Problem Set 2

Due by 3/7, 11:59PM

Consider a simple regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

for $i = 1, 2, \ldots, n$. Stacking the observations, we can write the model for the entire sample as

$$\boldsymbol{y} = \beta_0 \boldsymbol{l}_n + \beta_1 \boldsymbol{x} + \boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{l}_n$ is the n by 1 vector of 1's; and $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{\varepsilon}$ are all n by 1 vectors. Here $\{y_i, x_i\}$'s are random observations (i.i.d.), and β_0 and β_1 are unknown coefficients. The least-squares methodology tries to infer about the unknown coefficients by finding a best fit to sample data taken from the population. The best fit refers to values of β_0 and β_1 so that the sum of the squared residuals (or prediction errors) is the least possible, i.e.

$$(\widehat{\beta}_0, \widehat{\beta}_1)' = \underset{\beta_0, \beta_1}{\arg\min} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2.$$

It can be shown that the least squares estimator for the slope coefficient β_1 is given by

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

and the least-squares estimator for the intercept is given by

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x},$$

where $\bar{y}$ and $\bar{x}$ denote the sample means of y and x, respectively.

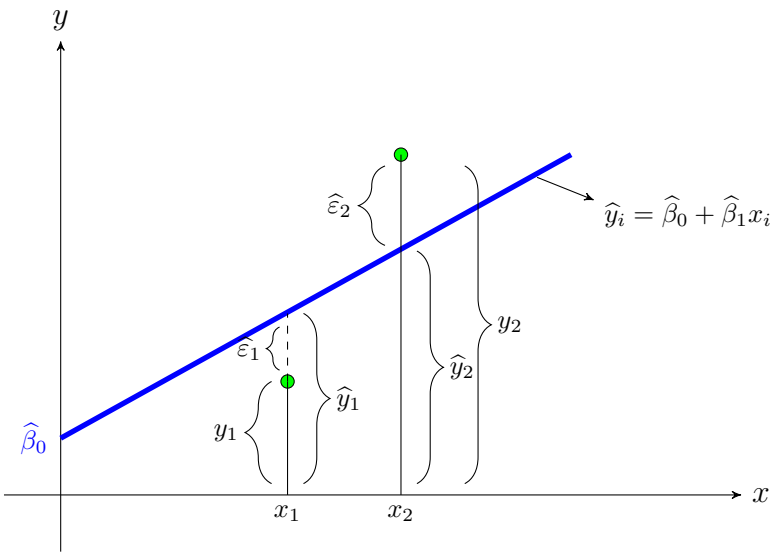

FIGURE 1. The least-squares regression line.

(0) Write a function which will take $\boldsymbol{y}$ and $\boldsymbol{x}$ as inputs and return estimates of β_0 and β_1 using the least-squares methodology.

(1) Set the seed to 37 for the random number generator in `numpy`, i.e., `np.random.seed(37)`.

(2) Generate an array of zeros of size 5000, and name it `tmp1`. For 5000 times, do the following in a loop:

   (a) generate 1000 observations on $\boldsymbol{x}$ by drawing randomly from the standard normal distribution, (i.e., $\boldsymbol{x}$ is a 1000 by 1 vector of standard Gaussian random variables);

   (b) generate 1000 observations on $\boldsymbol{\varepsilon}$ by drawing randomly from the standard normal distribution, (i.e., $\boldsymbol{\varepsilon}$ is a 1000 by 1 vector of standard Gaussian random variables);

   (c) generate the $\boldsymbol{y}$ vector using

$$\boldsymbol{y} = 0.5\,\boldsymbol{l}_n + 1.8\,\boldsymbol{x} + \boldsymbol{\varepsilon}$$

   using the draws from (a) and (b);

   (d) estimate the simple linear regression model using your function;

   (e) save β_1 estimate to `tmp1`.

(3) Calculate the mean of `tmp1`. Is it 1.8? Is it close to 1.8?

(4) Redo (2) and (3), but this time in 2(b), generate 1000 observations on $\boldsymbol{\varepsilon}$ as

$$\varepsilon_i = -0.5x_i + v_i$$

   for $i = 1, 2, \ldots, 1000$, where v_i is drawn from the standard normal distribution.

(5) Calculate the mean of `tmp1` again. Compared to your finding in (3), is it further away from 1.8? Why?