Introduction to Classical Linear Regression Model (CLRM)

Robert Asiimwe

The simple linear regression model: Motivation

- Let us assume two variables Y and X represent some population
  - We want to know:
    - How does Y change when X changes?
    - What is the causal effect (ceteris paribus effect) of X on Y?

- *Examples of possible Y and X (in general or in your research)?*

Y	X
Beef demanded	Beef price
Food security status	Income (Household head employment)
Child nutrition	HH receives cash transfers (nutrition education)
HH resilience to floods	HH receives early warning information
Forest conservation	HH receives carbon credit transfers

Simple Linear Regression Model

The general form of the model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \mu_i \qquad (1)$$

$$\boxed{y = mx + c}$$

- β_0 (intercept) and β_1 (slope coefficient); are the population parameters to be estimated

- μ for "unobserved disturbance represents all factors other than X that affect Y

- The i subscript denotes the ith observation

Other terminologies on Y and X

- Y_i is dependent, explained, regressand or outcome variable

- X is independent, explanatory, regressors, predictors or covariates

Alternative form of equation in matrix notation is; $Y_i = X\beta + \mu_i$ (2)

- Where $X\beta$ is a short form of $\beta_0 + \beta_1 X_{1i}$ and X is a vector of regressors

Multiple Linear Regression Model

The general form of the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ \ldots + \beta_k X_{ki} + \mu_i \qquad (1a)$$

Or

$$Y_i = X\beta + \mu_i \qquad (2a)$$

- Each slope coefficient measures
  - The (partial) rate of change in the mean value of Y for a unit change in the value of a regressor, holding the values of all other regressors constant (ceteris paribus)

Population (true) model

- Eq 1 or its short form $eq.2$ is known as the populations or true model
  - The term population refers to a well-defined entity (people, firms, districts, countries, etc)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \mu_i$$

- Populations (true) model consists of two components
  1. deterministic component bx (the conditional mean of Y, or $E(Y|X)$)
  2. non-systematic, random or stochastic component μ_i

- An individual Y_i value is equal to
  1. the mean value of the population of which it is a member
  2. plus or minus a random error term

- For example
  - If Y represents family expenditure on food, X is family income
  - Eq 2 states that, the food expenditure of an individual family is equal to
    1. the mean expenditure of all the families with the same level of income
    2. plus or minus a random component
       - varies from individual to individual and depends on several factors

Regression analysis

- Primary objective is to explain the mean, average, behavior of Y in relation to the regressors (X)
  - How mean of Y responds to change in the values of the X variables
  - An individual Y value will hoover around its mean value
- The casual relationship between Y and Xs, if any, should be based on the relevant theory
  - having Y on the left and X on the right is based on theory
- How many regressors in the model depends on the nature of the problem
- The error term μ
  - A catchall for all variables that cannot be introduced in the model
  - The average influence of these variations is assumed to be negligible

Sample regression function: Notation

- The sample counterpart of equation 1 is

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ \ldots + \hat{\beta}_3 X_{3i} + e_i \qquad (3)$$

- Or, as written in short form

$$\hat{Y}_i = X\hat{\beta} + e_i \qquad (4)$$

  - where e is the residual (what cannot be measured)

- The deterministic component is written as
  - Also known as the empirical model

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ \ldots + \hat{\beta}_3 X_{3i} \qquad (5)$$

Assumptions of the OLS estimator

To use data to get unbiased estimates of β_0 and β_1, we have to make some assumptions about the relationship between X and μ

1. The regression model is linear in parameters; $Y = \beta_0 + \beta_1 X_1 + \mu$

- Though my not be linear in variables
  - Semi-logarithmic relationships

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 log X_2 + \cdots + \beta_n X_n + \mu$$

$$log Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \mu$$

  - Double-logarithmic relationship $log Y = \beta_0 + \beta_1 X_1 + \beta_2 log X_2 + \cdots + \beta_n X_n + \mu$
  - Polynomial relationships $Y = \beta_0 + \beta_1 X_1^{k_1} + \beta_2 X_2^{k_2} + \cdots + \beta_n X_n^{k_n} + \mu$
  - Inverse relationships $Y = \beta_0 + \beta_1 X_1 + \beta_2 (\frac{1}{X_2}) + \cdots + \beta_n X_n + \mu$
- Non-linear relationship $Y = (\frac{\beta_0}{\beta_1})\ X_1 + \mu$

Assumptions of the OLS estimator

2. Fixed X values or X values independent of the error term

3. Expected value of the error term is zero: $E(u) \ = \ 0$

4. The expected value of the error term is zero $E(\mu|X) = 0$

#3 & #4 → $\boldsymbol{E(\mu|X) = E(u) \ = \ 0}$ **(zero conditional mean)**

If this holds, x is “exogenous”; but if x is correlated with u, x is “endogenous”

5. No autocorrelation between the disturbance

$$cov\left(\mu_i, \mu_j \middle| x_i, x_j\right) = 0$$

Assumptions of the OLS estimator

6. Homoskedasticity or constant Variance of μ

Variance of the error term is the same regardless of X

$$Var(\mu_i) = E[\mu_i - E(\mu_i|X_i]^2$$

$$= \sigma^2$$

Assumptions about relationship between X variables

7. Full Rank no Multicollinearity

- No independent variable is a perfect linear combination

8. The number of observations n must be greater than the number of parameters (explanatory variables)

9. The nature of X variables

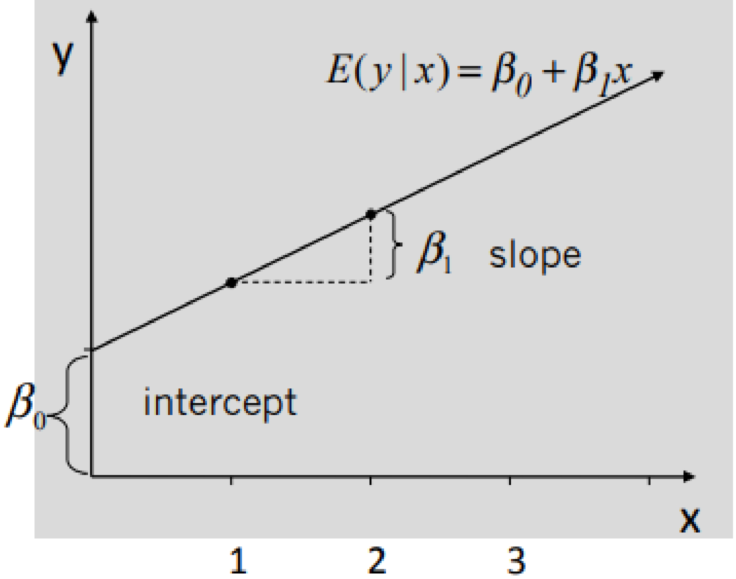
- Values of a given X in the sample must all not be the same
- There are no outliers in the values of X

Var (x)>0

- X values are exogenously generated

Interpretation of OLS Estimates –General

- Given that assumptions 1 to 7 hold, OLS gives best linear unbiased estimators BLUE
  1. Estimators are linear functions of the dependent variable Y
  2. Estimators are unbiased, in repeated samples the estimators approach their true value
  3. In the class of linear estimators, OLS estimators have the minimum variance i.e. they are efficient or the best estimators



Interpretation on the estimates

β_1: is the expected value of y given a one point increase in x; ceteris peribus (slope)

β_0: is the excepted value of y when x = 0 (intercept)

OLS Estimators of the estimates

- Ordinary) least squares (OLS) approach
- The estimated values of β_0 and β_1 are the values that minimize the sum of squared residuals
- “Fitted” values of y and residuals:

Fitted (estimated, predicted) values of y:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

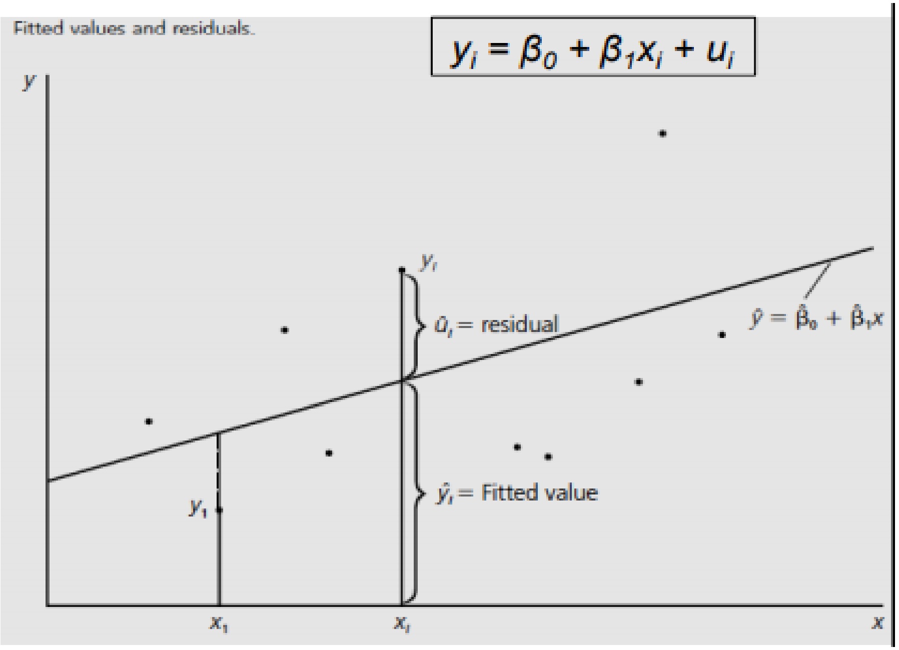
Residuals:

$$\begin{aligned}\hat{\mu}_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i\end{aligned}$$

OLS:

Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize:

$$\sum_{i=1}^{N} \hat{\mu}_i^2 = \sum_{i=1}^{N} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i)^2$$


Fitted values and residuals.
$y_i = \beta_0 + \beta_1 x_i + u_i$
y
y_i
$\hat{u}_i$ = residual
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
y_1
$\hat{y}_i$ = Fitted value
x_1
x_i
x

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

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

Model fit

- Total sum of squares: $SST \equiv \sum_{i=1}^{N}(y_i - \bar{y})^2$

- Explained sum of squares: $SSE \equiv \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$

- Residuals sum of squares: $SSR \equiv \sum_{i=1}^{N} \hat{\mu}^2$

$$SST = SSE + SSR$$

- Coefficient of determination or R^2: $= SSE/SST = 1 - (SSR/SST)$

Interpretation:

- The proportion of the sample variation in y that is explained by x (or a set of them)

- **Adjusted R^2**

Example: Modeling Willingness to Pay

- Assume we wish to estimate a regression model that explains people's WTP
  - Here WTP if the dependent variable i.e. the variable we want the model to explain
  - Factors that determine how much an individual is willing to pay could include age, income, family size, among others
- We collect data from three districts
  - This data is provided in folder Day 1 "practice1.dta"
- We will use this data to run the regression in Stata
  - Follow the do-file "day2" in the same folder to run the regression

Before the regression

- Think carefully about what you are trying to show, or the hypothesis you want to test.

- Use economic theory, previous work and your own common sense (intuition) to decide
  - What your independent variable should be
  - Which dependent variables you might need to include
  - The functional form of the regression
  - Try to specify a model before running a regression.

- Know the data as best as you can
  - Know what the variables mean.
  - Which are continuous and which are categorical
  - Create new variables if necessary
  - Know the means and standard deviations.
  - Know maximums and minimums. Are there any outliers? Should they be deleted?

After the regression ….(1)

- Look at number of observations
  - Is it what you expect? If not, you should find out why and determine if you should address the concern

Stata uses list wise deletion for missing data

- Look at the r2
  - If you have a "very low" r2;
    - Have you omitted some important variables;
    - Do other people doing similar work also have low r2?
    - Are you missing variables that many other people have included?
  - If very high (tending to 1);
    - If a high r2 is combined with many insignificant variables, you might have a multicollinearity problem
    - It might be an indication that you have mis-specified your model
  - The adjusted r2:

After the regression ….(2)

- Look at the F-test
  - you want a high F-value, and a low corresponding p- value
- Interpret the signs of the coefficients
- Interpret the size of the coefficients where relevant
- Look at the significance of the coefficients (most important?)
  - You want a low standard error, a high t-value, and a low p-value.
- Other tests follow,
  - testing for normality of error terms
  - checking for existence of heteroskedasticity
  - performing specification and robustness tests

Comments, questions, clarifications