Bivariate Regression

Nicholas Ray

2023-09-15

Summary of Bivariate Regression Estimators

Intuition

Suppose that we are interested in the relationship between two variables (say Y and X). We can estimate this relationship in a straightforward way if we assume that the relationship is linear, where β_1 is our intercept and β_2 is our slope parameter, such that $Y_i = \beta_1 + \beta_2 X_i$. Unless we correctly include every relevant influence on our outcome of interest (Y), we will have a mismatch (ε) between our assumed relationship and reality.

Given our relationship of interest given above, we can estimate the parameters of our supposed relationship (i.e., β_1 and β_2) with the estimators discussed in lecture (equations (1) and (2) below, respectively).

$$\hat{\beta}_1 = \bar{Y} - \bar{X}\hat{\beta}_2 \tag{1}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{2}$$

Application

To give a concrete example, lets use a built-in dataset in `R`. This dataset is called `mtcars`, and has several variables (type `View(mtcars)` in your console to see them). Let's suppose we are interested in the relationship between miles per gallon (mpg) and horsepower (hp):

$$mpg = \beta_1 + \beta_2 hp + \varepsilon.$$

In `R`, we can estimate the parameters of this relationship with the `lm()` function. Below I show how to estimate this relationship with five random observations from the `mtcars` database. I select only 5 so that I can later more easily show the equivalence between the `R` estimates and the analytical ones using equations (1) and (2) above.

```
library(tidyverse) #loading "tidyverse," containing the "dplyr" package and "slice_sample" function

#most itemized way to do it, creating more "objects"
set.seed(1) #using a particular random number generator (1) to ensure reproducibility
data<-slice_sample(mtcars,n=5) #randomly selecting five observations from built-in R dataset
model<-lm(mpg~hp,data) #estimating a linear model between mpg (Y) and hp (X)
summary(model) #function that creates more output about lm object

#can do the same thing with less creation of objects, using "pipe" (%>%) operator from dplyr package
set.seed(1)
mtcars %>%
  slice_sample(n=5) %>%
  lm(mpg~hp,data=.) %>%
  summary()

#can also do all this in one nested line of code
set.seed(1)
summary(lm(mpg~hp,data=slice_sample(mtcars,n=5)))
```

To confirm that these estimates correspond to our analytic expectations:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{3}$$

$$= \frac{((175-150)(19.2-19.38)) + ((110-150)(21.4-19.38)) + ((245-150)(14.3-19.38)) + ((110-150)(21-19.38)) + ((110-150)(21-19.38))}{(175-150)^2 + (110-150)^2 + (245-150)^2 + (110-150)^2 + (110-150)^2} \tag{4}$$

$$= \frac{-697.5}{14,450} \tag{5}$$

$$= -0.048 \tag{6}$$

Thus, our analytic estimate in (5) equals that from our `lm()` function in `R`.

$$\hat{\beta}_1 = \bar{Y} - \bar{X}\hat{\beta}_2 \tag{7}$$

$$= 19.38 - (150 * -0.048) \tag{8}$$

$$= 26.58 \tag{9}$$

Again, our analytic estimate in (9) equals that from the `R` output (given rounding differences).

Presenting Results

Tables

```
library(stargazer) #package for presenting some objects using stargazer() function
stargazer(model,header = FALSE)
```

Table 1:

	Dependent variable:
	mpg
hp	-0.048^{***}
	(0.006)
Constant	26.620^{***}
	(0.937)
Observations	5
R^2	0.957
Adjusted R^2	0.943
Residual Std. Error	0.707 (df = 3)
F Statistic	67.348^{***} (df = 1; 3)
Note:	*p<0.1; **p<0.05; ***p<0.01

Plots

```
#basic plot
plot(data$mpg~data$hp);
abline(model)
```