Homework 1, STAT 5205

Zongyi Liu

Jan 31, 2025

Question 1

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ with $\hat{\beta}_0, \hat{\beta}_1$ being the least square estimators (LSE). Let $\hat{\epsilon}_i = y_i - \hat{y}_i$. Please show the two claims below.

- $\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i$
- $\sum_{i=1}^n x_i \hat{\epsilon}_i = 0$

Answer

Property 1 For $\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i$, $\hat{\epsilon}_i = y_i - \hat{y}_i$, and thus $\hat{y}_i = y_i - \hat{\epsilon}_i$. In linear regression, the residual holds a property that $\sum_{i=1}^n \hat{\epsilon}_i = 0$. Thus $\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n (y_i - \hat{\epsilon}_i)$, and $\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i$.

Mathematically, it means that the regression line is designed to 'balance out' the errors above and below the line, which leads to the sum of predicted values equaling the sum of observed values.

Property 2 Here we start with the definition of residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i$. Multiplying both sides by x_i and summing: $\sum_{i=1}^n x_i \hat{\epsilon}_i = \sum_{i=1}^n x_i (y_i - \hat{y}_i)$

To calculate the sum of residuals: $\sum_{i=1}^n (\hat{\epsilon}_i) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))$, so $\sum_{i=1}^n (\hat{\epsilon}_i) = \sum_{i=1}^n (y_i) - n\beta_0 - \beta_1 \sum_{i=1}^n (x_i)$. Since the average of y (which is $\bar{y}$) is used to calculate the intercept (β_0) and the average of x (which is $\bar{x}$) is used in the slope calculation (β_1), the sum of residuals becomes: $\sum_{i=1}^n (\hat{\epsilon}_i) = n\bar{y} - n\beta_0 - \beta_1 n\bar{x} = 0$

Then to prove the sum of x multiplied by epsilon is zero, we have $\sum_{i=1}^n (x_i \hat{\epsilon}_i) = \sum_{i=1}^n (x_i (y_i - (\beta_0 + \beta_1 x_i)))$. $\sum_{i=1}^n (x_i \hat{\epsilon}_i) = \sum_{i=1}^n (x_i y_i) - \beta_0 \sum_{i=1}^n (x_i) - \beta_1 \sum_{i=1}^n (x_i^2)$. Using the property that the sum of residuals is zero ($\sum_{i=1}^n \epsilon_i = 0$), and considering the average x value again, this sum becomes zero: $\sum_{i=1}^n (x_i \hat{\epsilon}_i) = \sum_{i=1}^n (x_i y_i) - \beta_0 n\bar{x} - \beta_1 n\bar{x}^2 = 0$. Since $\frac{1}{n} \sum_{i=1}^n x_i y_i = \bar{x}\bar{y}$, we get $\sum_{i=1}^n (x_i \hat{\epsilon}_i) = n\bar{x}(\bar{y} - \beta_0 - \beta_1 \bar{x}) = 0$

Mathematically, it means that the residuals and x values are uncorrelated, which is one of the most important premises of linear regression model.

Question 2

Let $k_i = \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Please show the following properties of k_i s.

- $\sum_{i=1}^{n} k_i = 0$.
- $\sum_{i=1}^{n} k_i x_i = 1$.
- $\sum_{i=1}^{n} k_i^2 = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Answer

For property 1, we have $\sum_{i=1}^{n} k_i = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. For $\bar{x}$ is the mean of $x_1, x_2, \ldots, x_n$, given by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Since the numerator is the sum of deviations from the mean, we have $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, so the total number is zero, it follows that $\sum_{i=1}^{n} k_i = \frac{0}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0$.

Property 2, here we have $\sum_{i=1}^{n} k_i x_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Rewrite x_i as $(x_i - \bar{x}) + \bar{x}$, we get $\sum_{i=1}^{n} k_i x_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(\bar{x} + (x_i - \bar{x}))}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Expanding the product we get $\sum_{i=1}^{n} k_i x_i = \sum_{i=1}^{n} \frac{\bar{x}(x_i - \bar{x}) + (x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Then split the sum $\sum_{i=1}^{n} k_i x_i = \frac{\bar{x}\sum_{i=1}^{n}(x_i - \bar{x}) + \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, the first term cancelled, and we have $\sum_{i=1}^{n} k_i x_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 1$.

Property 3, we have $\sum_{i=1}^{n} k_i^2 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$. Then factor out the denominator $\sum_{i=1}^{n} k_i^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Thus we have $\sum_{i=1}^{n} k_i^2 = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Question 3

Please watch these videos and try to understand the material.

- https://www.youtube.com/watch?v=27vT-NWuw0M
- https://www.youtube.com/watch?v=t-n4a18AW08