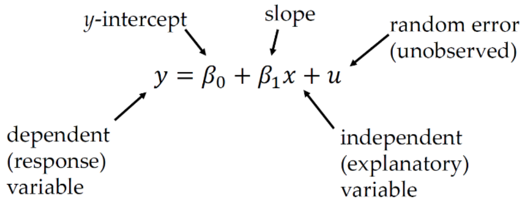
Simple Linear Regression Model

Definition and introduction

- Linear regression is a simple method of examining the relationship between y and x
    - Specifically, we explain variable y in terms of variable x.



Examples: X and Y

- Sales (y) explained by price (x)
- Sales (y) explained by advertising (x)
- Wages (y) explained by education (x)
- Household exp (y) explained by income (x)
- Prices (y) explained by costs (x)

What's the point?

- Suppose x and y are two variables representing some population. The objective is to:
    - Explain y in terms of x;
    - Study how y varies with changes in x;
- Issues:
    - How do we account for other factors that affect y?
    - What is the functional relationship between x and y?
    - How can we ensure that we are capturing a ceteris paribus (causal) result … if that is the goal?

The Error Term

- u represents factors other than x that affect y (unobserved)
    - randomness in behaviour
    - variables left out of the model
    - departures from linearity
    - errors in measurement

EXAMPLE: Soybean yield and fertilizer

$$Yield = B_0 + B_1 fertilizer + u$$

- B_1 – measures the effect of fertilizer on yield, holding all other factors fixed
- u – rainfall, land quality, presence of parasites

EXAMPLE: a simple wage equation

$$wage = B_0 + B_1 educ + u$$

- B_1 – measures the change in hourly wage given another year of education holding all other factors fixed
- u – labor force experience, tenure with current

Interpretation

- Interpretation in the simple linear regression model:
    - The goal is to understand how y varies with changes in x:

$$\frac{dy}{dx} = B_1 \ \ as\ long\ as\ \frac{du}{dx} = 0$$

- $\frac{dy}{dx} = B_1$ – by how much does the dependent variable change if the independent variable is increased by 1 unit
- $\frac{du}{dx} = 0$ – interpretation only correct if all other things remain equal when the independent variable is increased by one unit

Conditional mean independence

Question: when is it reasonable to assume that ceteris paribus holds? - Answer: requires conditional mean independence

- That is, since u and x are random variables, we can define the conditional distribution of u given any value of x.
    - In particular, for any x, we can obtain the expected (or average) value of u for that slice of the population described by the value of x.
- Crucial Assumption: $E(u|x) = E(u)$

    - Or that the average value of u does not depend on the value of x and
        - the average value of the unobservable is the same across all slices of the population determined by the value of x and that the common average is necessarily equal to the average of u over the entire population.
    - This gives that u is mean independent of x.
- Further: $E(u) = 0$
    - As long as the intercept b_0 is included in the equation, nothing is lost by assuming that the average value of u in the population is zero

Zero Conditional mean independence

- Two equations give: $E(u|x) = E(u) = 0$
    - Aka the zero conditional mean independence assumption

Causality?

- When is there a causal interpretation?

When the Conditional mean independence assumption holds:

$$E(u|x) = 0$$

- The explanatory variable must not contain information about the mean of the unobserved factors
- However, often unrealistic in simple models such as $wage = B_o + B_1 educ + u$
    - The conditional mean independence assumption is unlikely to hold here because individuals with more education will also be more intelligent on average.

Population Regression Function (PRF)

- The conditional mean independence assumption ALSO implies

$$\begin{aligned} E(y|x) &= E(B_0 + B_1 x + u|x) \\ &= B_0 + B_1 x + E(u|x) \\ &= B_0 + B_1 x \end{aligned}$$

- This means that the average value of the dependent variable y across the population can be expressed as a linear function of the explanatory variable x.
    - A one-unit increase in x changes the expected value of y by the amount of B_1
- Note: This tells us how Y changes with X on average, (i.e. Expected outcomes), not individual outcomes

Deriving OLS Estimates

- Data is required to estimate the regression model, with a random sample of n observations
- Plot the data, and fit as good as possible a regression line through the data points
- None of our points are on the line, therefore line represents the average with the residual representing the disparity of a point from the line

Ordinary Least Squares

- What does as good as possible mean? Regression residuals
- $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

Deriving OLS estimator: See Appendix 2A and https://are.berkeley.edu/courses/EEP118/current/derive_ols.pdf

- Start by defining fitted values for y: $\hat{y}_i = \hat{B}_0 + \hat{B}_1 x_i$
- The residual for observation i is thus: $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{B}_0 + \hat{B}_1 x_i$
- Choose parameters $\hat{B}_0$ and $\hat{B}_1$ to minimise: $\min \sum_{i=1}^{n} (\hat{u}_i)^2 = \sum_{i=1}^{n} \left(y - \hat{B}_0 - \hat{B}_1 x_i\right)^2$
- Take derivatives and set them equal to 0. This leads to the first order conditions
    - $\frac{dW}{d\hat{E}_0} = \sum_{i=1}^{n} -2\left(y - \hat{B}_0 - \hat{B}_1 x_i\right) = 0$
    - $\frac{dW}{d\hat{E}_1} = \sum_{i=1}^{n} -2x\left(y - \hat{B}_0 - \hat{B}_1 x_i\right) = 0$
- First solve for $\hat{B}_0$
    - Use first equation recalling $\sum_{i=1}^{n} y_i = N\bar{y}$
    - Leaves us with
        - $N\hat{B}_0 = N\bar{y} - N\hat{B}_1 \bar{x}$
        - $\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}$
- Solve for $\hat{B}_1$ using the second of the first order conditions substituting in $\hat{B}_0$
    - $\sum_{i=1}^{n} x_i(y_i - (\bar{y} - \hat{B}_1 \bar{x}) - \hat{B}_1 x_i) = 0$
    - $\sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{B}_1 \sum_{i=1}^{n} x_i(x_i - \bar{x})$
    - $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \hat{B}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2$
- Leaving us with: $\hat{B}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, Provided that $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$