§2. Simple Linear Regression

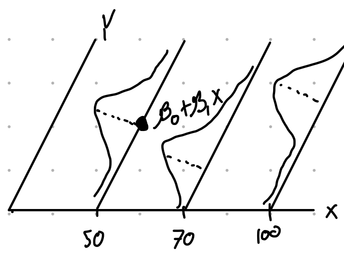
1. Introduction

$E(Y|X=x)=\beta_0+\beta_1 x$ $\quad Var(Y|X=x)=\sigma^2$

$Y=\beta_0+\beta_1 x+\varepsilon$, β_0 - intercept, β_1 - slope, ε - random error

① There is a probability distribution of Y, for each level of X.

② The means of these distributions vary, but the variances stay the same with different levels of x.



sample regression model $Y_i=\beta_0+\beta_1 x_i+\varepsilon_i$

Y_i – the value of the response (random variable)

x_i – the predictor

ε_i – random error with $E(\varepsilon_i)=0$

$Var(\varepsilon_i)=\sigma^2$, ε_i and ε_j are uncorrelated so that

$Cov(\varepsilon_i,\varepsilon_j)=0$, $i\neq j$ $\left(Cov(X,Y)=E[(X-\mu_X)(Y-\mu_Y)] \text{ or } Cov(X,Y)=E(XY)-E(X)E(Y) \right)$

Important features of model.

① Y_i is a random variable $Y_i=\beta_0+\beta_1 X_i+\varepsilon_i$

② Since $E(\varepsilon_i)=0$, so $E(Y_i)=\beta_0+\beta_1 x_i$. Thus, the response Y_i when the level of X in the i^{th} trial is x_i, comes from a distribution whose mean is $\beta_0+\beta_1 x_i$.

③ The response Y_i in the i^{th} trial exceeds or falls short of the value of the regression function by the error term ε_i.

④ The error term ε_i are assumed to have constant variance σ^2. Therefore, $Var(Y_i) = \sigma^2$.

⑤ Slope β_1: Change in the mean of Y for one unit change in X.

Intercept β_0: The mean of Y when $X=0$.

2. OLS Estimation

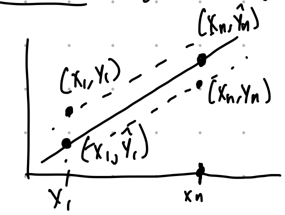
Consider we have a data set with n pairs

$(X_1, Y_1), (X_2, Y_2), \dots (X_n, Y_n)$

Parameters to be estimated: $\beta_0, \beta_1, \sigma^2$

Consider $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Define $e_i = Y_i - \hat{Y}_i$ residuals



minimize $Q = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$Q = \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2$$

$$= \sum_{i=1}^{n} [Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i]^2$$

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = 0 \Leftrightarrow -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = 0 \Leftrightarrow -2 \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 = \sum_{i=1}^{n} (X_i Y_i) - \dfrac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n} \cdot \left(\dfrac{1}{\sum_{i=1}^{n} X_i^2 - \dfrac{(\sum_{i=1}^{n} X_i)^2}{n}} \right) \end{cases}$$

Define $S_{XY} = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$

$$S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}\left[x_iy_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}\right]$$

$$= \sum_{i=1}^{n} x_iy_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y}$$

$$= \sum_{i=1}^{n} x_iy_i - \bar{y}\cdot n\bar{x} - \bar{x}\cdot n\bar{y} + n\bar{x}\bar{y}$$

$$= \sum_{i=1}^{n} x_iy_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$$

$$S_{xx} = \sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\cdot n\bar{x} + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

The ordinary least square (OLS) Estimation is: $\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \\ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \end{cases}$

and the fitted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Estimation of σ^2 since $\varepsilon_i \sim N(0,\sigma^2)$, residual sum of squares (RSS)

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

$$RSS = \sum_{i=1}^{n}\left[y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right]^2$$

$$= \sum_{i=1}^{n}\left[y_i - (\bar{y} - \hat{\beta}_1\bar{x}) - \hat{\beta}_1 x_i\right]^2$$

$$= \sum_{i=1}^{n}\left[(y_i-\bar{y}) - \hat{\beta}_1(x_i-\bar{x})\right]^2$$

$$= \sum_{i=1}^{n}(y_i-\bar{y})^2 + \hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2 - 2\hat{\beta}_1\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$= S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 S_{xy}$$

$$= S_{yy} + \frac{S_{xy}^2}{S_{xx}^2}S_{xx} - 2\frac{S_{xy}}{S_{xx}}\cdot S_{xy}$$

$$= S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

02/01/2022

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$