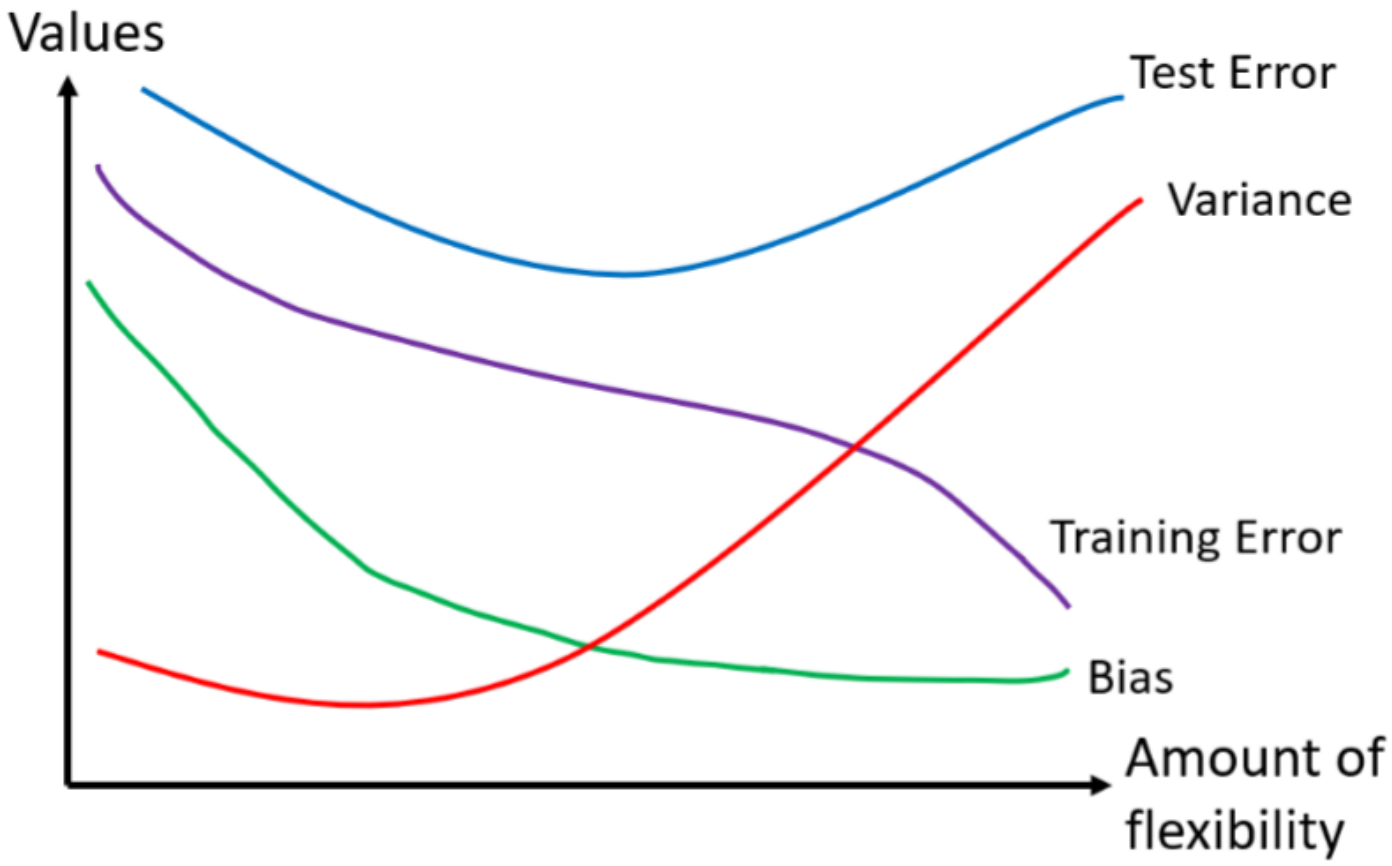
Assignment 1 Solution

Question 1

(a) Better than an inflexible method. Because the more flexible method will fit the data closer and will obtain a better fit than the inflexible method with a large sample size.
(b) Worse than an inflexible method. Because flexible methods will overfit a small number of observations.
(c) Better than an inflexible method. A flexible model will get a better fit because it has more degrees of freedom.
(d) Worse than an inflexible method. Because the flexible method fits the noise in the error term and increases the variance.

Question 2

(a)



(b)
For **training error**, when flexibility increases, the model may fit the training data set more accurately, training error decreases and attains the optimal constant value. Hence, the training error value decrease when increasing the flexibility. Downward sloping occurs.
For **test error**, at first, when the model uses restrictive method which has less flexibility, the model

has less accuracy to fit the test data set due to the simplicity. It causes high value of test error. After that, when the model uses a flexible method which more flexible than the restrictive method, the accuracy become larger. The test error comes to the minimum and it almost tends to explain the true test data set. However, when the model uses an over-flexible method that has extremely large flexibility, overfitting may occur. More and more errors may be found and easily affected by bias. The accuracy become smaller and the test error start to increase. Hence, the graph is downward sloping before reaching the minimum test MSE. After reaching the test MSE, the graph changes to upward sloping. A U-shape case occurs.

For **(squared)bias**, (squared)bias is defined as the error caused by approximating a real-life problem. Since most of the real-life problem perform non-linearly, linear regression (less flexibility method) may not explain the true function and results in a high (squared)bias. When the flexibility become larger, it may explain more accurate to the true function and result in a small (squared)bias. Hence, there is a upward sloping curve on (squared)bias.

For **Variance**, variance is defined as the uncertainly due of the randomness of the training data. As for the given equation [Test MSE = Bias^2 + Variance + Var(e)]. That means when bias increase, variance may decrease. On the other hand, when bias decrease, the variance may increase. Hence, an upward sloping curve on (squared) bias may tend to a downward sloping curve of variance.

Question 3

$$Y = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_1X_2 - 10X_1X_3$$

(a) (iii) is correct.

Male:

$$Y = 50 + 20X_1 + 0.07X_2 + 0.01X_1X_2$$

Female:

$$Y = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_1X_2 - 10X_1$$

Therefore, men earn more on average once their GPA is high enough.

(b) $Y(X_1 = 4.0, X_2 = 110, X_3 = 1) = 137.1$ (in thousands of dollars)

(c) That is false. We must examine the p-values of the regression coefficients to determine whether the interaction term is statistically significant.

Question 4

(a)

```
Call:
lm(formula = Sales ~ Price + Urban + US)

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)

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(formula = Sales ~ Price + US)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9269 -1.6286 -0.0574  1.5766  7.0515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
Price       -0.05448    0.00523 -10.416  < 2e-16 ***
USYes        1.19964    0.25846   4.641 4.71e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared:  0.2393,    Adjusted R-squared:  0.2354
F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-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)

	2.5 %	97.5 %
(Intercept)	11.79032020	14.27126531
Price	-0.06475984	-0.04419543
USYes	0.69151957	1.70776632

(h) All studentized residuals seem to be bounded from -3 to 3, thus the linear regression of (e) is not a potential outlier, it fits the data better. On the leverage chart, there are a few observations that greatly exceed $(p+1)/n (0.0076)$, which indicates that the corresponding point has high leverage.