QF603 Assignment 2

CHANJUNG KIM

Problem 1

A. In order to derive OLS regression line, the following sum should have minimum value.

$$\sum_{i=1}^{n}\left(Y_i-(\hat{a}+\hat{b}X_i)\right)^2 \qquad \because Y_i = a + bX_i + e_i$$

The minimum occurs at a point where both partial derivatives are equal to zero.

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{a}} = -2\sum_{i=1}^{n}\left(Y_i - \hat{a} - \hat{b}X_i\right) = 0 \qquad (a)$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \hat{b}} = -2\sum_{i=1}^{n} X_i\left(Y_i - \hat{a} - \hat{b}X_i\right) = 0 \qquad (b)$$

The simplification of (a) is as below

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n}\hat{a} + \sum_{i=1}^{n}\hat{b}X_i$$

$$\Rightarrow n\bar{Y} = n\hat{a} + n\hat{b}\bar{X}$$

$$\Rightarrow \bar{Y} = \hat{a} + \hat{b}\bar{X}.$$

$\therefore$ The point $(\bar{x}, \bar{y})$ is on the OLS regression line.

B. $\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (Y_i - \hat{a} - \hat{b} X_i)$

Here, $\hat{a} = \overline{Y} - \hat{b}\overline{X}$

Then $\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (Y_i - \overline{Y} + \hat{b}\overline{X} - \hat{b}X_i)$

$= \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \overline{Y} + \sum_{i=1}^{n} \hat{b}\overline{X} - \sum_{i=1}^{n} \hat{b}X_i$

$= 0$

$(\because \sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \overline{Y}, \ \sum_{i=1}^{n} \hat{b}\overline{X} = \sum_{i=1}^{n} \hat{b}X_i)$

C. By definition $\hat{e}_i = Y_i - \hat{Y}_i$

Therefore, $\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{Y}_i$

In the question above, it is proved that $\sum_{i=1}^{n} \hat{e}_i = 0$.

Therefore, $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$

By dividing both side with n, we can get

$\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i$

$\therefore$ We can say that $\overline{Y_i} = \overline{\hat{Y}_i}$

D) $\sum_{i=1}^{n} \hat{y}_i \hat{e}_i$

$$= \sum_{i=1}^{n} (\hat{a} + \hat{b}x_i) \hat{e}_i$$

$$= \sum_{i=1}^{n} \hat{a}\hat{e}_i + \sum_{i=1}^{n} \hat{b}x_i\hat{e}_i = \sum_{i=1}^{n} \hat{b}x_i\hat{e}_i \quad (\because \hat{a} \text{ is constant, } \sum_{i=1}^{n}\hat{e}_i = 0)$$

From the partial derivative of $\sum_{i=1}^{n} e_i^2$, we know that

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b} = -2\sum_{i=1}^{n} x_i(y_i - a - bx_i) = -2\sum_{i=1}^{n} x_i \cdot e_i = 0$$

$$\therefore \sum_{i=1}^{n} \hat{b}x_i\hat{e}_i = 0$$

Therefore, we can say that $\sum_{i=1}^{n} \hat{y}_i \hat{e}_i = 0$

Problem 2

if the linear regression is without intercept, the property B in problem 1 doesn't hold anymore

From the first FOC in problem 1-A, we derived

$$\sum_{i=1}^{n}(Y_i - a - bX_i) = \sum_{i=1}^{n} e_i = 0$$

This expression implies that $\sum_{i=1}^{n} e_i$ can go to zero only when there is a constant term in the regression.

Therefore, if regression does not have a constant, $\sum_{i=1}^{n} e_i$ may not be zero.

Problem 3.

According to the chart, X and Y seem to be evenly distributed around the regression line $y = x. (0 \leqq x \leqq 100)$

If we add another dot to this chart in line with the distribution of the plot, K% confidence interval of the dot's Y value will be

$$Y_{1001} \pm t_{998, \frac{100+k}{2}\%} \times \underbrace{29}_{\text{given } \hat{\sigma}_e} \sqrt{1 + \frac{1}{n} + \frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Since we have fairly distributed 1000 X values in the chart, we can say that $\frac{(X_{n+1} - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ is negligibly small.

And locating the new dot in the positive territory of the Y-axis means that the confidence interval follows one-tailed test.

Therefore, when the lower bound of 90% confidence interval is greater than 0, the dot has the positive value with 95% chance

$$Y_{1001} - 1.646 \times 29 \sqrt{1 + \frac{1}{1000}} > 0 \quad (\text{where } t_{998, 95\%} \doteqdot 1.646)$$

$$\therefore Y_{1001} > 47.758$$

The regression line of x and y is $y = x$.

In conclusion, when X value of the dot is greater than 47.758, the dot will be in positive territory of the Y axis with 95% chance.