Applied Exercises: Homework 1 Emma Taylor

9. Exercise 9 in 3.7

a. Scatterplot Matrix

library(MASS)

library(ISLR)

pairs(Auto)

b. cor(Auto[1:8])

c. Multiple Linear Regression

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

summary(lm.fit)

1. We can see that there is a relationship between the predictors and the response because the F-statistic is 247, much larger than one, and has a p-value close to zero.
2. The intercept, displacement, weight, year, and origin are statistically significant.
3. The positive coefficient for year (.75) suggests that the newer a car is, the higher mpg it is likely to get. For every additional .75 years of the year a car was produced, mpg will increase by 1.

d. Diagnostic Plots

par(mar=c(1,1,1,1))

plot(lm.fit)

We can see some possible outliers in the top right corner of the Residuals vs. Fitted plot, and Data Point 14 in the Residuals vs. Leverage Plot is a high leverage point.

e. Interaction Terms

lm.fit2=lm(mpg~year\*weight+year\*origin,data=Auto[,1:8])

summary(lm.fit2)

In this model, there is a statistically significant effect from the interaction term of year and weight.

f. Transforms

lm.fit3=lm(mpg~horsepower+I(horsepower^2), data=Auto[,1:8])

summary(lm.fit3)

I transformed horsepower by squaring it in order to see if horsepower had a better fitting quadratic relationship with mpg and found that horsepower and horsepower squared were both significant, yet horsepower had a small negative effect on mpg, and horsepower squared had a very small positive relationship with mpg.

lm.fit4=lm(mpg~log(year),data=Auto[,1:8])

summary(lm.fit4)

I then did a log transformation of year on mpg, which was still statistically significant and had a large positive coefficient of 92.

10. Exercise 10 in 3.7

a. Multiple Linear Regression

lm.fit=lm(Sales~Price+Urban+US, data=Carseats)

summary(lm.fit)

b. The intercept coefficient is significant and equal to 13.04, meaning the baseline level of sales is 13. This is not meaningful in the real world, as there is no true baseline of sales. The Price coefficient is significant and equal to -.05, meaning for every $1 dollar increase in price, Sales will go down by .05 dollars. The Urban coefficient is not statistically significant, so it has an effect on Price equal to zero. The US coefficient is significant and equal to 1.2, so for the US market, Sales are augmented to 1.2 units compared to non-US markets.

c. Sales = 13.04 -.05(Price) -0.02(UrbanYes) + 1.2(USYes) + error

d. The null hypothesis can be rejected for Price and US, given their small p-values.

e. Smaller Model

lm.fit2=lm(Sales~Price+US, data=Carseats)

summary(lm.fit2)

f. The models in (a) and (e) have a multiple R^2 of .239, and the model in (a) has an adjusted R^2 of .2335, while the model in (e) has an adjusted R^2 of .2354. These are in general poor fits, but the model in (e) has a slightly better fit.

g. confint(lm.fit2,level=.95)

h. Diagnostic Plots

par(mar=c(1,1,1,1))

plot(lm.fit2)

The diagnostic plots show a decently good relationship, but we can see some high leverage points in the Residuals vs. Leverage plot.

14. Exercise 14 from 3.7

a. Y = 2+2(X1)+0.3(X2)+error

Coefficient for X1 = 2

Coefficient for X2 = 0.3

b. Correlation between X1 and X2

cor(x1,x2) = 0.835

plot(x1,x2)

They are highly linearly related.

c. Regression

lm.fit=lm(y~x1+x2)

summary(lm.fit)

The coefficient for x1 became 1.4, while the coefficient for x2 became 1. In context of the model, these are relatively large changes. The X1 coefficient was significant at the .01 level, while X2 was not significant. The intercept became 2.13, which is close to its original 2. This was highly significant. You can reject the null hypothesis that Beta 1 equals 0, but you cannot do so for Beta 2.

d. X1 Model

lm.fit2=lm(y~x1)

summary(lm.fit2)

This produces a more significant model with coefficients close to their true value, and it allows us to reject the null hypothesis for Beta 1 equals 0.

e. X2 model

lm.fit3=lm(y~x2)

summary(lm.fit3)

This produces a model with a significant intercept and coefficient, but the value of X2’s coefficient is not close to its true value. This model does however allow us to reject the null hypothesis.

f. Because the value of X2 relies heavily on X1 (they are highly correlated), these results make sense. For our complete model, multiple linear regression struggles with collinearity, so it makes sense that we would get poor results. For our X1 model, this model could be more accurate because it can be predicted independent of X2. For our X2 model, this makes sense because the model is trying to predict y without X1 and struggles to do so, but because X2 is affected by X1 and X1 can predict Y well, we will get a falsely well-fitted model.

g. Additional Observation

x1=c(x1, 0.1)

x2=c(x2, 0.8)

y=c(y,6)

lm.fit=lm(y~x1+x2)

summary(lm.fit)

lm.fit2=lm(y~x1)

summary(lm.fit2)

lm.fit3=lm(y~x2)

summary(lm.fit3)

This new addition changed our first model, making X1 insignificant and X2 significant, and making X2’s coefficient much larger than X1’s. It did not drastically change the second or third models.

plot(lm.fit)

Plotting our first model, we see that our new observation is an outlier and a high leverage point, as it is on the top far right of the Residuals vs. Leverage Plot.

plot(lm.fit2)

In the plot of model 2, the new observation is far from the fitted line, but so are many other observations, so it is likely not an outlier or a high leverage point.

plot(lm.fit3)

In the plot of model 3, the new observation is close to the fit line of the Residuals vs. Leverage plot, but out to the far right, so it is a high leverage point.