SDSC5001 Statistical Machine Learning I
Assignment #1 Solutions

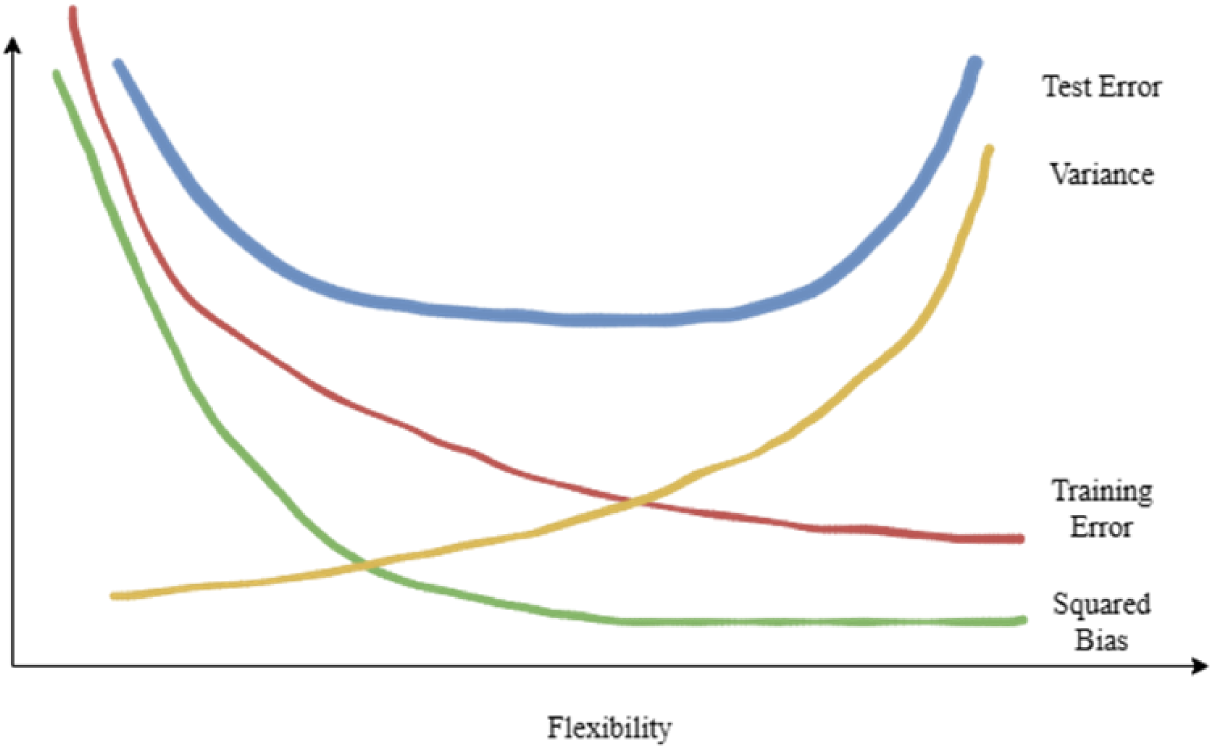
1. (a) Better. Flexible statistical learning method can better fit data with large sample size and small number of predictors.

   (b) Worse, flexible method is easy to overfit facing large number of predictors and small number of observations.

   (c) Better, flexible methods have more degrees of freedom and can capture the nonlinear relationships better.

   (d) Worse, the flexible methods will fit the noise points, and thus increases the variance.

2. (a)



   (b) **Squared bias** is monotonically decreasing due to more flexible methods have better fitting capacity; **Variance** is monotonically increasing because more flexible methods fit the data better. Any change on the data points may cause the estimated function change considerably; **Training error** is monotonically decreasing because more flexible methods have better fitting capacity; **Test error** decreases until it reaches an optimum as flexibility increases, but further increase in flexibility would make the model overfitting.

3. (a) The Euclidean distance between observation X_i and X_j can be calculated by

$$dist(i,j) = \sqrt{(X_{i,1} - X_{j,1})^2 + X_{i,2} - X_{j,2})^2 + X_{i,3} - X_{j,3})^2} \tag{1}$$

   So, the distances between each observation and the test data point are:

   $dist(1, test) = 3, dist(2, test) = 2, dist(3, test) = 3.162, dist(4, test) = 2.236, dist(5, test) = 1.414, dist(6, test) = 1.732$

   (b) Green, the nearest neighbor is observation 5.

   (c) Red, the nearest neighbors are observation 2 (Red), 5 (Green), and 6 (Red).

(d) We would expect K to be small, because in KNN, a smaller value of K will result in a more flexible model, which yields a non-linear decision boundary.

4. (a) Quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration, year;
Qualitative: name, origin;

(b)

mpg	**cylinders**	**displcement**	**horsepower**	**weight**	**acceleration**	**year**
[9.0,46.6]	[3,8]	[68,455]	[46,230]	[1613,5140]	[8,24.8]	[70,82]

Table 0-1

(c)

	mpg	**cylinders**	**displcement**	**horsepower**	**weight**	**acceleration**	**year**
mean	23.446	5.472	194.412	104.469	2977.584	15.541	75.980
std	7.805	1.706	104.644	38.491	849.403	2.759	3.684

Table 0-2

(d)

	mpg	**cylinders**	**displcement**	**horsepower**	**weight**	**acceleration**	**year**
range	[11,46.6]	[3,8]	[68,455]	[46,230]	[1649,4997]	[8.5,24.8]	[70,82]
mean	24.368	5.382	187.754	100.956	2939.644	15.718	77.132
std	7.881	1.658	99.939	35.896	812.650	2.694	3.110

Table 0-3

(e) Hint: Any reasonable tool and comment will be accepted. Below are some instances: According to the plots, it can be found that there exist linear relationships between several variables; The variable mpg is negatively linear related with displacement, horsepower, and weight; ...

(f) Yes, one is able to predict the mpg using the plots above:

```
* Increases in variables displacement, horsepower, and weight will yield a smaller mpg;
* Models with newer years tend to have a higher mpg;
* ...
```

Fence 0-1

5. (a) We can know that

$$Y = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_1 \times X_2 - 10X_1 \times X_3 \tag{2}$$

for male: $Y = 50 + 20X_1 + 0.07X_2 + 0.01X_1 \times X_2$

for female: $Y = 50 + 20X_1 + 0.07X_2 + 35 + 0.01X_1 \times X_2 - 10X_1$

So, if GPA is high enough, males tend to earn more on average.

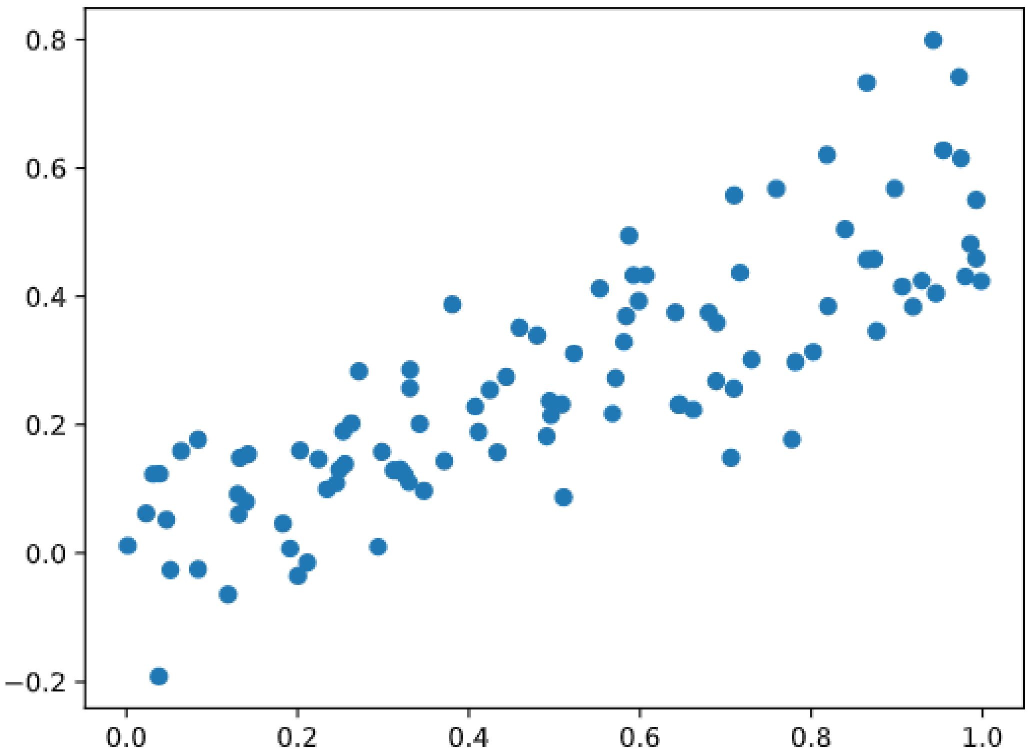
iii is correct.

(b) Follow the Equation in (a), we can derive: $Y = 50 + 20 * 4 + 0.07 * 110 + 35 + 0.01 * 4 * 110 - 10 * 4 = 137.1$

(c) False. Statistical significance is different from the magnitude of the interaction.

6. (a) Setting a random seed ensures consistent results each time.

(b) The correlation between X_1 and X_2 is 0.855 (the answer vary by the random seed).



(c) $\hat{\beta}_0$=2.0203, $\hat{\beta}_1$=1.6967, and $\hat{\beta}_2$=1.2622. Since the p-value of β_1 is relatively small, we can reject the null-hypothesis; Since the p-value of β_2 is large, we fail to reject the null-hypothesis;

(d) $\hat{\beta}_0$=2.0043, $\hat{\beta}_1 = 2.3553$. p-value of β_1 is small, we can reject the null hypothesis that $H_0 : \beta_1 = 0$. Multicollinearity suppresses the individual effects of each independent variable.

(e) $\hat{\beta}_0 = 2.2726, \hat{\beta}_1 = 3.6366$. The p-value of β_1 is almost 0, we can reject the null hypothesis that $H_0 : \beta_1 = 0$. The comments are same as (d).

(f) No, they are not contradictory. The reason is related to the high correlation between $X1$ and $X2$, and that multicollinearity suppresses the individual effects of each independent variable, i.e., when using two variables that are highly collinear, the effect on the response of one variable can be masked by another.

(g) Adding the new observation will cause different effects on these 3 regression models:

- For the modified regression model with $X1$ and $X2$, the model prediction ability is affected. The new observation is not an outlier (outlying Y observation), but is a high leverage point (outlying X observation) according to the studentized residual.

- For the modified regression model with $X1$, the model prediction ability is affected. The new observation is identified as an outlier (outlying Y observation) according to the studentized residual, but not a high leverage point (outlying X observation).

- For the modified regression model with $X2$, the model prediction ability is improved. The new observation has high leverage (outlying X observation) but is not an outlier (outlying Y observation) according to the studentized residual.