4. 最小二乘估计

$$\sum_{i=1}^{n}[Y_i-(\beta_0+\beta_1 x_i)]^2$$

a. 令 $\mu_i=\beta_0+\beta_1 x_i$，我们估计 $\hat{\mu}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$

$$L=\min\sum_{i=1}^{n}(Y_i-\mu_i)^2=\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)^2+2\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)(\hat{\mu}_i-\mu_i)+\sum_{i=1}^{n}(\hat{\mu}_i-\mu_i)^2$$

假设 $L=\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)(\hat{\mu}_i-\mu_i)=0$

then. $L=\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)^2+\sum_{i=1}^{n}(\hat{\mu}_i-\mu_i)^2\geq\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)^2$

因此，若 $\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)(\hat{\mu}_i-\mu_i)=0$， $\mu_i=\beta_0+\beta_1 x_i$, $\hat{\mu}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$

$$Y=\hat{\beta}_0+\hat{\beta}_1 X$$

水平

1° 令 $\beta_1=0$ 则 $\hat{\beta}_1=0$， $\hat{\beta}_0=\bar{Y}$

$$\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)(\hat{\mu}_i-\mu_i)=\sum_{i=1}^{n}(Y_i-\hat{\beta}_0)(\hat{\beta}_0-\beta_0)$$

$$=(\hat{\beta}_0-\beta_0)\sum_{i=1}^{n}(Y_i-\hat{\beta}_0)$$

上式 $=0$，计 $\sum_{i=1}^{n}(Y_i-\hat{\beta}_0)=n\bar{Y}-n\hat{\beta}_0=0$ $\quad\hat{\beta}_0=\bar{Y}$.

2° 令 $\hat{\beta}_0=0$，$\beta_0=0$ 经过原点

$$\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)(\hat{\mu}_i-\mu_i)=\sum_{i=1}^{n}(Y_i-\hat{\beta}_1 x_i)(\hat{\beta}_1 x_i-\beta_1 x_i)$$

$$=(\hat{\beta}_1-\beta_1)\sum_{i=1}^{n}(Y_iX_i-\hat{\beta}_1X_i^2)$$

3.6

上式 $=0$，计 $\sum_{i=1}^{n}(Y_iX_i-\hat{\beta}_1X_i^2)=\sum_{i=1}^{n}Y_iX_i-\hat{\beta}_1\sum_{i=1}^{n}X_i^2=0$

then $\hat{\beta}_1=\dfrac{\sum_{i=1}^{n}Y_iX_i}{\sum_{i=1}^{n}X_i^2}$.

3°. 一般 $\mu_i = \beta_1 x_i + \beta_0$

$$\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(\hat{\beta}_0 + \hat{\beta}_1 x_i - \beta_0 - \beta_1 x_i)$$

$$= (\hat{\beta}_0 - \beta_0)\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) + (\beta_1 - \beta_1)\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)X_i$$

$$0 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = n\bar{Y} - n\hat{\beta}_0 - n\hat{\beta}_1\bar{X} \Rightarrow \hat{\beta}_0 = \bar{Y} - \beta_1\bar{X}$$

then

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i = \sum_{i=1}^{n}(Y_i - \bar{Y} + \hat{\beta}_1\bar{X} - \hat{\beta}_1 X_i)X_i$$

$$= \sum_{i=1}^{n}\{(Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})\}X_i$$

thus.

$$\sum_{i=1}^{n}(Y_i - \bar{Y})X_i - \hat{\beta}_1 \sum_{i=1}^{n}(X_i - \bar{X})X_i = 0$$

则有

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})X_i}{\sum_{i=1}^{n}(X_i - \bar{X})X_i} = \frac{\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})(X_i - \bar{X})} = cor(Y, X)\frac{SD(X)}{SD(Y)}$$

因为 $\hat{\beta}_0 = Y - \hat{\beta}_1 X_i$

结论:

1. $\hat{\beta}_1 = cor(Y, X)\frac{SD(Y)}{SD(X)}$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$

经过点 $(\bar{X}, \bar{Y})$

center data后线段slope不变.

if normalize $\left\{\frac{X_i - \bar{X}}{SD(X)}, \frac{Y_i - \bar{Y}}{SD(Y)}\right\}$, slope is $cor(Y, X)$.