Wednesday   Week 15.2   April 10th

Inference in simple linear regression

Example: Runoff revisited (based on Spring 2019 final)

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

Y_i: runoff (mm/yr) in year i

x_i: precipitation (mm/yr) in year i

ε_i: noise/error in year i

Assume: $\hat{\beta}_0 = -1364$

$\hat{\beta}_1 = 1.08$

$s^2 = 156^2.$

$\boxed{\text{Qu 1:}}$ Show that $\hat{\mu}_0$ is unbiased

$\boxed{\text{Ans:}}$ $E[\hat{\mu}_0] = E\left[\hat{\beta}_0 + \hat{\beta}_1 x_0\right]$

$= \beta_0 + \beta_1 x_1$ ✓

$\boxed{\text{Qu 2:}}$ Show $\mathrm{Var}(\hat{\mu}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$

(you may assume $\mathrm{Cov}(\bar{Y}, \hat{\beta}_1) = 0$)

$\boxed{\text{Ans:}}$

$$\mathrm{Var}(\hat{\mu}_0) = \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x_0)$$

$$= \mathrm{Var}(\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0)$$

$$= \mathrm{Var}(\bar{Y} + \hat{\beta}_1 (x_0 - \bar{x}))$$

$$= \mathrm{Var}(\bar{Y}) + \mathrm{Var}(\hat{\beta}_1)(x_0 - \bar{x})^2 + 2\,\mathrm{Cov}(\bar{Y}, \hat{\beta}_1 (x_0 - \bar{x}))$$

$$= \frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathrm{Var}(Y_i)}_{=\sigma^2} + \frac{\sigma^2 (x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + \underbrace{2(x_0 - \bar{x})\,\underbrace{\mathrm{Cov}(\bar{Y}, \hat{\beta}_1)}_{=0}}_{=0}$$

$\boxed{\begin{array}{l} \text{Recall:} \\ \mathrm{Cov}(aY, bX) \\ = ab\,\mathrm{Cov}(Y, X) \\ \mathrm{Var}(aY) \\ = \mathrm{Cov}(aY, aY) \\ = a^2 \mathrm{Cov}(Y, Y) \\ = a^2 \mathrm{Var}(Y) \end{array}}$

$$= \frac{1}{n}\sigma^2 + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sigma^2$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 + \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \checkmark$$

$\boxed{\text{Qu 3:}}$ Assume $\bar{x} = 3200$. Predict runoff for $x_0 = 3200$.

$\boxed{\text{Ans:}}$

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_0$$

$$= -1364 + 1.08 \cdot 3200$$

$$= \boxed{2092 \text{ mm/yr}}$$

$\boxed{\text{Qu } 4:}$ Give a 95% CI for μ_0 at $x_0 = 3200$.

$\boxed{\text{Ans:}}$ We know:

$$\frac{\hat{\mu}_0 - \mu_0}{\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}} \sim t_{n-2}.$$

Hence:

$$\hat{\mu}_0 - t_{.025,n-2}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)} < \mu_0 < \hat{\mu}_0 + t_{.025,n-2}\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}.$$

$$\sqrt{s^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)} = \sqrt{156^2\left(\frac{1}{25} + \frac{(3200-3200)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}$$

$$= \sqrt{156^2\left(\frac{1}{25} + \frac{0}{\cdots}\right)}$$

$$= \sqrt{\frac{156^2}{25}} = \frac{156}{5}.$$

Also, $t_{.025,n-2} = t_{.025,23} \approx 2.07$, and so

$t_{.025,23} \cdot \frac{156}{5} \approx 64.58$. Thus, with 95%

'confidence' we can say that

$$\boxed{\mu_0 \in (2027.42 \text{ mm/yr}, 2156.58 \text{ mm/yr})}.$$

Example: Hot chocolate revisited (based on Fall 2015 final)

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , \quad i = 1, \dots, 20$

Y_i: cups hot chocolate sold

x_i: ski conditions from $x_i = 1$ (bad) to 4 (excellent)

ε_i: noise

i: week (on each Sunday)

Qu 1: Would it be reasonable to construct a 95% CI for β_0, β_1?

Ans: It would be problematic. As discussed last Wednesday, the residuals are not homoscedastic (of equal variance) but are intead heteroscedastic (of different variance depending on x). The confidence intervals would therefore not be very accurate.

$\boxed{\text{Qu 2:}}$ Test if $\beta_1 > 8$ at the 95% level.

$\boxed{\text{Ans:}}$

We wish to test hypothesis:

$H_0: \quad \beta_1 \leq 8$

$H_1: \quad \beta_1 > 8$

at the 95% level ($\alpha = 0.05$).

$\hat{\beta}_1$ is approximately Gaussian with mean β_0 and variance $\dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \approx \dfrac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Hence,

Our test statistic will be:

$$Z = \frac{\hat{\beta}_1 - 8}{\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \;\dot{\sim}\; N(0, 1).$$

"approximately distributed as"

$$\approx \frac{9.51 - 8}{\sqrt{\frac{5.65^2}{24.95}}} \approx 1.33.$$

The critical value will be $z_{.05} \approx 1.64$, so there is insufficient evidence to reject the null hypothesis at the 95% level.

The p-value is $1 - \Phi(1.33) \approx 0.09$.