ISLR Chapter 3 Exercises

2023-07-01

Contents


```
library(MASS)
library(ggplot2)
library(GGally)
library(dplyr)
library(gridExtra)
library(ISLR)
library(ggfortify)
```

Conceptual

Question 1

The null hypothesis for rows 2-4 is that there is no linear relationship between the media type in the leftmost column and sales. The null hypothesis for the first row (the intercept) is that the mean response is 0 when all media spending is 0. In terms of coefficients, for each row i, indexed starting at 0, the null hypothesis is that $\beta_i = 0$.

Question 2

The KNN regression method uses the average response of the `k` neighboring points closest to a test point as the prediction for that point. The KNN classification method first computes the conditional probability that a test point is of each class by computing the fraction of the `k` neighboring points that belong to that class, and then assigns the test point to the class with the maximum conditional probability.

Question 3

a Answer iii) is correct. The main effect of gender increases female salary relative to male salary by `35,000`, but the interaction effect decreases female salary by `10,000` per point of `GPA`. So if a female had a `GPA` of `4.0`, on average they would have a salary that differs from a male by $35,000 + -10,000 \cdot 4 = -5,000$, provided that the female and male both had the same `IQ` and `GPA`. Since this is a negative number, so we conclude that on average males earn more provided that the GPA is high enough.

b $50 + 20 \cdot 4 + 0.07 \cdot 110 + 35 + 0.01 \cdot 4 \cdot 110 - 10 \cdot 4 \cdot 1 = 137.1$, in thousands of dollars.

c This is false. The t value and the corresponding p-value for the interaction term needs to be examined, rather than the coefficient value itself. In general, the coefficient value will depend on the scaling of the variable rather than the significance of the term.

Question 4

a The training `RSS` will never increase as more variables are added to the model. This includes transformations of existing variables. So we would expect the cubic regression to have a lower training `RSS`.

b Given that the true relationship is linear, the cubic model will not result in a reduction of bias but will increase variance. So we would expect the linear regression to have a lower test error.

c See part a) above.

d Given that we don't know how far from linear the true relationship is, it is unclear how much bias will be reduced by using the cubic regression. So we can't tell which method would have lower test `RSS`.

Question 5

$$\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{j=1}^{n} x_j^2} = \sum_{i=1}^{n} \frac{x_i}{\sum_{j=1}^{n} x_j^2} y_i.$$

So $\alpha_i = \dfrac{x_i}{\sum_{j=1}^{n} x_j^2}$

Question 6

$\hat{y} = \beta_0 + \beta_1 x = \bar{y} - \beta_1 \bar{x} + \beta_1 x = \bar{y} + \beta_1 (x - \bar{x})$

This equality holds when $\hat{y} = \bar{y}$ and $x = \bar{x}$.

Question 7

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{Cov(X,Y)}{Var(X)} = \frac{Cor(X,Y)SD(Y)}{SD(X)}$$

Now $R^2 = \dfrac{\sum_{i=1}^{n} (\hat{y_i} - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \dfrac{\sum_{i=1}^{n} (\bar{y} + \beta_1 (x_i - \bar{x}) - \bar{y})^2}{SD(Y)^2} = \dfrac{\sum_{i=1}^{n} (\beta_1 (x_i - \bar{x}))^2}{SD(Y)^2} = \beta_1^2 \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{SD(Y)^2} =$

$$\frac{Cor(X,Y)^2 SD(Y)^2}{SD(X)^2} \cdot \frac{SD(X)^2}{SD(Y)^2} = Cor(X,Y)^2$$

Applied

Question 8

a

```
df_auto <- Auto
simple_linear_regression_model <- lm(mpg ~ horsepower, data = df_auto)

print(summary(simple_linear_regression_model))
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = df_auto)
##
## 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
```

i Since this is simple linear regression, the t-statistic and the F-statistic are testing the same thing: is there a relationship between the predictor and the response? Since the p-value associated with these statistics is tiny, we reject the null hypothesis that there is no linear association between `mpg` and `horsepower`.

ii This model has an R^2 of 0.61, indicating that 61% of the variance in the training data is explained by the model. The estimate of the standard error is `4.9` units, relative to the mean value of `23.4`, giving a relative error of roughly 20%.

iii Looking at the sign of the coefficient, the relationship is negative.

iv

```
dfs_interval <- lapply(
  c("confidence", "prediction"),
  function(interval, model) {
    df <- data.frame(predict(model, data.frame(horsepower = 98), interval = interval))
    df$interval <- interval

    df
  },
  model = simple_linear_regression_model
)
```

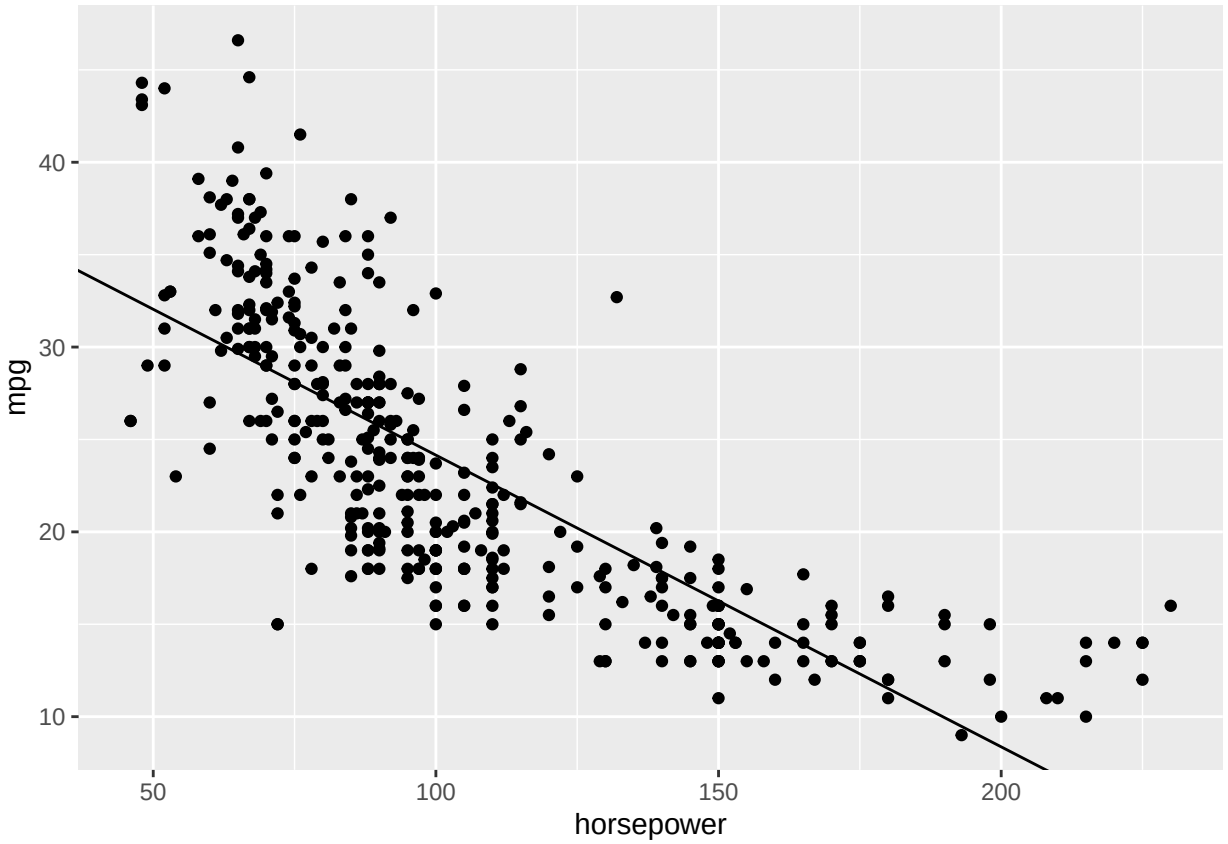
```
df_interval <- dplyr::bind_rows(dfs_interval)
rownames(df_interval) <- NULL

print(df_interval)
```

```
##        fit      lwr      upr   interval
## 1 24.46708 23.97308 24.96108 confidence
## 2 24.46708 14.80940 34.12476 prediction
```

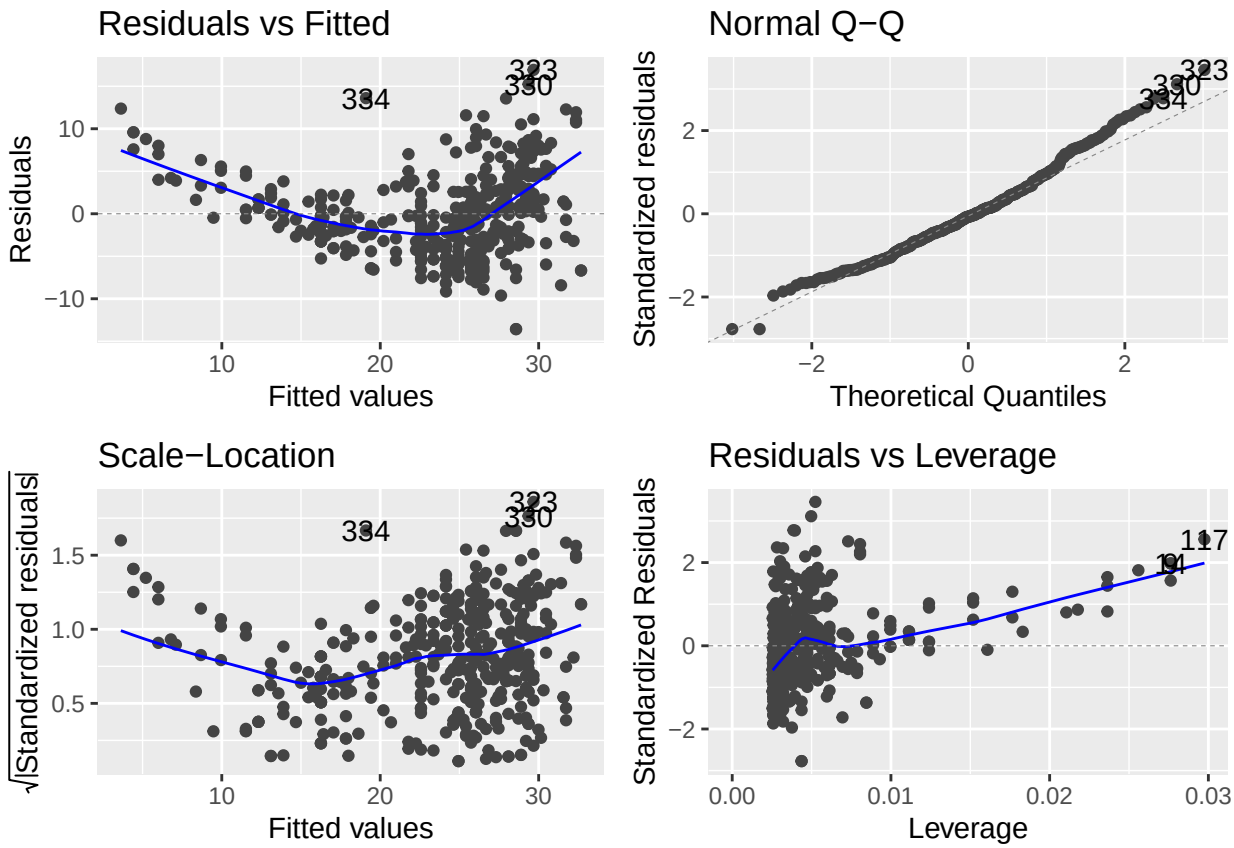
v

```
coefs <- coef(simple_linear_regression_model)
df_auto %>%
  ggplot2::ggplot(.) +
  ggplot2::geom_point(ggplot2::aes(x = horsepower, y = mpg)) +
  ggplot2::geom_abline(slope = coefs[[2]], intercept = coefs[[1]])
```



vi

```
ggplot2::autoplot(simple_linear_regression_model)
```



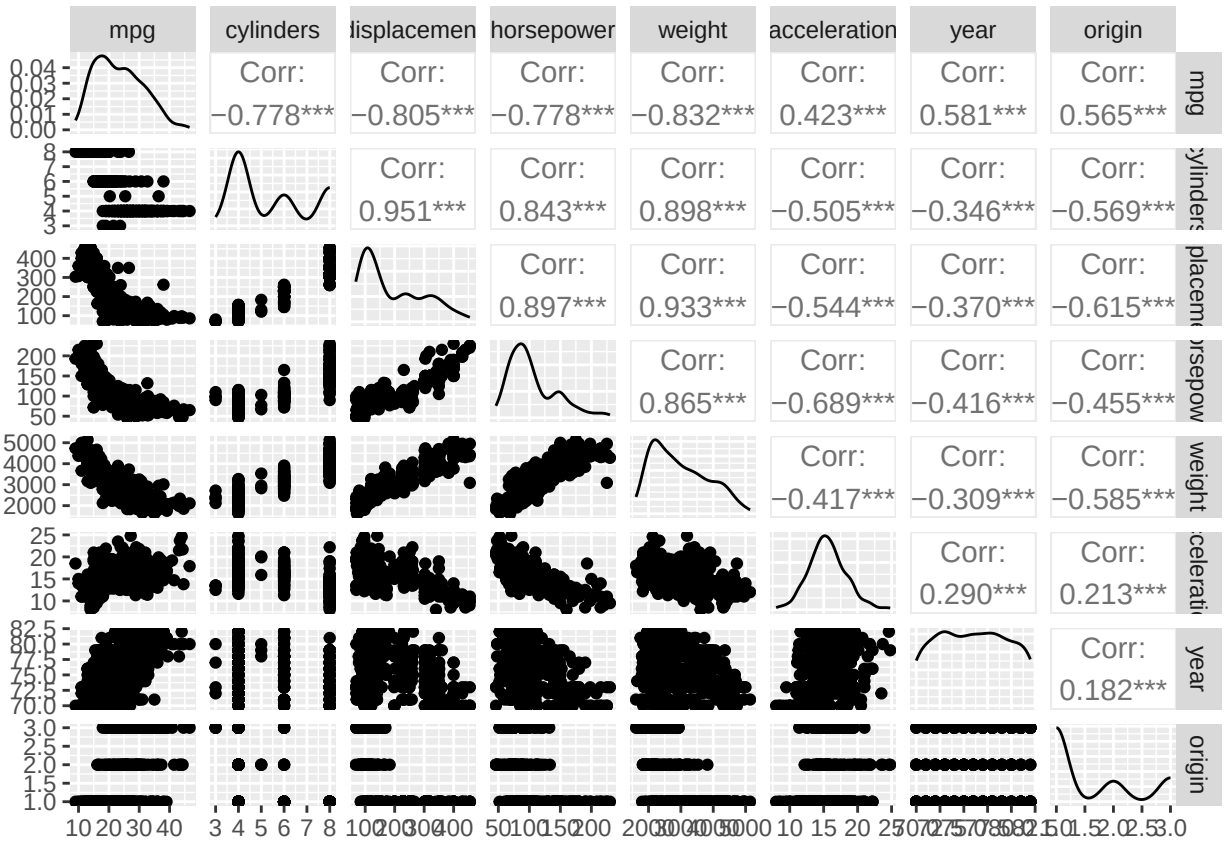
From this plot we notice a few things:

1. Evidence of non-linearity of data - The residuals vs fitted value plot shows evidence of a non-linear relationship, as there is a noticeable pattern to the residuals. You can also see this in the normal QQ plot.
2. No evidence of correlation of error terms - Since each measurement is independent of the other measurements in this data set, this is not a surprise.
3. Evidence of non-constant variance of error terms - The spread of the residuals is greater at higher values of the fitted values, which is related to point 1 above. You can also see this in the normal QQ plot.
4. Leverage - Looking at the standardized residuals vs leverage plot, there are a number of points with high leverage, but most of them have small residuals, i.e. are not outliers. The exceptions are observations `117` and two points which lie right on top of each other, all of which have high leverage and could also be considered borderline outliers (the combination of which gives a high value of Cook's distance).
5. Outliers - There are a few points with standardized residuals with magnitude greater than 2, but most of them have low leverage and thus do not affect the slope of the linear regression line greatly. As mentioned above, points `191` and `117` are exceptions.

Question 9

a

```
GGally::ggpairs(df_auto %>% dplyr::select(., -name))
```



b The correlations can be seen in the above plot.

c

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

print(summary(multiple_linear_regression_model))
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = df_auto)
##
## 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
```

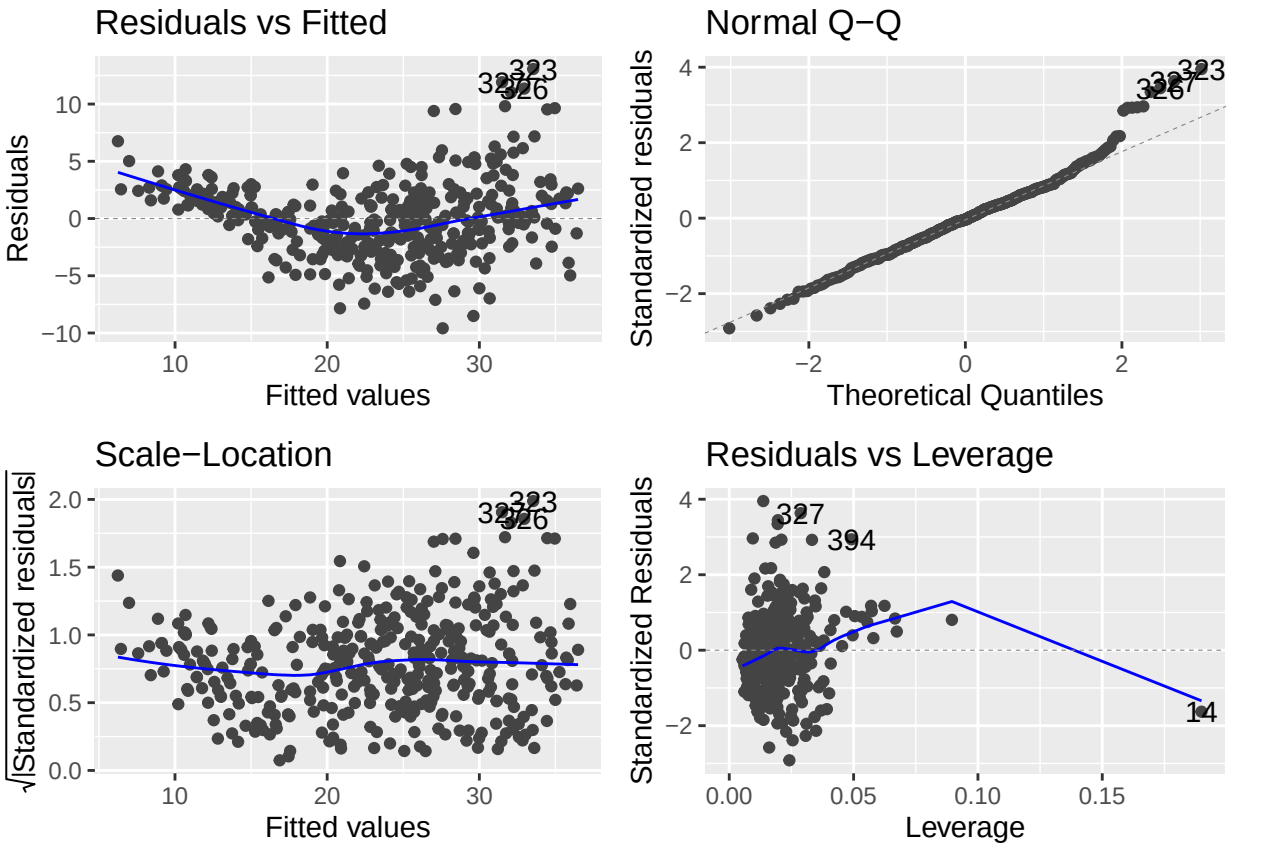
i The tiny p-value associated with the F-statistic indicates that not all of the coefficients are equal to 0, i.e. there is a relationship between at least one of the predictors and `mpg`.

ii Looking at the p-values, `displacement`, `weight`, `year`, and `origin` all appear to have a statistically significant relationship with `mpg`.

iii The coefficient for the `year` variable is positive, indicating that `mpg` has been increasing over time.

d

```
ggplot2::autoplot(multiple_linear_regression_model)
```



1. Evidence of non-linearity of data - The residuals vs fitted value plot shows evidence of a non-linear relationship, as in the case of the earlier simple linear regression. The non-linearity is not as severe as in the single variable case, however.
2. No evidence of correlation of error terms - Same as single variable case.
3. Evidence of non-constant variance of error terms - The spread of the residuals is greater at higher values of the fitted values, which is related to point 1 above. You can also see this in the normal QQ plot.

4. Leverage - Looking at the standardized residuals vs leverage plot, there is one point (point 14) with very high leverage and also a relatively large residual.
5. Outliers -there are a few points with standardized residuals with magnitude greater than 2, but most of them have low leverage and thus do not affect the slope of the linear regression line greatly. Point 14 is a borderline outlier and has high leverage.

e

```
interaction_model <- lm(mpg ~ . -name + acceleration:weight + displacement:horsepower, data = df_auto)

print(summary(interaction_model))
```

```
##
## Call:
## lm(formula = mpg ~ . - name + acceleration:weight + displacement:horsepower,
##     data = df_auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5578 -1.5848 -0.0741  1.4277 12.6105
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -6.803e+00  7.107e+00  -0.957  0.33906
## cylinders                 6.154e-01  3.040e-01   2.025  0.04361 *
## displacement             -7.132e-02  1.166e-02  -6.115 2.39e-09 ***
## horsepower               -1.896e-01  2.245e-02  -8.447 6.35e-16 ***
## weight                   -1.797e-03  1.685e-03  -1.067  0.28687
## acceleration              3.607e-02  3.010e-01   0.120  0.90468
## year                      7.433e-01  4.509e-02  16.487  < 2e-16 ***
## origin                    6.776e-01  2.531e-01   2.677  0.00774 **
## weight:acceleration      -8.668e-05  9.986e-05  -0.868  0.38595
## displacement:horsepower   4.901e-04  6.165e-05   7.950 2.12e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.913 on 382 degrees of freedom
## Multiple R-squared:  0.8639, Adjusted R-squared:  0.8607
## F-statistic: 269.4 on 9 and 382 DF,  p-value: < 2.2e-16
```

The `displacement-horsepower` interaction term is significant.

f

```
polynomial_model <- lm(mpg ~ poly(horsepower, 2), data = df_auto)

print(summary(polynomial_model))
```

```
##
## Call:
## lm(formula = mpg ~ poly(horsepower, 2), data = df_auto)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -14.7135  -2.5943  -0.0859   2.2868  15.8961
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            23.4459     0.2209  106.13   <2e-16 ***
## poly(horsepower, 2)1 -120.1377     4.3739  -27.47   <2e-16 ***
## poly(horsepower, 2)2   44.0895     4.3739   10.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.374 on 389 degrees of freedom
## Multiple R-squared:  0.6876, Adjusted R-squared:  0.686
## F-statistic:   428 on 2 and 389 DF,  p-value: < 2.2e-16
```

```
sqrt_model <- lm(mpg ~ I(sqrt(horsepower)), data = df_auto)

print(summary(sqrt_model))
```

```
##
## Call:
## lm(formula = mpg ~ I(sqrt(horsepower)), data = df_auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9768  -3.2239  -0.2252   2.6881  16.1411
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           58.705      1.349   43.52   <2e-16 ***
## I(sqrt(horsepower))   -3.503      0.132  -26.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.665 on 390 degrees of freedom
## Multiple R-squared:  0.6437, Adjusted R-squared:  0.6428
## F-statistic: 704.6 on 1 and 390 DF,  p-value: < 2.2e-16
```

In the polynomial_model, both the linear and quadratic term are statistically significant. The overall fit of the model is superior to the simple linear model, at least using adjusted R-squared. The square root term is also significant in the square root model, although this model does not perform as well as the quadratic model using adjusted R-squared.

Question 10

a

```
df_carseats <- Carseats
model <- lm(Sales ~ Price + Urban + US, data = df_carseats)
print(summary(model))
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US, data = df_carseats)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

b

1. Price - for a 1 unit change in price, sales go down on average by -0.05 units.
2. Urban - sales in urban populations are on average lower by -0.02 units, although this effect is non-significant.
3. US - sales in the US are on average 1.2 units higher than sales outside the US.

c $\hat{y} = 13.043469 - 0.054459 \cdot Price - 0.021916 \cdot \mathbf{1}_{Urban==Yes} + 1.200573 \cdot \mathbf{1}_{US==Yes}$