CV Homework 2

Code

Ramesh Kanakala

PROBLEM 1

Step 1

Hide

```
#exploring Wage instead of Auto
#load ISLR and take a look at Wage
library(ISLR)
data(Wage)
names(Wage)
summary(Wage)
#divide data randomly into train and test sets 75/25
set.seed(1234)
i <- sample(1:nrow(Wage), nrow(Wage)*0.75,
replace=FALSE)
train <- Wage[i,]
test <- Wage[-i,]
```

Step 2

Hide

```
#simple linear regression on the train data with wage as the response and age as the predictor
lm1 <- lm(wage~age, data=train)
summary(lm1)
#mse
pred <- predict(lm1, newdata=train)
mse <- mean((pred - train$wage)^2)
mse
rmse <- sqrt(mse)
rmse
```

Step 3

a. 79.16284 - .76360x
b. As the p-value is very small so we can reject the null hypothesis that there is no relationship between age and wage. Also, the F-statistic is far from 0, and it's p-value very low, meaning that age and wage are related.
c. Positive correlation d.An RSE of 40.9 means the average error of the model was about $40.9 wage, which I believe is not the greatest in this context as wage has a range from 20.09 to 318.34 and almost $40 is a big difference, however, it's not too bad. The adjusted R^2 is 0.04386 which means 4.39% of the variance in wage can be explained by our predictor; that's not good. And again, the F-statistic, of 104.2, is quite far from 0 with a very low associated p-value, meaning that age and wage are related.

d. The MSE by itself is hard to interpret in isolation but when I square root it, it is approximately 40 (exactly 40.88583), meaning that it was off by about $40 wage on average, very similar to the RSE. Not too bad but not good.

Step 4

Hide

```
plot(train$age, train$wage, pch=20, col="black",
main="age vs. wage", xlab="age", ylab="wage")
abline(lm1, lty=5, lwd=2, col='blue')
#predict 65
pred65 <- predict(lm1, data.frame(age=65))
pred65
```

The predicted value of 128.797 for an age of 65 certainly seems in line with the plot between age and wage; the middle of the points between 60 and 70, as well as the line, are aroung $125 for wage.

Step 5

Hide

```
#test on test data
pred <- predict(lm1, newdata=test)
cor1 <- cor(pred, test$wage)
cor1
mse <- mean((pred - test$wage)^2)
mse
rmse <- sqrt(mse)
rmse
```

A correlation of 0.1503787 is not good and means that there is little positive association between predicted values and test values of wage. Again, MSE by itself is hard to interpret but square rooting it and looking at the RMSE tells us that our test data was off $41.02583 wage on average. Compared to the MSE of the training data (1671.651) this MSE of 1683.119 is a little higher and this makes sense because the model was fitted to the train data but the test data us slightly different, leading to more error.

Step 6

Hide

```
#residual plots
par(mfrow=c(2,2))
plot(lm1)
```

In the first graph, the line is quite straight but isn't all the way horizontal, instead is curving downward at the beginning and end. The residuals have some variation at the beginning but vary more at the end, meaning that as age increases, wage can vary greatly. The second graph is a fairly straight diagonal line except at the end where it diverges far from it; the residuals are normally distributed except, again, at the beginning and mostly the end; there is great variation that the model does not capture. Graph 3 has primarily straight line but we see that the residual

points vary slightly at the end. Finally, the fourth graph has a line that falls at the end and many points singled out in the beginning suggesting there are leverage points at those areas. Overall, though it seems primarily normally distributed, there is much evidence of non-linearity especially in the ends of data

Step 7

Hide

```
#linear model with log(wage) as target (interestingly, data comes with a logwage variables)
lm2 <- lm(logwage~age, data=train)
summary(lm2)
```

The Adjusted R^2 of the second model is higher (0.05361 > 0.04386) meaning that the variance is better explained by age when wage is logged.

Step 8

Hide

```
#plot abline for second model where target is log(wage)
plot(train$age, log(train$wage), pch=20, col="black",
main="age vs. log(wage)", xlab="age", ylab="log(wage)")
abline(lm2, lty=5, lwd=3, col='red')
```

Using the log function damped down the x values across the axis bringing the wage values closer to the linear regression line; this line fits the data much better, or closer, than the first model, though it's not too much of an improvement with this data.

Step 9

Hide

```
#test on test data with lm2
pred2 <- exp(predict(lm2, newdata=test))
cor1 <- cor(pred2, log(test$wage))
cor1
mse <- mean((pred2 - test$wage)^2)
mse
```

A correlation is higher now (0.1607536 > 0.1503787) meaning that there is a stronger positive correlation between predicted values and and the log of the test values of wage. The MSE, however, actually increased from 1683.119 to 1739.485 meaning the test data somewhat more off than before.

Step 10

Hide

```
#residual plots for lm2
par(mfrow=c(2,2))
plot(lm2)
```

The second linear model has less variance than model 1 which we can see easily in graphs 1, 2, and 3 with the data points much closer to the line. Graph 4 still seems to be greatly off at the end but there are also less potential outliers and leverage points which we can see with less points being pointed out in these graphs.

PROBLEM 2

Step 1

Hide

```
#a scatterplot matrix for Wage
pairs(Wage)
?pairs()
```

As most of the variables are factors, there is a little number of correlations to observe here. There are positive correlations for age vs wage and logage vs wage

Step 2

Hide

```
#matrix of correlations (minus "name")
cor(Wage$age, Wage$wage)
```

Only numerical wage and age (and logage); correlation is weak and positive.

Step 3

Hide

```
#multiple linear regression with wage as the response and all other variables except name as pre
dictors
lm3 <- lm(wage~age+race+education+jobclass, data=train)
summary(lm3)
```

Decided to add race, educaiton, and job class variables to model as they seemed most impactful. Age, race black, all education levels, and job classes appear to have the most statistically significant relationship to the response as they have the lower p-values.

Step 4

Hide

```
#diagnostic plots of lm3
par(mfrow=c(2,2))
plot(lm3)
Wage[327,]
```

The residuals go off the line in the second Q-Q plot at the beginning and especially at the end.The rest of the plots are much closer to the lines except perhaps a little at the beginnings and ends, but overall much less variance here.

Step 5

Hide

```
#diagnostic plots of lm3
lm4 <- lm(wage~age+race*education+race*jobclass, data=train)
summary(lm4)
#compare lm3 and lm4
anova(lm3, lm4)
```

Race generally plays a part with education and job class so I chose race*education and race*jobclass. This model has a slightly better R^2 (0.279 > 0.2744) meaning it is a slightly better model of the data. The anova() function shows that the new model outperformed the last as model 2 has a lower RSS.