

Solutions to Tutorial Questions 1

1. For model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

assume that $X = 0$ is within the scope of the model. What is the implication for the regression function $Y_i = \beta_0 + \beta_1 X_i$ if $\beta_0 = 0$ so that the model is $Y_i = \beta_1 X_i + \varepsilon_i$? How would the regression function $Y_i = \beta_0 + \beta_1 X_i$ plot on the graph? [hint: what is the interpretation of β_0]

The intercept is 0, indicating when $X_i = 0$ the expected value of Y_i is 0. The plot will pass through the origin $(0,0)$.

2. Refer to the regression model above, what is the implication for the regression function if $\beta_1 = 0$ so that the model is

$Y_i = \beta_0 + \varepsilon_i$? How would the regression function $Y_i = \beta_0 + \beta_1 X_i$ plot on the graph? [hint: what is the interpretation of β_1]

The slope is 0, indicating when X_i increases Y_i does not change linearly. The plot will be a horizontal line passing through $(0, b_0)$. Statistically, X has no linear relation with Y .

3. (Grade Point average): For graduate students, X is ACT test score, Y is GPA. There are 120 students, and their ACT and GPA are recorded. The data can be found at ([dataTutorial1a.dat](#)). Suppose we predict Y based on X by a simple linear regression model above.

- Obtain the LSE of β_0 and β_1
- plot the estimated regression function and the data. Does the estimated regression appear to fit the data well?
- Obtain a point prediction of GPA for a student with ACT $X = 30$
- What is the change of the mean response when ACT increases by one point?
- plot the fitted residuals e_i against X_i
- calculate $\sum_{i=1}^{120} e_i^2$

(a)

$$\hat{b}_0 = 2.11405, \quad \hat{b}_1 = 0.003883$$

(b) see code ([code](#)), it looks not so good (we will discuss the issue in lectures later)

(c) $\hat{Y} = 3.278863$

(d) 0.03883

(e) see the code

(f) $\sum_{i=1}^{120} e_i^2 = 45.81761$

4. Derive the expression for

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

from the normal equations.

It is easy to see from the normal equations

$$b_1 = \frac{\sum_{i=1}^n X_i(Y_i - \bar{Y})}{\sum_{i=1}^n X_i(X_i - \bar{X})} = \frac{\sum_{i=1}^n X_i(Y_i - \bar{Y}) - \sum_{i=1}^n \bar{X}(Y_i - \bar{Y})}{\sum_{i=1}^n X_i(X_i - \bar{X}) - \sum_{i=1}^n \bar{X}(X_i - \bar{X})} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

In the second equation

$$\sum_{i=1}^n \bar{X}(X_i - \bar{X}) = 0, \quad \sum_{i=1}^n \bar{X}(Y_i - \bar{Y}) = 0$$

are used.

5. For the model in the first question, prove that the sum of Y_i 's is the same as the sum of fitted values

By the first normal equation

$$\sum_{i=1}^n Y_i = \sum_{i=1}^n (\hat{Y}_i + e_i) = \sum_{i=1}^n \hat{Y}_i + \sum_{i=1}^n e_i = \sum_{i=1}^n \hat{Y}_i$$

6. For the model in the first question, prove that $\sum_{i=1}^n e_i \hat{Y}_i = 0$

the two normal equations are

$$\sum_{i=1}^n e_i = 0 \tag{1}$$

$$\sum_{i=1}^n X_i e_i = 0 \tag{2}$$

the result follows from $(1) \times b_0 + (2) \times b_1$

7. Show that the least squares regression line fitted to data $(5, Y_{1,1}), (5, Y_{1,2}), (5, Y_{1,3}), (10, Y_{2,1}), (10, Y_{2,2}), (10, Y_{2,3}), (15, Y_{3,1}), (15, Y_{3,2}), (15, Y_{3,3})$, is the same as a model fitted to the three points $(5, \bar{Y}_1), (10, \bar{Y}_2)$ and $(15, \bar{Y}_3)$, where $\bar{Y}_1 = (Y_{1,1} + Y_{1,2} + Y_{1,3})/3, \bar{Y}_2 = (Y_{2,1} + Y_{2,2} + Y_{2,3})/3$, and $\bar{Y}_3 = (Y_{3,1} + Y_{3,2} + Y_{3,3})/3$.

Note that

$$\bar{X} = \frac{1}{9}(5 + 5 + 5 + 10 + 10 + 10 + 15 + 15 + 15) = \frac{1}{3}(5 + 10 + 15) = 10,$$

and

$$\bar{Y} = \frac{1}{9}(Y_{1,1} + Y_{1,2} + Y_{1,3} + Y_{2,1} + Y_{2,2} + Y_{2,3} + Y_{3,1} + Y_{3,2} + Y_{3,3}) = \frac{1}{3}(\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3).$$

Thus the estimate of b_1 for the 9 observations is

$$\begin{aligned} \sum_{i=1}^9 (X_i - \bar{X})(Y_i - \bar{Y}) &= -5(Y_{1,1} - \bar{Y}) - 5(Y_{1,2} - \bar{Y}) - 5(Y_{1,2} - \bar{Y}) \\ &\quad + 0(Y_{2,1} - \bar{Y}) + 0(Y_{2,2} - \bar{Y}) + 0(Y_{2,3} - \bar{Y}) \\ &\quad + 5(Y_{3,1} - \bar{Y}) + 5(Y_{3,2} - \bar{Y}) + 5(Y_{3,3} - \bar{Y}) \\ &= 15(\bar{Y}_3 - \bar{Y}_1) \end{aligned}$$

and

$$\sum_{i=1}^9 (X_i - \bar{X})^2 = 6 * 5^2$$

We have the estimate of b_1 based on the 9 observations is

$$b_1 = \frac{\bar{Y}_3 - \bar{Y}_1}{10}$$

Similarly, we can show that the estimator of β_1 based on the averaged 3 values is also

$$b'_1 = \frac{\bar{Y}_3 - \bar{Y}_1}{10}$$

Note that

$$b_0 = \bar{Y} - b_1 \bar{X}.$$

Their estimator of β_0 are also identical.

8. In fitting regression model in the first question, it is found that observation (X_i, Y_i) fell directly on the fitted regression line (i.e. $Y_i = \hat{Y}_i$). If this observation is deleted, would the least square regression line fitted to the remaining $n-1$ cases be changed?
[hint: what is the contribution of the observation to the function Q]

Let b_0, b_1 the estimators of β_0, β_1 based on all n observations $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$.

Without loss of generality, suppose the observation on the fitted line is the first one

$$Y_1 = \hat{Y}_1 = b_0 + b_1 X_1$$

where

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X}) Y_i}{\sum_{i=1}^n (X_i - \bar{X}) X_i}$$

Let

$$\bar{X}' = \sum_{i=2}^n X_i / (n-1) = \frac{n}{n-1} \bar{X} - \frac{1}{n-1} X_1 = \bar{X} + \frac{1}{n-1} (\bar{X} - X_1)$$

and b'_1 be the estimator of β_1 based on $(n-1)$ observations $(X_2, Y_2), \dots, (X_n, Y_n)$, it is

$$b'_1 = \frac{\sum_{i=2}^n (X_i - \bar{X}') Y_i}{\sum_{i=2}^n (X_i - \bar{X}') X_i} = \frac{\sum_{i=2}^n X_i Y_i - \bar{X}' \sum_{i=2}^n Y_i}{\sum_{i=2}^n X_i^2 - \bar{X}' \sum_{i=2}^n X_i}$$

For the numerator, we have

$$\sum_{i=2}^n X_i Y_i - \bar{X}' \sum_{i=2}^n Y_i = \sum_{i=1}^n X_i Y_i - \bar{X} \sum_{i=1}^n Y_i - \frac{n}{n-1} (X_1 - \bar{X})(Y_1 - \bar{Y})$$

and

$$\sum_{i=2}^n X_i^2 - \bar{X}' \sum_{i=2}^n X_i = \sum_{i=1}^n X_i^2 - \bar{X} \sum_{i=1}^n X_i - \frac{n}{n-1} (X_1 - \bar{X})^2$$

Note that

$$\frac{\sum_{i=1}^n X_i Y_i - \bar{X} \sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i^2 - \bar{X} \sum_{i=1}^n X_i} = b_1$$

and

$$\frac{n}{n-1} (X_1 - \bar{X})(Y_1 - \bar{Y}) / \left\{ \frac{n}{n-1} (X_1 - \bar{X})^2 \right\} = b_1$$

[Because $Y_1 - \bar{Y} = b_0 + b_1 X_1 - (b_0 + b_1 \bar{X}) = b_1 (X_1 - \bar{X})$]. Thus

$$b'_1 = b_1$$

Similarly we can prove the estimators of β_0 is unchanged.