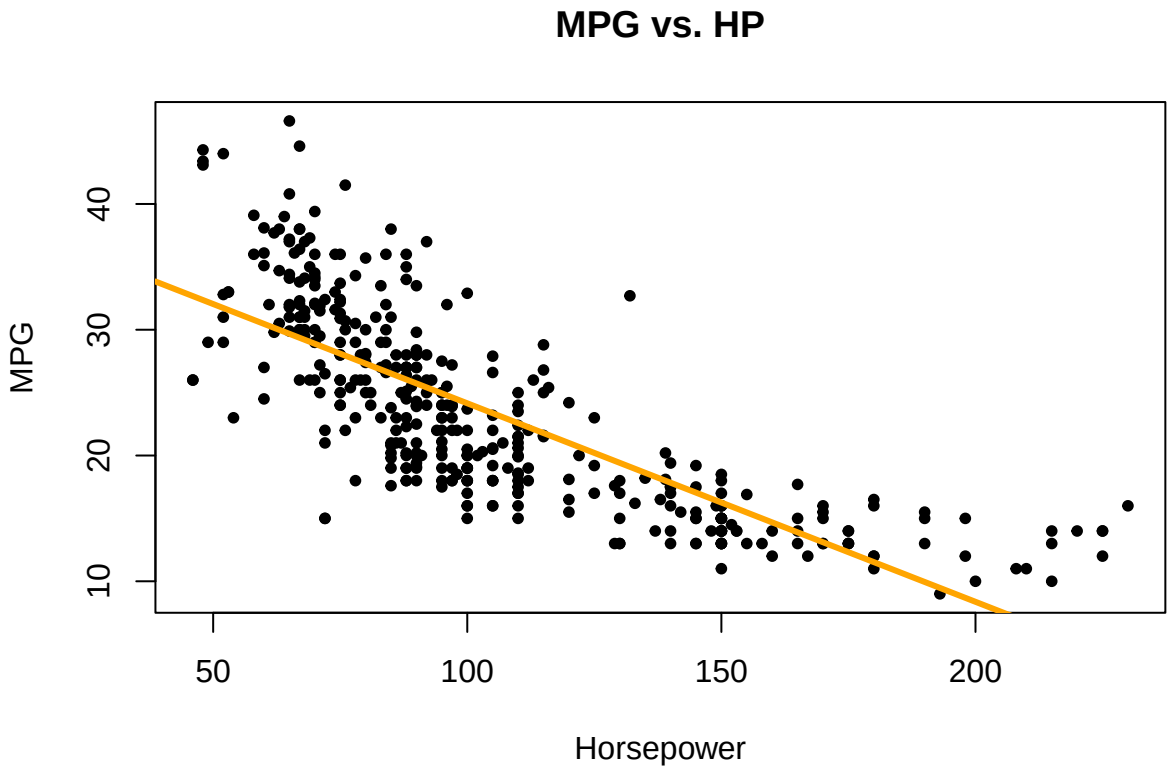
```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

```r
predict(fit, newdata = data.frame(horsepower = 98), interval="confidence") # Confidence intervals
```

```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

b.

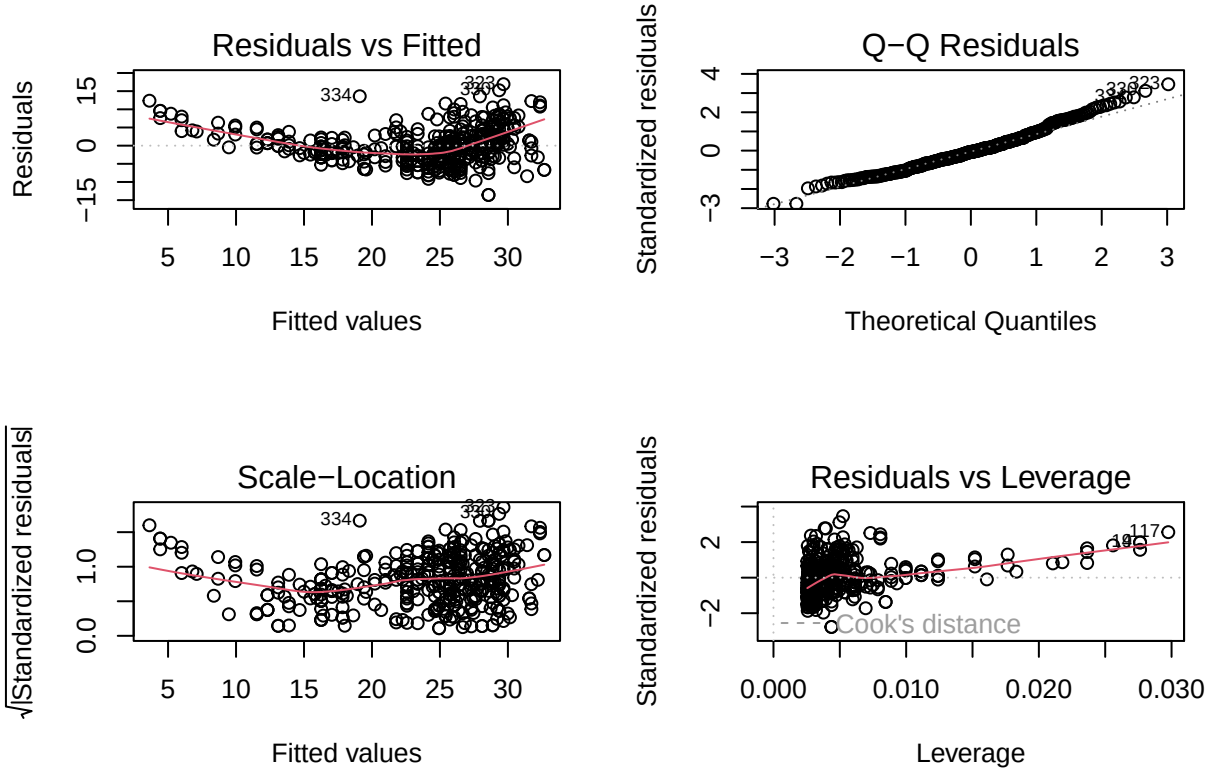
```r
plot(data$horsepower, data$mpg, pch=20, xlab="Horsepower", ylab="MPG", main= "MPG vs. HP")
abline(coef=coef(fit), lwd=3, col="orange")
```



c.

A problem easily noticeable, both from the diagnostic plots and the regression plot, is that the linear specification does not quite provide the best estimate of the data. Nonlinearity is inferred from the shape of the residuals vs fitted scatterplot, and also the general shape of the scatterplot of MPG vs. HP, so a possible improvement of the model could be to add a quadratic term to better capture that nonlinearity.

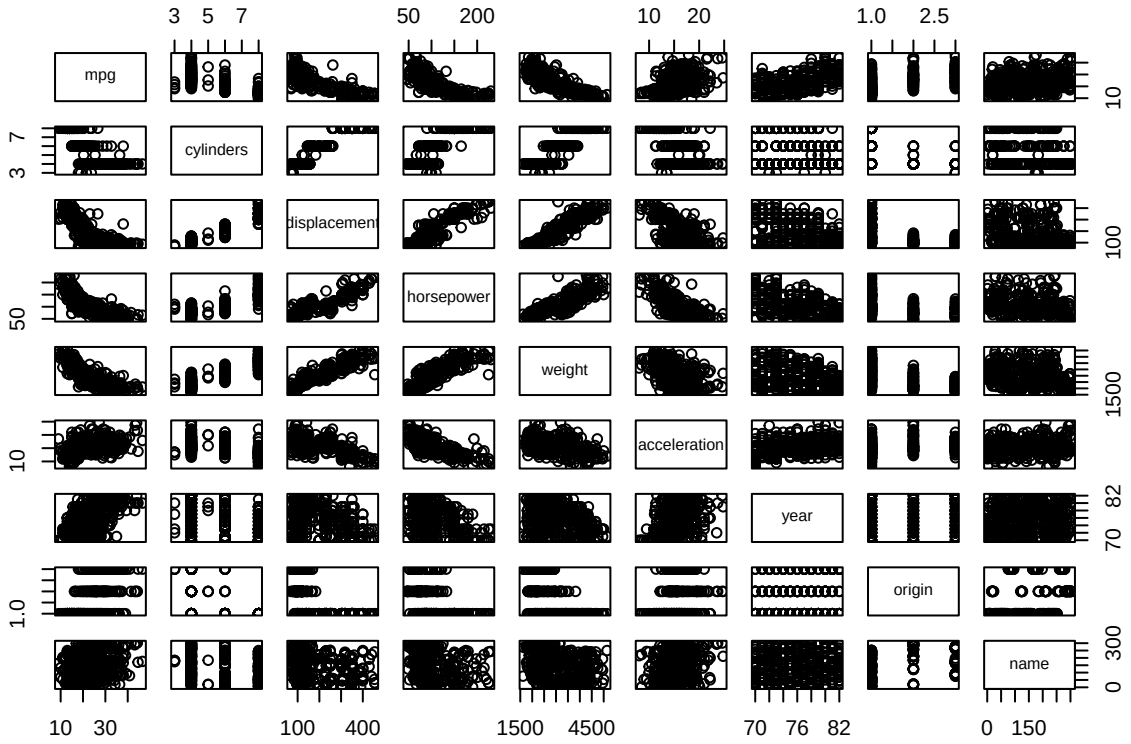
```
par(mfrow=c(2,2))
plot(fit)
```



Problem 9

a.

```
pairs(data)
```



b.

```
names(data)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
## [6] "acceleration" "year"         "origin"       "name"
```

```
cor(data[1:8])
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
```

```
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

c.

i.

Yes, there is a statistically significant relationship between the response and the predictors.

ii.

The predictors that have a statistically significant relationship are **displacement, weight, year,** and **origin** (and the **Intercept**).

iii.

The year coefficient suggests that for every increase in year, we expect a .751 increase in mpg of the car on average. From this, we can infer that newer cars tend to have better gas mileage than older cars.

```
fit <- lm(mpg ~ .-name, data=data)

summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

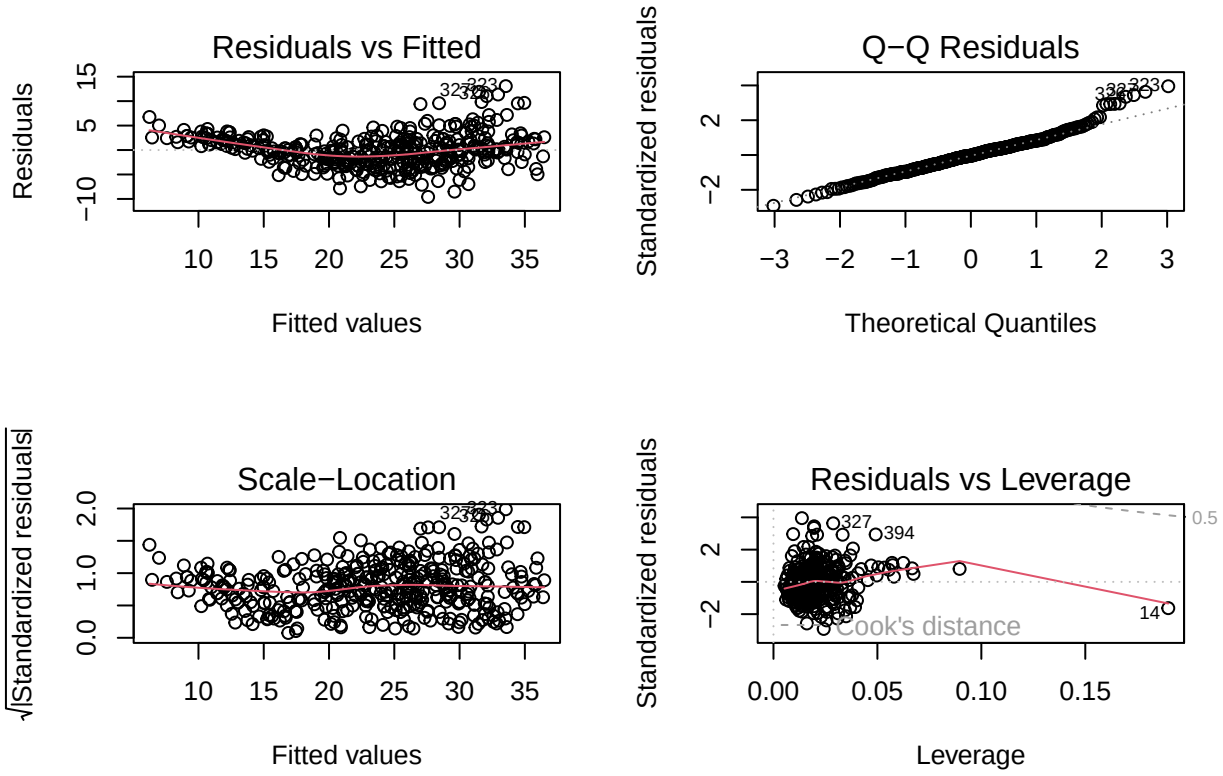
d.

The Q-Q plot mostly follows along the line trend up until the second quantile, where the variance of the standardized residuals start to increase, which can also be seen in the residuals vs fitted. This indicates that the model provides a weaker fit for cars with high mpg.

There doesn't seem to be any extreme outliers, though it does seem that the standard errors seem to be heteroskeadastic.

From the leverage plot, there is clear evidence of an observation with extreme leverage.

```
par(mfrow=c(2,2))
plot(fit)
```



Problem 12

a.

The coefficient estimate $\hat{\beta}$ is equal when we regress X onto Y as when we regress Y onto X when $\|X\|^2 = \|Y\|^2$. We can use equation (3.38) to show this. Note that I will use linear algebra notation, as it is easier to read (in my opinion). The estimate for β when we regress X onto Y is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Equivalently, when we regress Y onto X, the estimate is

$$\hat{\beta} = (Y^T Y)^{-1} X^T Y$$

Assuming that $X^T Y \neq 0$, we set these equal to each other and find

$$(Y^T Y)^{-1} X^T Y = (X^T X)^{-1} X^T Y \implies$$
$$Y^T Y = X^T X$$

b.

```
# Simplest case. Both X and Y are 100x1 random vectors with mean 0 and variance sigma^2. X!=Y
set.seed(25)

X <- rnorm(100)
Y <- rnorm(100)

fitX <- lm(Y ~ X+0)
fitY <- lm(X ~ Y+0)

coef(fitX)
```

```
##           X
## 0.05889304
```

```
coef(fitY)
```

```
##           Y
## 0.07358347
```

c.

```
# Simplest Case: X is a randon 100x1 vector with mean 0 and variance sigma^2. In this case, Y=X.
set.seed(100)

X <- rnorm(100)
Y <- X

fitX <- lm(Y ~ X+0)
fitY <- lm(X ~ Y+0)

coef(fitX)
```

```
## X
## 1
```

```r
coef(fitY)
```

```
## Y
## 1
```