Methods of Data Analysis I

Derek Li

Contents

1 Review

1.1 Expectation

- $\mathbb{E}[a] = a, a \in \mathbb{R}$.
- $\mathbb{E}[aY] = a\mathbb{E}[Y]$.
- $\mathbb{E}[X \pm Y] = \mathbb{E}[x] \pm \mathbb{E}[Y]$.
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ if X and Y are independent.
- Tower rule: $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$.

1.2 Variance and Covariance

- $\mathrm{Var}[a] = 0, a \in \mathbb{R}$.
- $\mathrm{Var}[aY] = a^2\mathrm{Var}[Y]$.
- $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
- $\mathrm{Cov}(Y, Y) = \mathrm{Var}[Y]$.
- $\mathrm{Var}[Y] = \mathrm{Var}[\mathbb{E}[Y|X]] + \mathbb{E}[\mathrm{Var}[Y|X]]$.
- $\mathrm{Var}[X \pm Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] \pm 2\mathrm{Cov}(X, Y)$.
- $\mathrm{Cov}(X, Y) = 0$ if X and Y are independent.
- $\mathrm{Cov}(aX+bY, cU+dW) = ac\mathrm{Cov}(X, U)+ad\mathrm{Cov}(X, W)+bc\mathrm{Cov}(Y, U)+bd\mathrm{Cov}(Y, W)$.
- Correlation:
$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}.$$

2 Sample Linear Regression

2.1 Statistical Model

$$Y = \beta_0 + \beta_1 X + e,$$

where Y is dependent or response variable, X is independent or explanatory variable, β_0 is intercept parameter, β_1 is slope parameter, and e is random error or noise (variation in measures that we cannot account for).

Given a specific value of $X = x$, we want to find the expected value of Y

$$\mathbb{E}[Y|X = x].$$

2.2 Estimating β_0, β_1

Given n pairs bivariate data $(x_1, y_1), \cdots, (x_n, y_n)$, we want to use $\widehat{\beta}_0$ and $\widehat{\beta}$ to estimate β_0 and β_1.

Consider the residual sum of squares (RSS)

$$\text{RSS} = \sum_{i=1}^{n} \widehat{e}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i \right) \right]^2,$$

we can use least squares method that minimizes the criterion RSS to find the estimators.

2.2.1 Least Squares Method

Least squares method makes no statistical assumptions. We have

$$\frac{\partial \text{RSS}}{\partial \widehat{\beta}_0} = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) \text{ and } \frac{\partial \text{RSS}}{\partial \widehat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) x_i.$$

Let $\frac{\partial \text{RSS}}{\partial \widehat{\beta}_0}$ and $\frac{\partial \text{RSS}}{\partial \widehat{\beta}_1}$ be 0, we get the normal equations

$$\sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) = 0 \text{ and } \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right) x_i = 0.$$

Therefore, we have

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \widehat{\beta}_0 - \sum_{i=1}^{n} \widehat{\beta}_1 x_i = n\overline{y} - n\widehat{\beta}_0 - n\widehat{\beta}_1\overline{x} = 0 \Rightarrow \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}.$$

Besides,

$$\begin{aligned}\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \widehat{\beta}_0 x_i - \sum_{i=1}^{n} \widehat{\beta}_1 x_i^2 &= \sum_{i=1}^{n} x_i y_i - \left(\overline{y} - \widehat{\beta}_1\overline{x}\right)\sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} x_i y_i - n\overline{xy} + n\widehat{\beta}_1\overline{x}^2 - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0,\end{aligned}$$

i.e.,

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} := \frac{SXY}{SXX}.$$

2.2.2 Interpretation

$\widehat{\beta}_0$: The expected value of y when $x = 0$. No practical interpretation unless 0 is within the range of the predictor values.

$\widehat{\beta}_1$: When x changes by 1 unit, the corresponding average change in y is the slope.

2.3 Properties of Fitted Regression Line

Property 2.1.

$$\sum_{i=1}^{n} \widehat{e}_i = 0.$$

Proof. By definition,

$$\begin{aligned}\sum_{i=1}^{n} \widehat{e}_i &= \sum_{i=1}^{n}(y_i - \widehat{y}_i) = \sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) = \sum_{i=1}^{n}\left(y_i - \overline{y} + \widehat{\beta}_1\overline{x} - \widehat{\beta}_1 x_i\right) \\ &= n\overline{y} - n\overline{y} + n\widehat{\beta}_1\overline{x} - n\widehat{\beta}_1\overline{x} = 0.\end{aligned}$$

□

Property 2.2. The sum of squares of residuals is not 0 unless the fit to the data is perfect.

Property 2.3.

$$\sum_{i=1}^{n} \widehat{e}_i x_i = 0.$$

Proof. By definition,

$$\begin{aligned}\sum_{i=1}^{n} \widehat{e}_i x_i &= \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) x_i = \sum_{i=1}^{n} x_i y_i - \overline{y} \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \overline{x} \sum_{i=1}^{n} x_i - \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 \\ &= \sum_{i=1}^{n} x_i y_i - n\overline{xy} - \widehat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = 0.\end{aligned}$$

□

Property 2.4.

$$\sum_{i=1}^{n} \widehat{e}_i \widehat{y}_i = 0.$$

Proof. By definition,

$$\begin{aligned}\sum_{i=1}^{n} \widehat{e}_i \widehat{y}_i &= \sum_{i=1}^{n} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) \\ &= \sum_{i=1}^{n} \left(\overline{y} - \widehat{\beta}_1 \overline{x}\right) y_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \left(\overline{y} - \widehat{\beta}_1 \overline{x}\right)^2 - 2\left(\overline{y} - \widehat{\beta}_1 \overline{x}\right) \widehat{\beta}_1 \sum_{i=1}^{n} x_i - \widehat{\beta}_1^2 \sum_{i=1}^{n} x_i^2 \\ &= n\overline{y}^2 - n\widehat{\beta}_1 \overline{xy} + \widehat{\beta}_1 \sum_{i=1}^{n} x_i y_i - n\overline{y}^2 + 2n\widehat{\beta}_1 \overline{xy} - n\widehat{\beta}_1^2 \overline{x}^2 - 2n\widehat{\beta}_1 \overline{xy} + 2n\widehat{\beta}_1^2 \overline{x}^2 - \widehat{\beta}_1^2 \sum_{i=1}^{n} x_i^2 \\ &= \widehat{\beta}_1 \left(\sum_{i=1}^{n} x_i y_u - n\overline{xy}\right) - \widehat{\beta}_1^2 \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) = 0.\end{aligned}$$

□

Property 2.5.

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

Proof. By definition,

$$\sum_{i=1}^{n} \widehat{y}_1 = \sum_{i=1}^{n} \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i\right) = \sum_{i=1}^{n} \left(\overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 x_i\right) = n\overline{y} = \sum_{i=1}^{n} y_i.$$

□

2.4 Assumptions

The Gauss-Markov conditions are:

1. $\mathbb{E}[e_i] = 0$.

2. $\mathrm{Var}[e_i] = \sigma^2$, i.e., homoscedastic.

3. The errors are uncorrelated or $\mathrm{Cov}(e_i, e_j) = \rho(e_i, e_j) = 0$.

Theorem 2.1 (Gauss-Markov Theorem)**.** Under the conditions or the simple linear regression model, the least-squares parameter estimators are best linear unbiased estimators.

We assume that Y is relate to x by the simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, \cdots, n.$$

Under the conditions we have

$$\mathbb{E}[Y|X = x_i] = \beta_0 + \beta_1 x_i$$

and

$$\mathrm{Var}[Y|X = x_i] = \mathrm{Var}[\beta_0 + \beta_1 x_i + e_i|X = x_i] = \mathrm{Var}[e_i] = \sigma^2.$$

2.5 Estimating the Variance of the Random Error Term

The variance σ^2 is another parameter of the SLR model and we want to estimate σ^2 to measure the variability of our estimates of Y, and carry out inference on the model.

An unbiased estimate of σ^2 is

$$S^2 = \frac{\sum_{i=1}^{n} \widehat{e}_i^2}{n-2} = \frac{\mathrm{RSS}}{n-2}.$$

2.6 Properties of Least Squares Estimators

Since $\sum\limits_{i=1}^{n}(x_i - \overline{x}) = 0$,

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i - \overline{y}\sum_{i=1}^{n}(x_i - \overline{x}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i.$$

Let $c_i = \frac{x_i - \overline{x}}{SXX}$, we can rewrite $\widehat{\beta}_1$ as

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

which is a linear combination of y_i.

We have

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta}_1 | X\right] &= \mathbb{E}\left[\sum_{i=1}^{n} c_i y_i | X = x_i\right] = \sum_{i=1}^{n} c_i \mathbb{E}[y_i | X = x_i] \\
&= \sum_{i=1}^{n} c_i \mathbb{E}[\beta_0 + \beta_1 x_i] = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i \\
&= \frac{\beta_0}{SXX}\sum_{i=1}^{n}(x_i - \overline{x}) + \beta_1 \sum_{i=1}^{n}\frac{(x_i - \overline{x})x_i}{SXX} \\
&= \beta_1 \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{SXX} = \beta_1.
\end{aligned}$$

Therefore, $\widehat{\beta}_1$ is unbiased for β_1. Besides,

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\beta}_1 | X\right] &= \mathrm{Var}\left[\sum_{i=1}^{n} c_i y_i | X\right] = \sum_{i=1}^{n} c_i^2 \mathrm{Var}[y_i | X = x_i] \\
&= \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \frac{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2}{SXX^2} = \frac{\sigma^2}{SXX}.
\end{aligned}$$