Problem 1. Consider the simple linear regression $y_i = a + b\,x_i + \varepsilon_i$. Show that

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

Answer. The residual sum of squares (RSS) is given by

$$\text{RSS} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - a - b\,x_i\right)^2.$$

The first-order conditions with respect to a and b result in the estimates $\hat{a}$ and $\hat{b}$ that satisfy

$$\frac{\partial \text{RSS}}{\partial a} = -2\sum_{i=1}^{n} \left(y_i - \hat{a} - \hat{b}\,x_i\right) = 0, \tag{1}$$

$$\frac{\partial \text{RSS}}{\partial b} = -2\sum_{i=1}^{n} x_i\left(y_i - \hat{a} - \hat{b}\,x_i\right) = 0. \tag{2}$$

Working on equation (1), which is

$$\sum_{i=1}^{n} y_i - n\hat{a} - \hat{b}\sum_{i=1}^{n} x_i = 0,$$

and since $n\bar{y} = \sum_{i=1}^{n} y_i$ and $n\bar{x} = \sum_{i=1}^{n} x_i$, we obtain

$$n\bar{y} - n\hat{a} - \hat{b}\,n\bar{x} = 0.$$

Rewriting this result, we have

$$\bar{y} = \hat{a} + \hat{b}\,\bar{x}.$$

In other words, $\boxed{\text{the point } (\bar{x}, \bar{y}) \text{ is on the OLS regression line, } y_i = \hat{a} + \hat{b}\,x_i.}$ □

B. The OLS residuals add up to zero, i.e., $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$.

Answer. The residual $\hat{\varepsilon}_i$ is $y_i - \hat{a} - \hat{b}\,x_i$. Since $\hat{a} = \bar{y} - \hat{b}\,\bar{x}$, the residual $\hat{\varepsilon}_i$ equals $y_i - \bar{y} + \hat{b}\,\bar{x} - \hat{b}\,x_i$. Applying $n\bar{y} = \sum_{i=1}^{n} y_i$ and $n\bar{x} = \sum_{i=1}^{n} x_i$, we have

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} \left(y_i - \bar{y} + \hat{b}\,\bar{x} - \hat{b}\,x_i\right) = n\bar{y} - n\bar{y} + n\hat{b}\,\bar{x} - \hat{b}\,n\bar{x} = 0.$$

□

C. $\bar{y} = \bar{\hat{y}}$, i.e., the sample average of the actual y_i is the same as the sample average of the fitted values.

Answer. Since $\hat{\varepsilon}_i = y_i - \hat{a} - \hat{b}\,x_i$ and $\hat{y}_i := \hat{a} + \hat{b}\,x_i$ hence $y_i = \hat{y}_i + \hat{\varepsilon}_i$. Consequently,

$$\frac{1}{n}\sum y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i + \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i.$$

We have proved earlier that $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$. Accordingly, we have

$$\boxed{\bar{y} = \bar{\hat{y}}.}$$

□

D. $\sum_{i=1}^{n} \hat{y}_i \hat{\varepsilon}_i = 0.$

Answer. Note that $\hat{y}_i = \hat{a} + \hat{b}\,x_i$. Therefore,

$$\sum_{i=1}^{n} \hat{y}_i \hat{\varepsilon}_i = \sum_{i=1}^{n} \big(\hat{a} + \hat{b}\,x_i\big)\hat{\varepsilon}_i = \hat{a}\sum_{i=1}^{n} \hat{\varepsilon}_i + \hat{b}\sum_{i=1}^{n} x_i\hat{\varepsilon}_i$$

We have already shown that $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$. All we need to do is to show that $\sum_{i=1}^{n} x_i\hat{\varepsilon}_i = 0$. Indeed, from equation (2), which is the second first-order condition with respect to b, we get

$$0 = \sum_{i=1}^{n} x_i\big(y_i - \hat{a} - \hat{b}\,x_i\big) = \sum_{i=1}^{n} x_i\big(y_i - \hat{y}\big) = \sum_{i=1}^{n} x_i\hat{\varepsilon}_i.$$

□

Problem 2. Does the property B in Question 1 still hold if the linear specification is without the intercept, i.e., $y_i = b\,x_i + \varepsilon_i$? Explain your answer.

Answer. The OLS regression leads to

$$y_i = \hat{a} + \hat{b}\,x_i.$$

Consider a related regression $y_i = \alpha + \beta\,x_i + e_i$ for the same set of data of x_i and y_i, $i = 1, 2, \ldots, n$. Since the point $(\bar{x}, \bar{y})$ lies on the OLS line, we have

$$\bar{y} = \hat{\alpha} + \hat{\beta}\,\bar{x}. \tag{3}$$

Now,

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} \big(y_i - \hat{b}\,x_i\big) = n\bar{y} - \hat{b}\,n\bar{x}.$$

This is re-expressed as

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i + \hat{b}\,\bar{x}.$$

Substitute this $\bar{y}$ into equation (3), we have

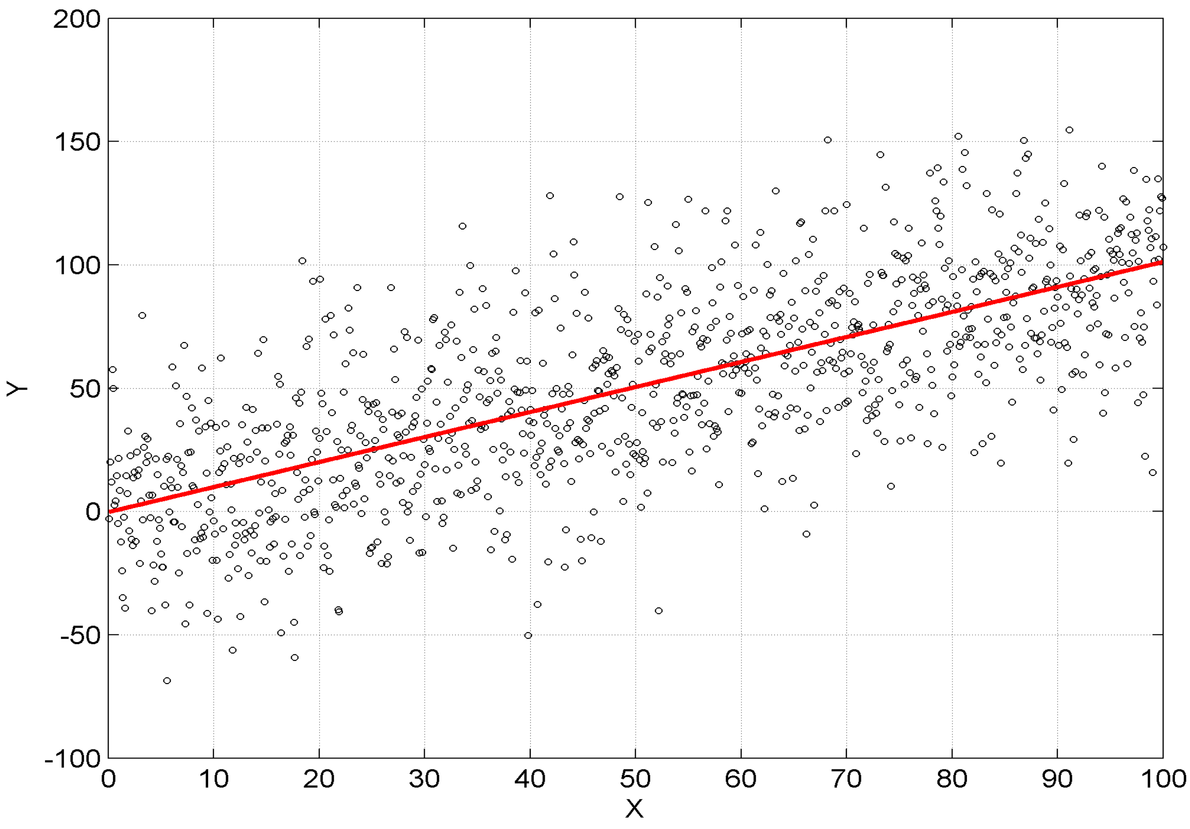
$$\sum_{i=1}^{n}\hat{\varepsilon}_i + n\hat{b}\,\bar{x} = n\big(\hat{\alpha} + \hat{\beta}\,\bar{x}\big).$$

After re-arranging, we obtain

$$\sum_{i=1}^{n}\hat{\varepsilon}_i = n\big(\hat{\alpha} + (\hat{\beta} - \hat{b})\bar{x}\big).$$

In general, $\hat{\alpha} + (\hat{\beta} - \hat{b})\bar{x} \neq 0$. So if the intercept is omitted, $\sum_{i=1}^{n}\hat{\varepsilon}_i \neq 0$. □

Problem 3. Consider an OLS regression of Y on X using 1,000 observations. The straight line through the plot below is $\widehat{Y} = \hat{a} + \hat{b}X$, and the standard error of the regression, typically denoted by $\hat{\sigma}_e$, is 29.



Now, another dot is going to be added to this chart, in line with the distribution of the plot. Choose the X value of the dot in such a way that a Y value of greater than zero is obtained. More precisely, at what value of X are you going to have a 95% chance of getting a dot such that it is in the positive territory of the Y axis? Note that all the information required to answer this question is already given in the chart (plus the fact that $\hat{\sigma}_e = 29$). Provide the arguments and workings by which you arrive at your answer.

Answer. First, we note that the OLS line is $\hat{a} = 0$ and $\hat{b} = 1$, i.e., $Y_i = X_i$, $i = 1, 2, \ldots, 1000$. When X_{1001} is given, the predicted value $\hat{Y}_{1001}$ is such that

$$\mathbb{P}\big[\hat{Y}_{1001} > 0 \big| X_{1001}\big] > 95\%.$$

Therefore, we are examining the lower bound (LB) of this point prediction:

$$\text{LB} = \hat{Y}_{1001} - t_{998,95\%} \times 29\sqrt{1 + \frac{1}{1000} + \frac{(X_{1001} - \overline{X})^2}{\sum_{i=1}^{1000}(X_i - \overline{X})^2}}$$

Now, $n = 1000$ is large,

$$\text{LB} \approx \hat{Y}_{1001} - t_{998,95\%} \times 29.$$

Since $Y_i = 0 + 1 \times X_i$ is the OLS line, we write

$$\text{LB} \approx X_{1001} - t_{998,95\%} \times 29.$$

Now, in the context of this question, the critical value of $t_{998,95\%}$ is referring to one-tail critical value at the 5% significance level! Checking the table, one-tail $t_{998,95\%} \approx 1.646$ Therefore, for $\text{LB} > 0$,

$$X_{1001} > 1.646 \times 29 = 47.73.$$

This question is taken from a nice article: "How economists get tripped up by statistics." □