Lesson 11. Inference for Simple Linear Regression Slope – Part 1

Note. In Part 2 of this lesson, you can run the R code that generates the outputs here in Part 1.

1 Overview

- Recall the simple linear regression model (population-level):

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \varepsilon \sim \text{iid } N(0, \sigma_\varepsilon^2)$$

- We want to **infer** something about the population based on our sample
- We've learned how to obtain and interpret **point estimates** of β_0, β_1 and σ_ε^2
- The parameter we're usually most interested in is
- Our main questions:

Do X and Y truly have a (linear) relationship at the population level?	
What can we infer about the nature of their relationship (size and direction) at the population level?	

2 Sampling distribution of $\hat{\beta}_1$

- We will see shortly that hypothesis testing and confidence interval computations for β_1 rely on the t-distribution
- Why?
- Under the conditions for simple linear regression:

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

- We can standardize:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim N(0,1)$$

- Since we don't know σ_ε^2, we estimate it with $\hat{\sigma}_\varepsilon^2 = \dfrac{SSE}{n-2}$:

$$\frac{\hat{\beta}_1 - \beta_1}{SE_{\hat{\beta}_1}} \sim t(n-2) \qquad \text{where} \qquad SE_{\hat{\beta}_1} = \sqrt{\frac{SSE/(n-2)}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

  - $\text{SE}_{\hat{\beta}_1}$ is the **standard error (SE)** of the estimated slope $\hat{\beta}_1$

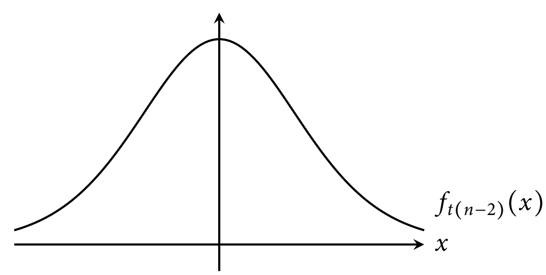
3 t-test for the slope of a simple linear regression model

- Question: **Does the predictor variable X have a significant association with the response variable Y?**
- Formal steps:
  1. State the hypotheses:

  2. Calculate the test statistic:

  3. Calculate the p-value:
     - If the conditions for simple linear regression hold, then the test statistic t follows

     $\Rightarrow$ p-value =



  4. State your conclusion, based on the given significance level α:

     If we reject H_0 (p-value $\leq \alpha$):

     > We reject H_0 because the p-value is less than the significance level $\underline{\alpha}$. We see significant evidence that $\underline{X}$ is associated with $\underline{Y}$.

     If we fail to reject H_0 (p-value $> \alpha$):

     > We fail to reject H_0 because the p-value is greater than the significance level $\underline{\alpha}$. We do not see significant evidence that $\underline{X}$ is associated with $\underline{Y}$.

     The underlined parts above should be rephrased to correspond to the context of the problem

Example 1. Let's look at the `AccordPrice` data again. Recall that we were interested in predicting `Price` from `Mileage`.

a. Fit a simple linear model predicting `Price` from `Mileage`.

Recall we did this in Lesson 7, using the following R code:

```
library(Stat2Data)
data(AccordPrice)

fit <- lm(Price ~ Mileage, data = AccordPrice)
```

b. Before we do any inference, it is important to make sure the **conditions** for a simple linear regression model are reasonably met.

   Recall that we already did this in Lesson 7.

c. Is the association between *Price* and *Mileage* significant? Use a significance level of $\alpha = 0.05$. Here is the output from `summary(fit)`:

```
Call:
lm(formula = Price ~ Mileage, data = AccordPrice)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5984 -1.8169 -0.4148  1.4502  6.5655 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  20.8096     0.9529   21.84  < 2e-16 ***
Mileage      -0.1198     0.0141   -8.50 3.06e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.085 on 28 degrees of freedom
Multiple R-squared:  0.7207,	Adjusted R-squared:  0.7107 
F-statistic: 72.25 on 1 and 28 DF,  p-value: 3.055e-09
```

- Other things to note:
  - What's happening in the `(Intercept)` line of the output?
  - If we want to do a **one-sided test** for β_1 (for example, $H_0 : \beta_1 \geq 0$ versus $H_a : \beta_1 < 0$ in the Accord example above), how could we use the R output to get the correct p-value?

4 Confidence interval for the slope of a simple linear regression model

- If the conditions for a simple linear regression model are met, then we can construct a $100(1-\alpha)\%$ **confidence interval for the slope** β_1 as follows:

Example 2. Use the output from Example 1 to do the following:

a. Construct a 95% confidence interval for β_1. Note that $t_{0.025,28} \approx 2.048$.
b. Interpret your confidence interval.

- You can compute the 95% CI for β_1 with this R code instead:

```
confint(fit, level=0.95)    # level is the confidence level
```

- The resulting output looks like this:

A matrix: 2 × 2 of type dbl

	2.5 %	97.5 %
(Intercept)	18.8577657	22.76146004
Mileage	-0.1486848	-0.09093915

- Other things to note:
  - Again, we could do something similar for β_0, but we usually don't
  - There is a direct connection between the hypothesis test and the confidence interval:

$$(1-\alpha)100\% \text{ CI for } \beta_1 \text{ does not contain } 0 \quad \Longleftrightarrow \quad t\text{-test for } \beta_1 \text{ will reject } H_0 \text{ at significance level } \alpha$$