Lecture 4

Wednesday, May 15, 2019 2:27 PM

Inference for Simple Linear Regression

- CI's /HT's for β_0, β_1, μ.
- It turns out that the estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\mu}$ follow normal distributions and this result is a consequence of the fact that linear combinations of normal random variables are also normally distributed

  ↳ $y_i \sim N(\mu_i, \sigma_i^2)$ for $i=1,2,\dots,n$ then

$$\sum_{i=1}^{n} a_i y_i \sim N\left(E\left[\sum_{i=1}^{n} a_i y_i\right], Var\left[\sum_{i=1}^{n} a_i y_i\right]\right)$$

for real numbers a_i, $i=1,2,\dots,n$.

- Using this result it can be shown that

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

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)\right)$$

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\mu_0 = \beta_0+\beta_1 x_0, \ \sigma^2\left[\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]\right)$$

* because $E[\hat{\beta}_1] = \beta_1$ and $E[\hat{\beta}_0 = \beta_0]$ and $E[\hat{\mu}_0] = \mu_0$, we call these estimates unbiased

Hypothesis tests and confidence arise by noticing:

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0,1)$$

Because σ is unknown we can't actually calculate this quantity. So let's replace σ with its LSE $\hat{\sigma}$:

$$\frac{\hat{\beta}_1 - \beta_1}{\underbrace{\hat{\sigma}/\sqrt{S_{xx}}}_{SE(\hat{\beta}_1) \text{ "standard error"}}} \sim t_{(n-2)}$$

* Aside: If $Z \sim N(0,1)$, $U \sim \chi^2_{(v)}$ and they're independent, then

$$Z/\sqrt{U/v} \sim t_{(v)}$$

And: $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$

We can similarly show that:

$$\frac{\hat{\beta}_0 - \beta_0}{\underbrace{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}}_{SE(\hat{\beta}_0)}} \sim t_{(n-2)} \quad \text{and} \quad \frac{\hat{\mu}_0 - \mu_0}{\underbrace{\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}}_{SE(\hat{\mu}_0)}} \sim t_{(n-2)}$$

In general a test statistic for a parameter θ in a hypothesis of the form:

$$H_0: \theta = \theta_0 \quad \text{vs.} \quad H_A: \theta \neq \theta_0$$

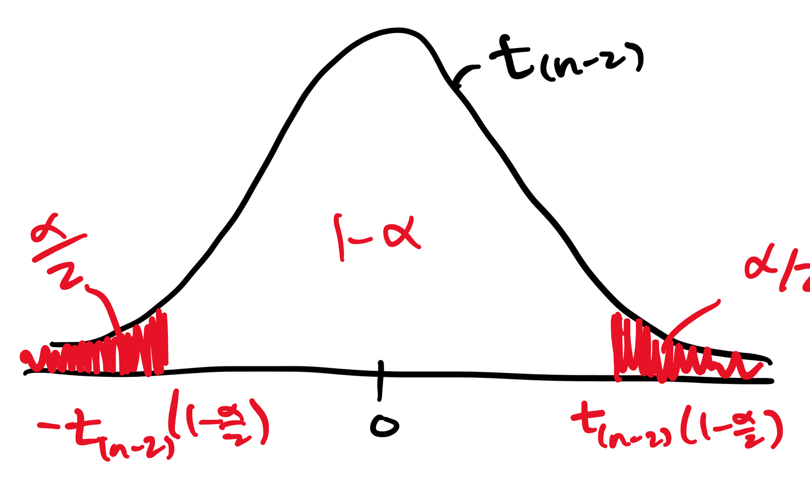
is $t = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$. P-values are calculated in the context of $t_{(n-2)}$.

The general form of a $100 \times (1-\alpha)\%$ confidence interval is:

$$\hat{\theta} \pm \text{critical value} \times SE(\hat{\theta})$$

where the critical value in confidence intervals for β_0, β_1 or μ_0 is $t_{(n-2)}(1-\alpha/2)$ where:

$$P\left(t_{(n-2)} \leq t_{(n-2)}\left(1-\frac{\alpha}{2}\right)\right) = 1 - \frac{\alpha}{2}$$



Common Inferences:

- 95% CI for β_0, β_1, μ_0
- $H_0: \beta_1 = 0$ vs. $H_A: \beta_1 \neq 0$

Example (Golf):

- $H_0: \beta_1 = 0$ vs. $H_A: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = -10.738 \to \text{p-value} = 2P(t_{(n-2)} \geq |t|)$$

$$= 2P(t_{(194)} \geq 10.738)$$

$$= 2.07 \times 10^{-21}$$

- 95% CI for β_1:

$$\hat{\beta}_1 \pm t_{(194)}(0.975) \times SE(\hat{\beta}_1)$$

$$= (-0.4637, -0.3198)$$

Prediction Intervals for y

- It is often of interest to predict a value of y for a given value of $x = x_0$. This predicted value is the same as the estimate of μ_0.

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

- We define prediction error as follows

$$y_0 - \hat{y}_0 = (\beta_0 + \beta_1 x_0 + \varepsilon_0) - (\hat{\beta}_0 + \hat{\beta}_1 x_0)$$

$y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$

$$= (\mu_0 - \hat{\mu}_0) + \varepsilon_0$$

This prediction error is the sum of two components that reflect two sources of uncertainty:

(1) Estimation uncertainty

(2) New random error ε_0

- $E[y_0 - \hat{y}_0] = \mu_0 - E[\hat{\mu}_0] + E[\varepsilon_0] = \mu_0 - \mu_0 + 0 = 0$

- $Var[y_0 - \hat{y}_0] = Var[\hat{\mu}_0] + Var[\varepsilon_0]$ (since $\hat{\mu}_0$ and ε_0 are independent)

$$= \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right) + \sigma^2$$

$$= \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right]$$

- The prediction error follows a normal distribution since $\hat{\mu}_0$ and ε_0 are normal random variables.

- A $(1-\alpha) \times 100\%$ PI for y_0 is:

$$\hat{y}_0 \pm t_{(n-2)}\left(1-\frac{\alpha}{2}\right) \underbrace{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}}_{SE(\hat{y}_0)}$$