Packet 5: Linear Regression

Chap 6.5 Linear Regression Model

Least Square method does not use any distn assumption

From mathematics to statistics: Up to this point our construction of the least square estimators is completely out of mathematical intuition. Without a probabilistic model (i.e. a set of distributional assumptions), however, we cannot evaluate the statistical properties of these estimators. Under the sampling view, what happens if we repeat the experiment many many times? In particular, we cannot address pertinent questions such as the reliability of the slope and intercept estimates in the presence of "noise". Next, let us introduce some assumptions, and evaluate the properties of the estimators accordingly.

The simplest statistical model for the data pairs (x_i, y_i) is the linear regression model.

$$Y_i = \underbrace{\alpha + \beta(x_i - \bar{x})}_{\text{expected}} + \underbrace{\epsilon_i}_{\text{randomness}}, \quad i = 1, 2, \ldots, n.$$

where the ϵ_i are i.i.d. random errors with zero mean and common variance

$$\epsilon_i \sim N(0, \sigma^2), \quad \text{for } i = 1, 2, \ldots, n.$$

Let us assume that the x's are fixed (not random).

It is important to note the set of assumptions we are making here, i.e., LINE.

Linear : the relationship between X and Y is linear, $f(x) = \alpha + \beta(x_i - \bar{x})$

Independence : ε_i's are independent, Y_i's are independent

Normality : $\varepsilon_i \sim$ Norm

Equal Variance : $Var(\varepsilon_i) = \sigma^2$

Under LINE Assumption

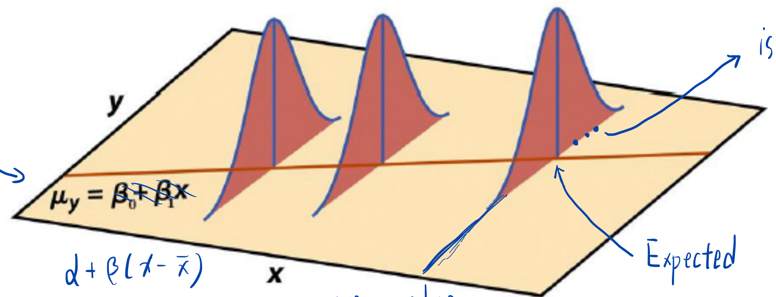
With the above assumptions, let us check properties of the least square estimators. The error terms ϵ_i's are i.i.d. Normal$(0, \sigma^2)$ r.v.s, i.e.,

$Y_i \sim N(\underline{\alpha + \beta(x_i - \bar{x})}, \sigma^2)$

$$Y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

observed value of y is a realization of $Y \sim N(\alpha + \beta(x - \bar{x}), \sigma^2)$

$E(Y_i \mid x_i)$ regression line



$\alpha + \beta(x - \bar{x})$

Expected

given value of x

In this case, let us find the maximum likelihood estimators, $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$.

defination of Norm distn, p.d.f. of Y_i $\quad f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \alpha - \beta(x_i - \bar{x}))^2}{2\sigma^2}\right)$

$$L(\alpha, \beta, \sigma^2) = \prod_{i=1}^{n} f(y_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{(y_i - \alpha - \beta(x_i - \bar{x}))^2}{2\sigma^2}\right)$$

$$= 2\pi^{-\frac{n}{2}} \times (\sigma^2)^{-\frac{n}{2}} \times \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta(x_i - \bar{x}))^2\right]$$

The likelihood function is

$$L(\alpha, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\sum_{i=1}^{n}(y_i - \alpha - \beta(x_i - \bar{x}))^2}{2\sigma^2}}.$$

$$\log L(\alpha, \beta, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\boxed{\sum_{i=1}^{n}(y_i - \alpha - \beta(x_i - \bar{x}))^2}$$

sum of squared errors

$S(\alpha, \beta)$

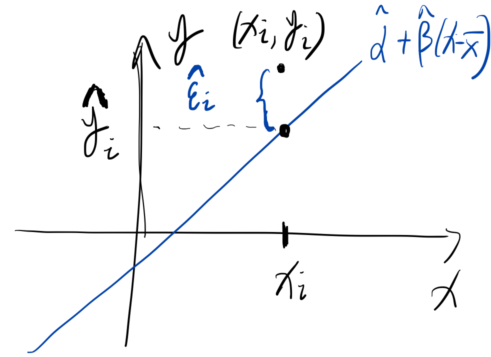
$\because -\frac{1}{2\sigma^2} < 0$ larger $S(\alpha, \beta) \rightarrow$ smaller $\log L(\alpha, \beta, \sigma^2)$

Maximizing this log-likelihood with respect to α and β is the same as minimizing $S(\alpha, \beta)$

$$\sum_{i=1}^{n}(y_i - \alpha - \beta(x_i - \bar{x}))^2.$$

More Examples: 6.5-1

for any value of σ^2, minimizing $S(\alpha, \beta)$ will lead to max Likelihood



So, the least square solutions must be the MLEs.

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{y}.$$

The MLE of σ^2 can be derived by taking a partial derivative on the log-likelihood with respect to σ^2.

$$\frac{\partial \log L(\alpha, \beta, \sigma^2)}{\partial \sigma^2} = 0 - \frac{n}{2}\frac{1}{\sigma^2} - \frac{1}{2} S(\alpha, \beta) \times \left(-\frac{1}{\sigma^4}\right)$$

$$= \frac{1}{2\sigma^4} S(\alpha, \beta) - \frac{n}{2\sigma^2} = 0$$

multiply $\frac{2\sigma^4}{n}$ on both sides

for any (α, β) this σ^2 maximizes likelihood.

$$\sigma^2 = \frac{S(\alpha, \beta)}{n}$$

plug in $\hat{\alpha}$, $\hat{\beta}$ (MLE)

$$\hat{\sigma}^2 = \frac{S(\hat{\alpha}, \hat{\beta})}{n} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})\right)^2 \text{ (MLE)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

Fitted value $\hat{y}_i = \hat{\alpha} + \hat{\beta}(x_i - \bar{x})$ Residuals $\hat{\varepsilon}_i = y_i - \hat{y}_i$

We note that the distribution of the residual sum of squares divided by σ^2 is χ^2_{n-2}, i.e.,

$$\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})\right)^2}{\sigma^2} \sim \chi^2(n-2)$$

# of free parameters besides σ^2

$\hat{\alpha}$, $\hat{\beta}$

Ex In Chap 5.5 Sample Variance

$$\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2(n-1)$$

we need to estimate the mean parameter by $\hat{\mu}_Y = \bar{Y}$ in order to estimate σ^2

$$E\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = n-2$$

$$E(\hat{\sigma}^2) = \frac{n-2}{n}\sigma^2 < \sigma^2$$

MLE $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ is a biased estimator for σ^2

$$\boxed{\text{Chap 5.3} \quad E\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i E(Y_i) \qquad Var\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i^2 Var(Y_i)}$$

* independence

Estimator properties: $\hat{\alpha}$ and $\hat{\beta}$ are both linear functions of random variables, Y_i's.

$E(Y_i) = \alpha + \beta(x_i - \bar{x}) \qquad Var(Y_i) = \sigma^2 \qquad$ normality

$$\hat{\alpha} = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$E(\hat{\alpha}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\left(\alpha + \beta(x_i - \bar{x})\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\alpha + \beta \boxed{\sum_{i=1}^{n}(x_i - \bar{x})} = 0$$

$$= \alpha \quad \text{unbiased}$$

$$Var(\hat{\alpha}) = \sum_{i=1}^{n} \frac{1}{n^2} Var(Y_i)$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

$$\hat{\alpha} \sim N\left(\alpha, \frac{\sigma^2}{n}\right)$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} \boxed{(x_i - \bar{x})Y_i} - \underbrace{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})}}_{=0}\bar{Y}}{S_{xx}} = \sum_{i=1}^{n} c_i Y_i$$

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

$$E(\hat{\beta}) = \sum_{i=1}^{n} c_i E(Y_i) = \sum_{i=1}^{n}\left[\frac{(x_i - \bar{x})}{S_{xx}}\left[\alpha + \beta(x_i - \bar{x})\right]\right]$$

$$= \alpha \underbrace{\boxed{\sum_{i=1}^{n}\frac{(x_i - \bar{x})}{S_{xx}}}}_{=0} + \beta \sum_{i=1}^{n}\frac{\boxed{(x_i - \bar{x})(x_i - \bar{x})}}{S_{xx}} \qquad = S_{xx}$$

$$= \beta \quad \text{unbiased}$$

$$Var(\hat{\beta}) = \sum_{i=1}^{n} c_i^2 Var(Y_i) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2 \times \sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{S_{xx}^2}$$

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

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$