Week 2 Wednesday Worksheet

Valentino Aceves

April 10, 2024

Intro to Linear Regression

As economists, often times we are interested in the relationship between two economic variables X and Y , say X = education and Y = income. At first, we should investigate the summary statistics we reviewed last week, for example, mean, variance, and correlation. However, focusing on the summary statistics won't get us far. That's when we need an economic model. As a starting point, we will focus on a linear model, that is we model the relationship between X and Y as a line. Suppose we have a sample $\{X_i, Y_i\}$ for $i = 1, ..., n$ observations. We have the model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where Y_i = dependent variable, X_i = independent variable, β_0 = intercept, β_1 = slope, and ε_i = error/ residual. Here, we assume that this model is the **true** relationship between X and Y , and we are interested in the **parameters** β_0 and β_1, which are **unknown**.

How do we estimate the unkown β_0 and β_1?

One commonly used criteria is to estimate β_0 and β_1 is by minimizing the sum of squared residuals, which is the so called ordinary least squares, or simply OLS. For example, if we arbitrarily choose some constants b_0 and b_1 and fit the line, for each observation $\{X_i, Y_i\}$, we will have the error

$$\varepsilon_i = Y_i - (b_0 + b_1 X_i)$$

To account for both positive and negative residuals we square the term. Squaring this error and adding them together gives us a sum of squared residuals/errors (SSE) depending on b1 and b2:

$$SSE(b_0, b_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

Our job is to find b_0, b_1 such that the SSE is minimized. In math notation, we denote the estimator with a "hat" symbol. The following read as "b_0 and b_1 are the estimators that minimizes the SSE among all possible values b_0 and b_1". (arg min denotes the values/arguments that minimize the object.)

$$(\hat{b}_0, \hat{b}_1) = \arg\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

$$SSE(\hat{b}_0, \hat{b}_1) \leq SSE(b_0, b_1) \qquad \forall\ b_0, b_1$$

Explicit Formulas for the Estimators

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$$

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$ are the mean of X and Y respectively.

Question 1

Let $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Show that

$$E[\varepsilon_i | X_i] = 0 \implies E[Y_i | X_i] = \beta_0 + \beta_1 X_i$$

Question 2

$\sum_{i=1}^{3} x_i y_i =$

$\sum_{i=1}^{2} x_i - y_i =$

Question 3

Let X and ϵ be independent normally distributed random variables such that $X \sim N(5, 4)$ and $\epsilon \sim N(0, 9)$. Let Y be a random variable given by $Y = 1 + 2X + \epsilon$. Compute $Cov(X, Y)$