Probability and Statistics Notes series - Note Set 2

Regression Models

By: @0.5mins

Last update: September 5, 2025

Abstract

This note on regression model is developed based on Dr KY Liu's treatment in the course STAT3008 Applied Regression Analysis at CUHK, incorporating the textbook *Introduction to Linear Regression Analysis* by Montgomery, Peck and Vining, and other relevant materials. In this note, we will have a complete treatment of the implementation of regression models.

Contents

1 Simple linear regression

1.1 Interpolation

Suppose we have a data set $\{(x_i, y_i) : i = 1, 2, \cdots, n\} \subset \mathbb{R}^2$, where x-part is the independent variable and y-part is the dependent variable (so naturally x_i's are distinct). We wish to construct a model to predict the y values using the x values. A naive idea would be to find a polynomial $y = f(x)$ that passes through all data points, i.e. satisfies $y_i = f(x_i)$. To do so, we can use **Lagrange interpolation formula**. Define the **basis polynomials** by

$$f_i(x) = \prod_{\substack{k=1 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k}.$$

Note that they satisfy

$$f_i(x_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

Thus we have

$$y_i f_i(x_j) = \begin{cases} y_i & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

Summing up, we have the polynomial

$$f(x) = \sum_{i=1}^{n} y_i f_i(x)$$

which is a polynomial of degree $n - 1$ satisfying $y_i = f(x_i)$.

This is not a good model actually, because we usually obtain y_i by measurement, which means there is error. We need to control the error of the measurement from our model, but the Lagrange interpolation formula would take the error of the measurement into account as well. This is why we are going to study regression models in this note, because regression models also help us to handle the error using statistical methods.

1.2 Simple linear regression

Let's build up the simple linear regression model. Consider a data set $\{(x_i, y_i) : i = 1, 2, \cdots, n\} \subset \mathbb{R}^2$, where x_i's are distinct. As lazy guys, we wish the data set to satisfy

$$y = \beta_0 + \beta_1 x$$

for some fixed constants β_i. Of course this is not always possible, so there should exist an error term ε such that

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{1}$$

This is called the **simple linear regression model**. In this equation,

- x is the independent variable, which is customarily called **regressor** variable in regression analysis, and
- y is the dependent variable, which is customarily called **response** variable in regression analysis.
- The **intercept** β_0 and the **slope** β_1 are unknown constants which we collectively refer them to as the **regression coefficients**.
- ε is the error term, which at this stage we assume $\varepsilon \sim N(0, \sigma^2)$ where σ is fixed.

The parameters β_i are unknown, and we need to estimate them by using sample data. This is where our data set comes into play. By our assumption, we may write

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

for which (x_i, y_i) are the sample data, and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. This is called the **sample regression model**.

In practice, the regressor x is controlled and measured with *negligible error*, while y is a random variable that we measure. Yes, y is a random variable, because $\varepsilon \sim N(0, \sigma^2)$, and β_i and x are fixed, so eventually $y \sim N(\beta_0 + \beta_1 x, \sigma^2)$:

$$\mathbb{E}[y|x] = \mathbb{E}[\beta_0 + \beta_1 x + \varepsilon] = \mathbb{E}[\beta_0] + \mathbb{E}[\beta_1 x] + \mathbb{E}[\varepsilon] = \beta_0 + \beta_1 x$$

$$\mathbb{V}(y|x) = \mathbb{V}(\beta_0 + \beta_1 x + \varepsilon) = \mathbb{V}(\varepsilon) = \sigma^2$$

Now we can see the interpretations of β_i:

- $\beta_0 = \mathbb{E}[y|x = 0]$ which is the mean of the distribution of the response y when $x = 0$, and
- $\beta_1 = [\beta_0 + \beta_1(x + 1)] - [\beta_0 + \beta_1 x] = \mathbb{E}[y|x + 1] - \mathbb{E}[y|x]$ which is the change in the mean of the distribution of y produced by a unit change in x.

1.2.1 Least square estimation

The **least square estimators** $\hat{\beta}_i$ of β_i are defined to be the minimizer of the sum of squares

$$S = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

What are the summands $(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$? Recall that y_i are the observed values of the response variable in our data set, and $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is called the **fitted simple linear regression model**. Using the fitted simple linear regression model, we can obtain a predicted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ for each x_i. The gap between y_i and $\hat{y}_i$ is called **residual**, which is denoted by

$$e_i = y_i - \hat{y}_i.$$

This e_i can be viewed as the *signed distance* from y_i to the corresponding point on the fitted simple linear regression model with the same x_i value. To minimize the *unsigned distance*, we consider the **sum of squares**. We prefer squares over absolute values because squares are smooth, but absolute values are not smooth and nonlinear.

Okay, after explaining the choice of estimator, we finally start deriving it. Here we provide a analytical method of derivation, and in a later section we will derive it by linear algebraic methods. The least square estimators $\hat{\beta}_i$ satisfies

$$\frac{\partial S}{\partial \hat{\beta}_0} = \frac{\partial S}{\partial \hat{\beta}_1} = 0$$

as given by first derivative test. We compute the derivatives as follows:

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2 \left\{ \sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i \right\}$$

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2 \left\{ \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \right\}$$

Thus we have

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \quad \text{and} \quad \sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2.$$

These equations are called the **least squares normal equations**. Looking at the first normal equation, if we consider the sample means $\overline{x} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_i$ and $\overline{y} = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} y_i$, then it becomes

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x} \quad \Longrightarrow \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

Putting it into the second normal equation, we have

$$\begin{aligned} \sum_{i=1}^{n} x_i y_i &= (\overline{y} - \hat{\beta}_1 \overline{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= n\overline{x}(\overline{y} - \hat{\beta}_1 \overline{x}) + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= n\overline{x}\ \overline{y} + \hat{\beta}_1 \left\{ \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right\} \end{aligned}$$

and hence

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\ \overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}.$$

Thus we have already obtained the least square estimators as

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\ \overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}. \tag{2}$$

Example 1.1. A study investigated whether the average number of tweets (or messages) per hour prior to the movie's release on Twitter.com could be used to forecast the opening weekend box office revenues of movies. The two variables of a sample of 23 movies were measured. Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where y_i are the weekend box office revenues and x_i are the average number of tweets per hour. It is given that

$$\sum_{i=1}^{n} x_i^2 = 4933199 \quad \sum_{i=1}^{n} y_i^2 = 35626.09 \quad \sum_{i=1}^{n} x_i y_i = 396603.2$$
$$\sum_{i=1}^{n} x_i = 6980.65 \quad \sum_{i=1}^{n} y_i = 576.3$$

Compute the least square estimators of β_0 and β_1. What are the practical meanings of the regression coefficients β_0 and β_1? Are the signs of the estimators sensible, based on the practical meanings of the regression coefficients?

Solution.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\ \overline{y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{(396603.2) - (23)\left(\frac{6980.65}{23}\right)\left(\frac{576.3}{23}\right)}{(4933199) - (23)\left(\frac{6980.65}{23}\right)^2} \approx 0.07876722$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \approx \left(\frac{576.3}{23}\right) - (0.07876722)\left(\frac{6980.65}{23}\right) \approx 1.150158$$

- β_0 is the average weekend box office revenues when the tweet rate is 0. So the sign of $\hat{\beta}_0$ makes sense because, even no one talks about a movie, there should be someone who watched the movie.
- β_1 is the change in average weekend box office revenues when the tweet rate is increased by 1 unit. So the sign of $\hat{\beta}_1$ makes sense because more people talks about a movie implies the movie is more well-known, and so it makes sense that more people would watch the movie. ■

The formula we have for $\hat{\beta}_1$ looks rather messy. To write down a cleaner formula, let

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 \quad \text{and} \quad S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

Then we have

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}.$$

Proof. Note that

$$\begin{aligned} S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^{n} (x_i y_i - \overline{x} y_i - x_i \overline{y} + \overline{x}\ \overline{y}) \\ &= \sum_{i=1}^{n} x_i y_i - \overline{x} \sum_{i=1}^{n} y_i - \overline{y} \sum_{i=1}^{n} x_i + n\overline{x}\ \overline{y} \\ &= \sum_{i=1}^{n} x_i y_i - \overline{x}(n\overline{y}) - \overline{y}(n\overline{x}) + n\overline{x}\ \overline{y} \\ &= \sum_{i=1}^{n} x_i y_i - n\overline{x}\ \overline{y} \end{aligned}$$

which is indeed the numerator of $\hat{\beta}_1$. Similarly,

$$\begin{aligned} S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 &= \sum_{i=1}^{n} (x_i^2 - 2\overline{x} x_i + \overline{x}^2) \\ &= \sum_{i=1}^{n} x_i^2 - 2\overline{x} \sum_{i=1}^{n} x_i + n\overline{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\overline{x}(n\overline{x}) + n\overline{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \end{aligned}$$

which is indeed the denominator of $\hat{\beta}_1$. ■

1.2.2 Properties of the least square estimators

As said earlier, we choose to consider the sum of squares because it satisfies a range of clean properties, which we showcase as follows.

Proposition 1.1. $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combination of y_i's.

Proof. Recall that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{S_{xx}}.$$

We manipulate the numerator S_{xy} as follows:

$$\begin{aligned} S_{xy} = \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) &= \sum_{i=1}^n \{y_i(x_i - \overline{x})\} - \overline{y} \sum_{i=1}^n (x_i - \overline{x}) \\ &= \sum_{i=1}^n \{y_i(x_i - \overline{x})\} - \overline{y} \left(\sum_{i=1}^n x_i - n\overline{x} \right) \\ &= \sum_{i=1}^n \{y_i(x_i - \overline{x})\} - \overline{y} \left(\sum_{i=1}^n x_i - \sum_{i=1}^n x_i \right) \\ &= \sum_{i=1}^n \{y_i(x_i - \overline{x})\} \end{aligned}$$

Therefore

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{S_{xx}} \\ &= \frac{\sum_{i=1}^n \{y_i(x_i - \overline{x})\}}{S_{xx}} \\ &= \sum_{i=1}^n \left\{ \frac{x_i - \overline{x}}{S_{xx}} \right\} y_i \\ &= \sum_{i=1}^n c_i y_i \qquad \text{where } c_i = \frac{x_i - \overline{x}}{S_{xx}} \end{aligned}$$

and hence

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = \frac{1}{n} \sum_{i=1}^n y_i - \overline{x} \sum_{i=1}^n c_i y_i = \sum_{i=1}^n \left\{ \frac{1}{n} - \overline{x} c_i \right\} y_i.$$

■

It is noteworthy that we have also deduced $\sum_{i=1}^n (x_i - \overline{x}) = 0$ in the middle of the proof, and similarly, we can deduce that $\sum_{i=1}^n (y_i - \overline{y}) = 0$. These facts will often be used in our proofs.

Using the linear property of $\hat{\beta}_i$ as well as the aforementioned fact, we can show that

Theorem 1.2. $\hat{\beta}_0$ and $\hat{\beta}_1$ are the unbiased estimators of the parameters β_0 and β_1 respectively.

Proof. First, by linearity, we can observe that

$$\begin{aligned} \mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\sum_{i=1}^n c_i y_i \right] &= \sum_{i=1}^n c_i \mathbb{E}[y_i] \\ &= \sum_{i=1}^n c_i \mathbb{E}[\beta_0 + \beta_1 x_i + \varepsilon_i] \\ &= \sum_{i=1}^n c_i \left\{ \mathbb{E}[\beta_0 + \beta_1 x_i] + \mathbb{E}[\varepsilon_i] \right\} \\ &= \sum_{i=1}^n c_i (\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum_{i=1}^n c_i + \beta_1 \sum_{i=1}^n c_i x_i \end{aligned}$$

Our goal is to show $\mathbb{E}[\hat{\beta}_1] = \beta_1$, which is equivalent to showing that $\sum_{i=1}^{n} c_i = 0$ and $\sum_{i=1}^{n} c_i x_i = 1$.

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \frac{x_i - \overline{x}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{S_{xx}} = 0$$

where here we used the fact that $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$. On the other hand,

$$\begin{aligned}
\sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{n} c_i x_i - \overline{x}\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} c_i(x_i - \overline{x}) \\
&= \sum_{i=1}^{n} \left\{ \frac{x_i - \overline{x}}{S_{xx}} \right\} (x_i - \overline{x}) \\
&= \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{S_{xx}} \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\
&= \frac{S_{xx}}{S_{xx}} = 1
\end{aligned}$$

Thus we have shown that $\mathbb{E}[\hat{\beta}_1] = \beta_1$, which means $\hat{\beta}_1$ is an unbiased estimator of β_1. We can deduce $\mathbb{E}[\hat{\beta}_0] = \beta_0$ in a similar way as follows. Recall

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left\{ \frac{1}{n} - \overline{x} c_i \right\} y_i := \sum_{i=1}^{n} d_i y_i \qquad \text{where } d_n := \frac{1}{n} - \overline{x} c_i.$$

Then we can derive that

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_0] = \mathbb{E}\left[\sum_{i=1}^{n} d_i y_i \right] = \sum_{i=1}^{n} d_i \mathbb{E}[y_i] \\
&= \sum_{i=1}^{n} d_i \mathbb{E}[\beta_0 + \beta_1 x_i + \varepsilon_i] \\
&= \sum_{i=1}^{n} d_i \left\{ \mathbb{E}[\beta_0 + \beta_1 x_i] + \mathbb{E}[\varepsilon_i] \right\} \\
&= \sum_{i=1}^{n} d_i (\beta_0 + \beta_1 x_i) \\
&= \beta_0 \sum_{i=1}^{n} d_i + \beta_1 \sum_{i=1}^{n} d_i x_i
\end{aligned}$$

and so here we need to show $\sum_{i=1}^{n} d_i = 1$ and $\sum_{i=1}^{n} d_i x_i = 0$.

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} \left\{ \frac{1}{n} - \overline{x} c_i \right\} = 1 - \overline{x} \sum_{i=1}^{n} c_i = 1$$

$$\sum_{i=1}^{n} d_i x_i = \sum_{i=1}^{n} \left\{ \frac{1}{n} - \overline{x} c_i \right\} x_i = \frac{1}{n} \sum_{i=1}^{n} x_i - \overline{x} \sum_{i=1}^{n} c_i x_i = \overline{x} - \overline{x}(1) = 0$$

and we have already completed the proof that $\mathbb{E}[\hat{\beta}_0] = \beta_0$. ■

We can of course compute $\mathbb{V}(\hat{\beta}_i)$ in a similar fashion. Recall the following formulas: if $T = \sum_{i=1}^{n} a_i X_i$ and $W = \sum_{j=1}^{m} b_j Y_j$, then

$$\operatorname{Cov}(T, W) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \operatorname{Cov}(X_i, Y_j),$$

moreover,

$$\mathbb{V}(T) = \operatorname{Cov}(T, T) = \sum_{i=1}^{n} a_i^2 \mathbb{V}(X_i) + 2 \sum_{i<j} a_i a_j \operatorname{Cov}(X_i, X_j).$$

Proposition 1.3. $\mathbb{V}(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$ and $\mathbb{V}(\hat{\beta}_0) = \sigma^2 \left\{ \dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}} \right\}$

Proof. Recall that

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$$

Furthermore, as $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, we know $\operatorname{Cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j$. Thus for $i \neq j$, we have

$$\begin{aligned} \operatorname{Cov}(y_i, y_j) &= \operatorname{Cov}(\beta_0 + \beta_1 x_i + \varepsilon_i, \beta_0 + \beta_1 x_j + \varepsilon_j) \\ &= \operatorname{Cov}(\varepsilon_i, \varepsilon_j) = 0 \end{aligned}$$

because $\beta_0 + \beta_1 x_i$ and $\beta_0 + \beta_1 x_j$ are constants. Hence we can compute that

$$\begin{aligned} \mathbb{V}(\hat{\beta}_1) = \sum_{i=1}^{n} c_i^2 \mathbb{V}(y_i) &= \sigma^2 \sum_{i=1}^{n} c_i^2 \\ &= \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{S_{xx}^2} \\ &= \frac{\sigma^2}{S_{xx}^2} \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ &= \frac{\sigma^2}{S_{xx}^2} S_{xx} \\ &= \frac{\sigma^2}{S_{xx}} \end{aligned}$$

Now recall that $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$, and so

$$\mathbb{V}(\hat{\beta}_0) = \mathbb{V}(\overline{y} - \hat{\beta}_1 \overline{x}) = \mathbb{V}(\overline{y}) + \overline{x}^2 \mathbb{V}(\hat{\beta}_1) + 2\overline{x} \operatorname{Cov}(\overline{y}, \hat{\beta}_1)$$

Beware that $\overline{x}$ is a fixed constant. Next, note that $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, so

$$\mathbb{V}(\overline{y}) = \mathbb{V}\left(\frac{1}{n} \sum_{i=1}^{n} y_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}(y_i) = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}.$$

On the other hand, we can compute that

$$\begin{aligned} \operatorname{Cov}(\overline{y}, \hat{\beta}_1) = \operatorname{Cov}\left(\frac{1}{n} \sum_{i=1}^{n} y_i, \sum_{j=1}^{n} c_j y_j \right) &= \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{c_j}{n} \operatorname{Cov}(y_i, y_j) \\ &= \sum_{i=1}^{n} \frac{c_i}{n} \mathbb{V}(y_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} c_i \sigma^2 \\ &= \frac{\sigma^2}{n} \sum_{i=1}^{n} c_i = 0 \end{aligned}$$

Putting these ingredients back, we have

$$\mathbb{V}(\hat{\beta}_0) = \mathbb{V}(\overline{y}) + \overline{x}^2 \mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{n} + \overline{x}^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left\{ \frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right\}.$$

■

Remark. Note that y_i's are not iid, so the proof are not as clean as we expect!

In the followng we will state a theorem regarding the quality of the least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$:

Theorem 1.4 (Gauss-Markov theorem)**.** The least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and have minimum variance among all unbiased linear estimators.

We will prove it in a later section. A fun fact is that we often call these least square estimators **BLUE** - best unbiased linear estimator. By 'best', we mean they have the minimum variance. We will revisit this point in a greater generality later.

There are many other useful properties about the fitted simple linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ that are noteworthy:

(1) The sum of residuals $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$ vanishes, i.e.

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

Proof. Recall $S = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$, so

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2 \sum_{i=1}^{n} e_i.$$

Recall the least square condition $\dfrac{\partial S}{\partial \hat{\beta}_0} = 0$, so

$$\sum_{i=1}^{n} e_i = \frac{1}{-2} \frac{\partial S}{\partial \hat{\beta}_0} = 0$$

■

(2) The sum of observed values y_i and the sum of the fitted values $\hat{y}_i$ are equal, i.e.

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i.$$

This is immediate from property (1).

(3) The fitted model always pass through the **centroid** $(\overline{x}, \overline{y})$, i.e.

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}.$$

This is the geometric meaning of our first normal equation.

(4) $\sum_{i=1}^{n} x_i e_i = 0$

Proof. Recall $S = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$, so

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = -2 \sum_{i=1}^{n} x_i e_i.$$

Recall the least square condition $\dfrac{\partial S}{\partial \hat{\beta}_1} = 0$, so

$$\sum_{i=1}^{n} x_i e_i = \frac{1}{-2} \frac{\partial S}{\partial \hat{\beta}_1} = 0$$

■

(5) $\sum_{i=1}^{n} \hat{y}_i e_i = 0$

Proof. $\sum_{i=1}^{n} \hat{y}_i e_i = \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i) e_i = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i e_i = 0$ ■