Assignment 1 Solution

Question 1

(a) better - a more flexible approach will fit the data closer and with the large sample size a better fit than an inflexible approach would be obtained.

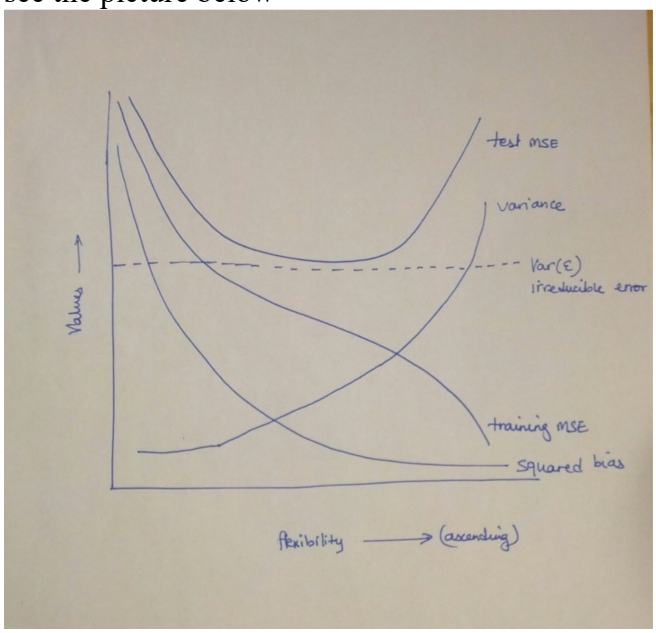
(b) worse - a flexible method would overfit the small number of observations.

(c) better - with more degrees of freedom, a flexible model would obtain a better fit.

(d) worse - flexible methods fit to the noise in the error terms and increase variance.

Question 2

(a) see the picture below



(b) all 5 lines >= 0

i. (squared) bias - decreases monotonically because increases in flexibility yield a closer fit.

ii. variance - increases monotonically because increases in flexibility yield overfit.

iii. training error - decreases monotonically because increases in flexibility yield a closer fit.

iv. test error - concave up curve because increase in flexibility yields a closer fit before it overfits.

v. Bayes (irreducible) error - defines the lower limit, the test error is bounded below by the irreducible error due to variance in the error (epsilon) in the output values (0 <= value). When the training error is lower than the irreducible error, overfitting has taken place.

The Bayes error rate is defined for classification problems and is determined by the ratio of data points which lie at the 'wrong' side of the decision boundary, (0 <= value < 1).

Question 3

$$Y = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_1 * X_2 - 10X_1 * X_3$$

(a) iii. is correct.
Male: $Y = 50 + 20X_1 + 0.07X_2 + 0.01X_1 * X_2$
Female: $Y = 50 + 20X_1 + 0.07X_2 + 35 + 0.01X_1 * X_2 - 10X_1$
Once the GPA is high enough, males earn more on average.

(b) $Y(X_1 = 4.0,\ X_2 = 110,\ X_3 = 1) = 50 + 20 * 4 + 0.07 * 110 + 35 + 0.01 * 4 * 110 - 10 * 4 = 137.1$ (in thousands of dollars)

(c) False. We must examine the p-value of the regression coefficient to determine if the interaction term is statistically significant or not.

Question 4

(a)

```
Library(ISLR2)
Summary(Carseats)
```

```
##      Sales          CompPrice       Income       Advertising
##  Min.   : 0.00   Min.   : 77   Min.   : 21.0   Min.   : 0.00
##  1st Qu.: 5.39   1st Qu.:115   1st Qu.: 42.8   1st Qu.: 0.00
##  Median : 7.49   Median :125   Median : 69.0   Median : 5.00
##  Mean   : 7.50   Mean   :125   Mean   : 68.7   Mean   : 6.63
##  3rd Qu.: 9.32   3rd Qu.:135   3rd Qu.: 91.0   3rd Qu.:12.00
##  Max.   :16.27   Max.   :175   Max.   :120.0   Max.   :29.00
##    Population      Price      ShelveLoc        Age          Education
##  Min.   : 10   Min.   : 24   Bad   : 96   Min.   :25.0   Min.   :10.0
##  1st Qu.:139   1st Qu.:100   Good  : 85   1st Qu.:39.8   1st Qu.:12.0
##  Median :272   Median :117   Medium:219   Median :54.5   Median :14.0
##  Mean   :265   Mean   :116
##  3rd Qu.:398   3rd Qu.:131
##  Max.   :509   Max.   :191
##  Urban       US
##  No :118   No :142
##  Yes:282   Yes:258
##
##
```

```
Attach(Carseats)
lm.fit = lm(Sales~Price+Urban+US)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.921 -1.622 -0.056  1.579  7.058
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.04347
## Price       -0.05446
## UrbanYes    -0.02192
## USYes        1.20057
## ---
0.65101   20.04  < 2e-16 ***
0.00524  -10.39  < 2e-16 ***
0.27165   -0.08     0.94
0.25904    4.63  4.9e-06 ***
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.47 on 396 degrees of freedom
## Multiple R-squared:  0.239,  Adjusted R-squared:  0.234
## F-statistic: 41.5 on 3 and 396 DF,  p-value: <2e-16
```

(c) **Price:** The linear regression suggests a relationship between price and sales given the low p-value of the t-statistic. The coefficient states a negative relationship between Price and Sales: as Price increases, Sales decreases.

UrbanYes: The linear regression suggests that there isn't a relationship between the location of the store and the number of sales based on the high p-value of the t-statistic.

USYes: The linear regression suggests there is a relationship between whether the store is in the US or not and the amount of sales. The coefficient states a positive relationship between USYes and Sales: if the store is in the US, the sales will increase by approximately 1201 units.

(c) Sales = 13.04 + -0.05 Price + -0.02 UrbanYes + 1.20 USYes

(d) Price and USYes, based on the p-values, F-statistic, and p-value of the F-statistic.

(e)

```
lm.fit2 = lm(Sales ~ Price + US)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.927 -1.629 -0.057  1.577  7.052
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079
## Price       -0.05448
## USYes        1.19964
## ---
0.63098   20.65  < 2e-16 ***
0.00523  -10.42  < 2e-16 ***
0.25846    4.64  4.7e-06 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.47 on 397 degrees of freedom
## Multiple R-squared:  0.239,  Adjusted R-squared:  0.235
## F-statistic: 62.4 on 2 and 397 DF,  p-value: <2e-16
```

(f) Based on the RSE and R^2 of the linear regressions, they both fit the data similarly, with linear regression from (e) fitting the data slightly better.

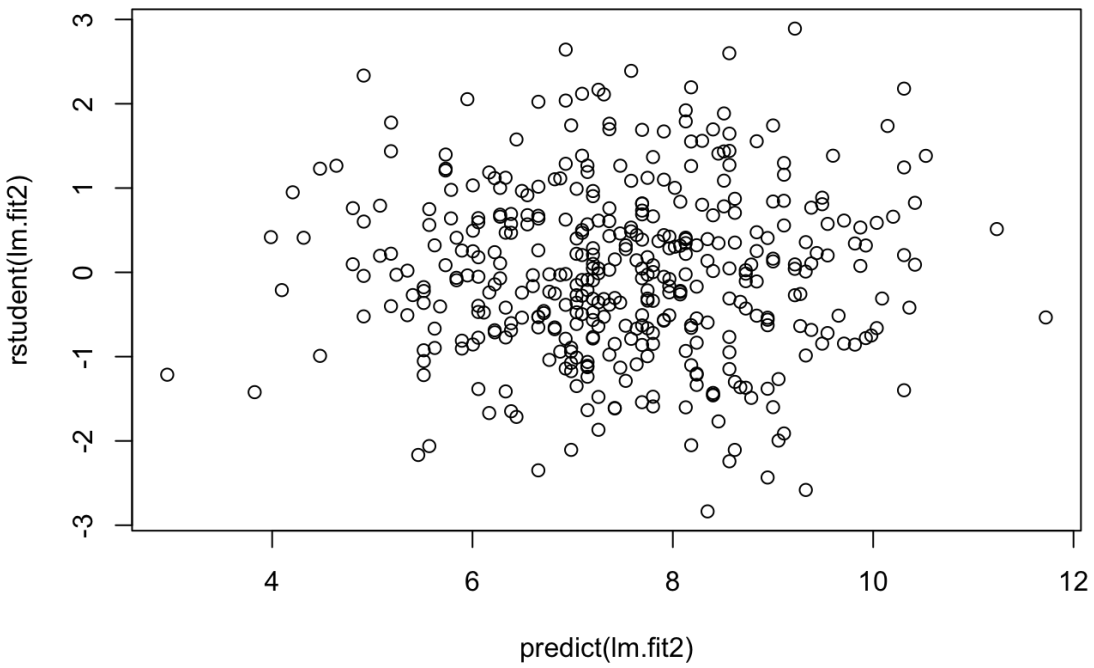
(g)

```
confint(lm.fit2)
```

```
##                2.5 %  97.5 %
## (Intercept) 11.79032 14.2713
## Price       -0.06476 -0.0442
## USYes        0.69152  1.7078
```

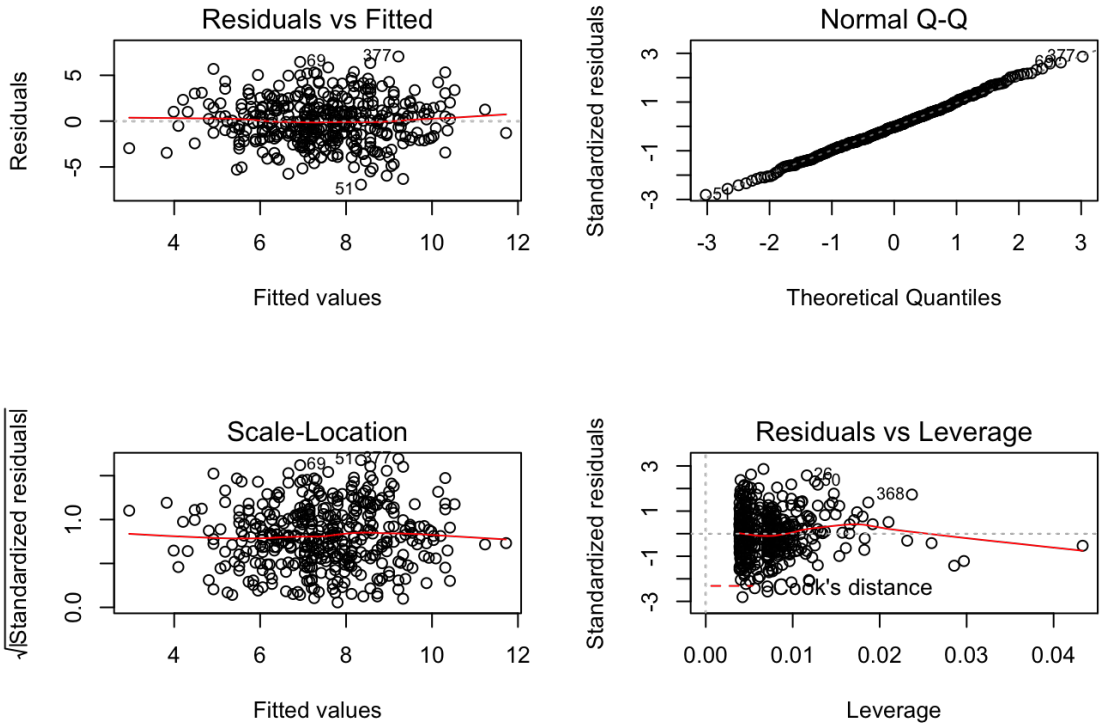
(h)

```
plot(predict(lm.fit2),rstudent(lm.fit2))
```



All studentized residuals appear to be bounded by -3 to 3, so not potential outliers are linear regression from (e) fitting the data slightly better.

```
par(mfrow=c(2,2))
plot(lm.fit2)
```



There are a few observations that greatly exceed $(p+1)/n(0.0076)$ on the leverage-statistic plot that suggest that the corresponding points have high leverage.