ST102 Week 21

Linear Regression Part II

Notations:

1) Total SS $:= \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$

2) Regression SS $:= \sum_{i=1}^{n} \hat{\beta}_1^2 (x_i - \bar{x})^2 = \hat{\beta}_1^2 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right)$

3) Residual SS $:= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

Facts:

1) Total SS = Regression SS + Residual SS

2) If $\beta_1 = 0$ and assume $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, then:

a) $$\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sigma^2} \sim \chi^2_{n-1}$$

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

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

Test $H_0: \beta_1 = 0$ v.s. $H_1: \beta_1 \neq 0$

(continued)

We can then propose the test statistic F, s.t. under H_0, $F \sim F_{1,n-2}$, by:

$$F := \frac{\text{Regression SS}/1}{\text{Residual SS}/(n-2)} = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}} = \left[\frac{\hat{\beta}_1}{E.S.E.(\hat{\beta}_1)}\right]^2 \underset{H_0}{\sim} F_{1,n-2}$$

Reject H_0, at significance level $\alpha \cdot 100\%$, if $f > F_{\alpha;1,n-2}$.

Def. (Coefficient of determination)

$$R^2 := \frac{\text{Regression SS}}{\text{Total SS}} \in [0, 1]$$

Better explanatory power $\Leftrightarrow R^2 \nearrow 1$.

<u>Confidence Interval of $E(y)$</u>

Recall our fitted model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Now let's arbitrarily fix a know value of x. Then we are actually interested in 2 terms:

1) $\mu(x) := E(y) = \beta_0 + \beta_1 x$ (recall we assume $E(\varepsilon) = 0$)

$\Rightarrow$ the mean response value in underlying truth

2) $y \Rightarrow$ the realized value in one experiment

(continued)

To gain an interval estimator, we further assume $\varepsilon \sim N(0, \sigma^2)$ and denote $\hat{\mu}(x) := \hat{\beta}_0 + \hat{\beta}_1 x$.

Reusing the intermediate results last week we have:

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

which, after normalization, would be

$$\frac{\hat{\mu}(x) - \mu(x)}{\left[\frac{\sigma^2}{n} \cdot \frac{\sum_{i=1}^{n}(x_i - x)^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right]^{\frac{1}{2}}} \sim N(0,1)$$

while, in practice, using $\hat{\sigma}^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 / (n-2)$,

$$\frac{\hat{\mu}(x) - \mu(x)}{\left[\frac{\hat{\sigma}^2}{n} \cdot \frac{\sum_{i=1}^{n}(x_i - x)^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}\right]^{\frac{1}{2}}} \sim t_{n-2}$$

As a result, the $(1-\alpha)\cdot 100\%$ confidence interval for $\mu(x)$ is

$$\hat{\mu}(x) \pm t_{\frac{\alpha}{2}, n-2} \cdot \hat{\sigma} \cdot \left[\frac{\sum_{i=1}^{n}(x_i - x)^2}{n \cdot \sum_{j=1}^{n}(x_j - \bar{x})^2}\right]^{\frac{1}{2}}$$

Prediction Interval for y

By problem settings we know $y - \hat{\mu}(x) \sim N(0, \sigma_1^2)$, with

$$\sigma_1^2 = Var(y) + Var[\hat{\mu}(x)] = \sigma^2 + \frac{\sigma^2}{n} \frac{\sum_{i=1}^{n}(x_i - x)^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

It can then be shown that

$$\frac{y - \hat{\mu}(x)}{\left\{ \hat{\sigma}^2 \left[1 + \frac{\sum_{i=1}^{n}(x_i - x)^2}{n \cdot \sum_{j=1}^{n}(x_j - \bar{x})^2} \right] \right\}^{\frac{1}{2}}} \sim t_{n-2}$$

Then the $(1-\alpha)\cdot 100\%$ prediction interval for y is

$$\hat{\mu}(x) \pm t_{\frac{\alpha}{2}, n-2} \cdot \hat{\sigma} \cdot \left[1 + \frac{\sum_{i=1}^{n}(x_i - x)^2}{n \cdot \sum_{j=1}^{n}(x_j - \bar{x})^2} \right]^{\frac{1}{2}}$$

Multiple (multi-variate) Linear Regression

Given i.i.d. random sample $\{(y_i, x_{i1}, x_{i2}, \dots, x_{ip})\}_{i=1}^{n}$, collected from the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i$$

with $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2 > 0$, $Cov(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.

(continued)

For any fixed point $(x_{i1}, \cdots, x_{ip})$, we know

$$E(y_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad \text{and} \quad Var(y_i) = \sigma^2,$$

and all y_i's uncorrelated.

LSE can similarly be obtained by minizing

$$\sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2$$

$\Rightarrow$ model fitting : $\hat{y} = \hat{\beta_0} + \sum_{j=1}^{p} \hat{\beta_j} x_j$.

By denoting
$$\begin{cases} \text{Total SS} := \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ \text{Regression SS} := \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\ \text{Residual SS} := \sum_{i=1}^{n} (y_i - \hat{\beta_0} - \sum_{j=1}^{p} \hat{\beta_j} x_{ij})^2 \end{cases},$$

we still have the decomposition :

$$\text{Total SS} = \text{Regression SS} + \text{Residual SS}.$$

Then unbiased estimator of σ^2 is

$$\hat{\sigma}^2 := \frac{\text{Residual SS}}{n-p-1}$$

Test : $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1: Some $\beta_j \neq 0$ at significance level α.

(continued)

Further assume $\varepsilon_i \sim N(0, \sigma^2)$, we design

$$F := \frac{\text{Regression SS}/p}{\text{Residual SS}/(n-p-1)} \underset{H_0}{\sim} F_{p,\,n-p-1}$$

Reject H_0 if $f > F_{\alpha;\,p,\,n-p-1}$.