Problem Set 2

Due by 2/19

Consider a simple regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

for $i = 1, 2, \ldots, n$. 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 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 easily 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.

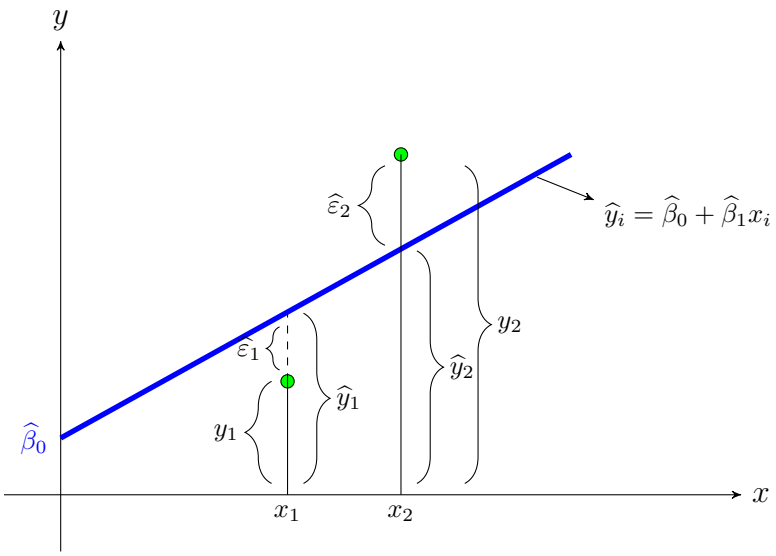

FIGURE 1. The least-squares regression line.

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

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

(3) Generate 1000 observations on x by drawing randomly from the standard normal distribution.

(4) Generate y as

$$y_i = 0.5 + 1.8 x_i + \varepsilon_i$$

for $i = 1, 2, \ldots, 1000$, where ε_i is a random draw from the standard normal distribution.

(5) Using your function estimate the simple linear regression model.

(6) Are the estimates you obtained for β_0 and β_1 the same as the true values you set, i.e. 0.5 and 1.8? Why?

(7) Regenerate y by regenerating ε. Fit the model again using your function. Are the estimates the same the estimates you obtained from the initial estimation? Why?