ST102 Week 20

Simple Linear Regression

Problem settings:

Given paired observations $\{(x_i, y_i)\}_{i=1}^{n}$ from model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

with $E(\varepsilon_i) = 0$, $\forall i \in [n]$ and $Var(\varepsilon_i) = \sigma^2 > 0$, $\forall i \in [n]$.
Also assume $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.

Here we focus on "fixed design" (then what is a "random design"?) $\{X_i\}_1^n$ v.s. $\{x_i\}_1^n$.

Parameters to understand: $\beta_0, \beta_1, \sigma^2$.

Fact: 1) $E(y_i) = \beta_0 + \beta_1 x_i$, $Var(y_i) = \sigma^2$, and all y_i's are uncorrelated.

2) If $\varepsilon_i \sim N(0, \sigma^2) \Rightarrow y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, and y_i's are independent.

<u>LSE of parameters</u>

Define the loss function:

$$\mathcal{L}(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Then let's find the minima of $\mathcal{L}$:

$$\frac{\partial}{\partial \beta_0} \mathcal{L}(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \qquad (1)$$

$$\frac{\partial}{\partial \beta_1} \mathcal{L}(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) \qquad (2)$$

$$\begin{cases} (1) = 0 \\ (2) = 0 \end{cases} \Rightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

What about the estimator of σ^2:

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

<u>Properties of estimators</u>

1) $E(\hat{\beta}_0) = \beta_0$

$$Var(\hat{\beta}_0) = \frac{\sigma^2}{n} \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

2) $E(\hat{\beta}_1) = \beta_1$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Inference for parameters in the case of Normal

Further assume: $\varepsilon_i \underset{i.i.d.}{\sim} N(0, \sigma^2)$.

$\Rightarrow$ $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and y_i's independent.

In addition, $$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

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

The estimated standard errors:

$$E.S.E.(\hat{\beta}_0) = \frac{\hat{\sigma}}{\sqrt{n}} \left(\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)^{\frac{1}{2}}$$

$$E.S.E.(\hat{\beta}_1) = \frac{\hat{\sigma}}{\left[\sum_{i=1}^n (x_i - \bar{x})^2\right]^{\frac{1}{2}}}$$

Lemma. 1) $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$

2) $\hat{\beta}_0 \perp\!\!\!\perp \hat{\sigma}^2$, and $\dfrac{\hat{\beta}_0 - \beta_0}{E.S.E.(\hat{\beta}_0)} \sim t_{n-2}$

(continued)

3) $\hat{\beta}_1 \perp\!\!\!\perp \hat{\sigma}^2$, and $\dfrac{\hat{\beta}_1 - \beta_1}{E.S.E.(\hat{\beta}_1)} \sim t_{n-2}$

Confidence Intervals

$(1-\alpha) \times 100\%$ confidence interval for β_0 is

$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2}, n-2} \cdot E.S.E.(\hat{\beta}_0)$$

$(1-\alpha) \times 100\%$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}, n-2} \cdot E.S.E.(\hat{\beta}_1)$$

Test the slope

$H_0: \beta_1 = b$ v.s. $H_1: \cdots$

$$T := \frac{\hat{\beta}_1 - b}{E.S.E.(\hat{\beta}_1)} \sim t_{n-2} \text{ under } H_0.$$

Verification by MLE

(still assume Normal)

$$\mathcal{L}(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

(continued)

Then we can further have the log-likelihood:

$$\ell(\beta_0, \beta_1, \sigma^2) = \frac{n}{2}\ln\left(\frac{1}{\sigma^2}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 + C.$$

Which part is flexible? How to maximize it?