Lesson 10. Inference for Regression Slope – Part 1

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

1 Overview

- Recall the simple linear regression model:

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

- We want to **infer** something about the population based on our sample
- We've seen 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:
  1. Do X and Y truly have a (linear) relationship at the population level?
  2. 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 the t-test and confidence interval computations 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 $\hat{\sigma}_\varepsilon^2 = \dfrac{SSE}{n-2}$:

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

  - The **standard error (SE)** is the standard deviation of a statistic

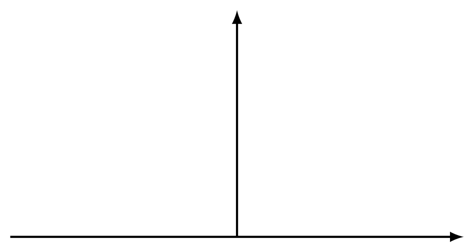
3 t-test for the slope of a simple linear regression model

1. State the hypotheses

2. Calculate the test statistic

3. Calculate the p-value

   - If the conditions for simple linear regression hold, then the sampling distribution of the test statistic under H_0 is



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

   - Provide your **decision**: e.g., reject H_0, fail to reject H_0
   - State your **conclusion** in terms of evidence: e.g.,
     - We see significant evidence that the true slope is not 0
     - We do not see significant evidence that the true slope is not 0

Example 1. Let's look at the `PorschePrice` 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 6, using the following R code:

```
library(Stat2Data)
data(PorschePrice)

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

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. Use the R output to test the hypothesis that $\beta_1 = 0$ at the $\alpha = 0.05$ significance level. Here is the output from `summary(fit)`:

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

Residuals:
     Min       1Q   Median       3Q      Max 
-19.3077  -4.0470  -0.3945   3.8374  12.6758 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 71.09045    2.36986    30.0  < 2e-16 ***
Mileage     -0.58940    0.05665   -10.4 3.98e-11 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 7.17 on 28 degrees of freedom
Multiple R-squared:  0.7945,  Adjusted R-squared:  0.7872 
F-statistic: 108.3 on 1 and 28 DF,  p-value: 3.982e-11
```

- 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 Porsche 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 form a CI for β_1 with the following formula:

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

a. Construct a 95% confidence interval for β_1
b. Interpret your 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)	66.2360186	75.9448869
Mileage	-0.7054401	-0.4733618

- 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
    - If we form a 95% CI for β_1 and it doesn't contain 0, then a hypothesis test of $\beta_1 = 0$ (versus $\beta_1 \neq 0$) will reject the null at the $\alpha = 0.05$ significance level