Regression Analysis
Simple Linear Regression

Buddhananda Banerjee

Department of Mathematics
Centre for Excellence in Artificial Intelligence
Indian Institute of Technology Kharagpur

bbanerjee@maths.iitkgp.ac.in

Simple linear regression

- Consider a data set $D = \{(x_i, y_i) | x_i \in \mathbb{R}, y_i \in \mathbb{R}, \ \forall i = 1, 2, \cdots, n\}$
- x_is are non stochastic
- y_is are stochastic and realized values of random variable Y_is
- ϵ_is are iid $N(0, \sigma^2)$ random variables
- Regression parameter (β_0, β_1) and error parameter σ^2 are unknown

Problem statement

Considering the linear model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ \forall \ i = 1, 2, \ldots n$$

we want to estimate (β_0, β_1) which will minimize least squared condition.

Least square estimate

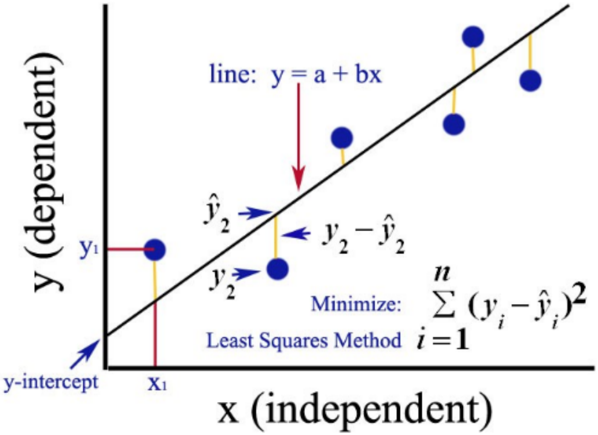

Figure: Least square

Least Squared condition

The least squared condition to estimate the model parameters is to minimize

$$S(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2. \tag{1}$$

with respect to β_0 andβ_1.

If $(\hat{\beta}_0, \hat{\beta}_1)$ minimizes $S(\beta_0, \beta_1)$ then their values can be obtained by solving the **normal equations**

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0 \implies n\hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i = \sum_i y_i \tag{2}$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0 \implies \hat{\beta}_0 \sum_i x_i + \hat{\beta}_1 \sum_i x_i^2 = \sum_i y_i x_i \tag{3}$$

Prediction

Estimated parameters

Defining $S_{xy} = \sum_i (y_i - \bar{y})(x_i - \bar{x})$ we have the solutions as

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Least squared prediction line

For any x such as old x_is or some x_{new} the prediction line is

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

Prediction Error

The **prediction error or residual** is defined as $e_i = y_i - \hat{y}_i$ and hence the residual sum of square (SSR) is

$$\begin{aligned}
\sum_i e_i^2 &= \sum_i (y_i - \hat{y}_i)^2 \\
&= \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= \sum_i [(y_i - (\bar{y} - \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)]^2 \\
&= \sum_i [(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})]^2 \\
&= S_{yy} + \hat{\beta}_1^2 S_{xx} - 2\hat{\beta}_1 S_{xy} \\
&= S_{yy} - \hat{\beta}_1 S_{xy} \\
&= S_{yy} - \frac{S_{xy}^2}{S_{xx}}
\end{aligned} \tag{4}$$

NOTE: We estimate σ^2 by $\hat{\sigma}^2 = \frac{SSR}{n-2} \equiv MSR$, mean residual sum of square. [Explanation will be studied under Multiple Linear Regression]

Prediction Error

- We estimate σ^2 by $\hat{\sigma}^2 = \frac{SSR}{n-2} \equiv MSR$, mean residual sum of square. [Explanation will be studied under Multiple Linear Regression]
- We can estimate σ^2 even without computing individual errors e_i.
- Prediction error $\mathbf{e} = (e_1, e_2, \ldots, e_n)$ and prediction vector $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ are independently distributed.

Linear estimator

If an estimator $T(\mathbf{y})$ can be expressed as a linear combination of $\mathbf{y}$ with non random coefficients i.e. $T(\mathbf{y}) = \sum_i \alpha_i y_i$ then $T(\mathbf{y})$ is called a linear estimator.

Note

1. $\hat{\beta}_1$ is a linear estimator of β_1. and $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$

2. $\hat{\beta}_0$ is a linear estimator of β_0 and $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$

3. Regression line passes through $(\bar{x}, \bar{y})$

4. $SSR \sim \sigma^2 \chi^2_{n-2}$

What more?

- Can we test for $H_0 : \beta_0 = b_0$ vs $H_1 : \beta_0 \neq b_0$?
- Can we test for $H_0 : \beta_1 = b_1$ vs $H_1 : \beta_1 \neq b_1$?
- Can we test for $H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 \neq \sigma_0^2$?
- Can we have a prediction interval for $\hat{y}$?

YES !!!! Now we can...

Bivariate normal$(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$

If (X, Y) has joint density function $f(x, y)$ as

$$f(x,y) = \frac{e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2+\left(\frac{y-\mu_y}{\sigma_y}\right)^2-2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]}}{2\pi\sigma_x, \sigma_y\sqrt{1-\rho^2}}$$

then (X, Y) is said to follow Bivariate normal.

- If (X, Y) follow Bivariate normal$(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ the show that $Y|X = x$ follows $N(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), (1 - \rho^2)\sigma_y^2)$
- Perform a large and small sample test for $H_0 : \rho = 0$ Vs $H_1 : \rho \neq 0$.
- Perform a large sample test for $H_0 : \rho = \rho_0$ Vs $H_1 : \rho \neq \rho_0$.


THANK
YOU