Ordinary Least Squares

Edicson Luna

University of Pennsylvania

Introduction

Types of data

In Econometrics, you will find three types of databases

1. **Cross-sectional data:** it consists of a sample of individuals, households, or other types of units, taken at a given point in time.

2. **Time series data:** this type of dataset consists of observations on a variable or several variables over time.

3. **Panel data:** (also called longitudinal data) This set consists of a time series for each cross-sectional member in the data set.

Motivation

In Econometrics we will focus on the relationship between two or more variables.

Some relevant questions that we can "solve" using Econometrics are the following:

- Which quiz was more difficult between the first two?
- Which vaccine was more effective for battling COVID?
- Do women earn lower wages than men?
- Are large firms more productive than small firms?

And almost all the questions you can imagine. Nowadays, Econometrics tools cover a huge range of methodologies.

Plug-in Estimator

A plug-in estimator replaces an unknown population parameter with its sample analogue. Suppose we want to estimate a parameter:

$$\theta_0 = g(F)$$

where F is the population distribution.
The plug-in estimator uses the empirical distribution $\hat{F}_n$:

$$\hat{\theta} = g(\hat{F}_n)$$

In simple words: we estimate the object we do not know by "plugging in" what we observe in the sample.

As an example, for $\mu = E[X]$, its plug-in estimator is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Two "types" of differences

First type: Take a random sample $(X_i, Y_i)_{i=1}^n$ such that $E[X_i] = \mu_X$ and $E[Y_i] = \mu_Y$. Sometimes, we may be interested in the following parameter θ:

$$\theta := \mu_Y - \mu_X = E[Y_i] - E[X_i]$$

Second type: Now imagine X_i is a dummy variable (i.e. it only takes $0-1$ values). We may also want to estimate this other parameter θ:

$$\theta := \mu_1 - \mu_0 = E[Y_i|X_i = 1] - E[Y_i|X_i = 0]$$

First case

Notice that in the first case, we can define a new RV as $Z_i = Y_i - X_i$. Then, we can compute the sample mean statistic ($\bar{z} = \frac{\sum(y_i - x_i)}{n}$) and use hypothesis testing to conclude if the difference is negligible or not.

In the question "Which quiz was more difficult between the first two?" define X_i as the grade of the person i in quiz 1 and Y_i as the analogous for quiz 2. Z_i would be the difference in the grades and we can apply what we studied before.

In conclusion, there is nothing new here.

Second case

In the second case, we can divide the sample into two subsamples. One sample will contain the observations with $X_i = 1$; and the other sample, the observations with $X_i = 0$. We can think of these new samples as random samples of size n_1 and n_0.

As long as the observations are i.i.d., we know that $\bar{Y}_{n_1}$ (i.e. the mean conditioned on the first subsample) is a consistent estimator for $E[Y_i|X_i = 1]$. In the same line, $\bar{Y}_{n_0}$ works for $E[Y_i|X_i = 1]$.

Second case

We can work with the "plug-in" estimator which will be a consistent estimator for our parameter of interest.

$$\hat{\theta} := \bar{Y}_{n_1} - \bar{Y}_{n_0}$$

It is easy to show that

$$\frac{\hat{\theta} - \theta}{\mathrm{SE}(\hat{\theta})} = \frac{\bar{Y}_{n_1} - \bar{Y}_{n_0} - (\mu_1 - \mu_0)}{\sqrt{\frac{1}{n_1}\sigma_1^2 + \frac{1}{n_0}\sigma_0^2}} \xrightarrow{d} \mathcal{N}(0, 1)$$

And we might use hypothesis testing. The limitations of this approach are that it is difficult to implement if X_i is continuous or if we want to add other "control variables".

Linear regression

Linear regression

Linear regressions don't have those problems. On the contrary, working with continuous RV or adding control variables is straightforward.

A linear regression model can be thought of as how Y_i changes when X_i changes on average.

$$\mathbb{E}[Y_i|X_i = x] = \beta_0 + \beta_1 x$$

Example

Define Y_i as the salary of a person, and X_i as the gender of the person (1 for women and 0 for men). Then,

$$\mathbb{E}[Y_i|X_i = 0] = \beta_0$$
$$\mathbb{E}[Y_i|X_i = 1] = \beta_0 + \beta_1$$

Hence, the difference in salaries between women and men becomes:

$$\theta = \mathbb{E}[Y_i|X_i = 1] - \mathbb{E}[Y_i|X_i = 0] = \beta_1$$

That is, β_1 captures our parameter of interest. β_1 can be interpreted as the difference in wages between women and men.

Example: adding more covariates

Now, add a new RV Z that captures the person's education. For simplification, consider $Z = 1$ if the person had an undergraduate degree. In this case

$$\mathbb{E}[Y_i|X_i = x, Z = z] = \beta_0 + \beta_1 x + \beta_2 z$$

It is straightforward to see that

$$\theta = \mathbb{E}[Y_i|X_i = 1, Z_i = z] - \mathbb{E}[Y_i|X_i = 0, Z_i = z] = \beta_1$$

The result seems similar, but the interpretation changes a little bit: β_1 is interpreted as the difference in wages between women and men when they have the same level of education.

Another look to β_1

$\mathbb{E}[Y_i|X_i = x] = \beta_0 + \beta_1 x$ can also be written as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \text{with } E[\epsilon_i|X_i] = 0 \tag{1}$$

Analyzing the covariance between Y_i and X_i

$$\begin{aligned}
\text{Cov}(X_i, Y_i) &= \text{Cov}(X_i, \beta_0 + \beta_1 X_i + \epsilon_i) \\
&= \text{Cov}(X_i, \beta_0) + \text{Cov}(X_i, \beta_1 X_i) + \text{Cov}(X_i, \epsilon_i) \\
&= \beta_1 \text{Var}(X_i)
\end{aligned}$$

Then

$$\beta_1 = \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)}$$

Another look to β_1

Working with the plug-in estimator for both $\text{Cov}(X_i, Y_i)$ and $\text{Var}(X_i)$, the estimator $\hat{\beta}_1$ for β_1 would be

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

If we take expectation in both sides of (1), we get $\beta_0 = \mathbb{E}[Y_i] - \mathbb{E}[X_i]\beta_1$. Its respective plug-in estimator would be

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

Notice that the estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are straightforward to compute for a sample.

Ordinary least squares

Dictionary

- **Outcome variable:** the variable we are trying to explain, y_i. I will also say, explained variable or dependent variable.
- **Independent variable:** the variable(s) used to explain the outcome variable. It will be our x_i. I will also refer to it as explanatory variable(s), control variable(s), or regressor(s).
- **Error term:** it is the factors that explain y_i that are not included in x_i. It will be symbolized as ϵ_i.

Introduction to OLS

As mentioned before, we are interested in finding the parameter in a regression model of the form

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

This equation allows us to address the functional question of the relationship between y and x. If the other factors are held fixed ($\Delta\epsilon_i = 0$), then x_i has a linear effect on y_i

$$\Delta y_i = \beta_1 \Delta x_i \text{ if } \Delta\epsilon_i = 0$$

We will first focus on how to derive the estimators using OLS, and then we will study the assumptions for this to work.

OLS

This way (more popular) to get estimators is to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared errors. Let's start analyzing the algebraic way of finding this estimator, and later we will analyze its properties.

Define the residual (error of the model) by $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. Then, the ordinary least squares (OLS) estimator is obtained by solving

$$\min_{b_0, b_1} \sum_{i=1}^{n} e_i^2 = \min_{b_0, b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

We solve it by taking First-order-conditions (FOC) wrt b_0 and b_1.

OLS

After taking FOC wrt b_0 and b_1, we obtain the minimizers

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Which are the same as the plug in formula we derived before. Notice that using the Law of Large Numbers (LLN) and by the continuous mapping theorem (CMT), we can deduce that these estimators converge to their analogous population parameters.

OLS: adding more covariates

Now consider the case in which

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \text{ with } E[\epsilon_i | X_{1i}, X_{2i}] = 0$$

We could start taking the FOC of $\min_{b_0, b_1, b_2} \sum_{i=1}^{n} e_i^2$ wrt to b_0, b_1, and b_2. But this is very inefficient. To make it simpler, rewrite the expression in the following way

$$Y_i = X_i' \beta + \epsilon_i$$

Where $X_i = [1 \quad X_{i1} \quad X_{i2}]'$ and $\beta = [\beta_0 \quad \beta_1 \quad \beta_2]'$.

OLS: adding more covariates

The error would be defined by

$$e_i = y_i - x_i' b$$

Hence,

$$e_i^2 = (y_i - x_i' b)^2$$

Then, the problem would be of the form

$$\min_b \sum_{i=1}^{n} e_i^2 = \min_b \sum_{i=1}^{n} (y_i - x_i' b)^2 = \min_b \sum_{i=1}^{n} y_i^2 - 2 y_i x_i' b + (x_i' b)^2$$

If we take FOC with respect to b, we get the estimator

$$\hat{\beta} = \left(\sum_{1}^{n} x_i x_i' \right)^{-1} \sum_{1}^{n} x_i y_i$$

OLS: matricial form

We can also look at the problem using matrices.

$$Y_{n\times 1} = X_{n\times K}\beta_{K\times 1} + \epsilon_{n\times 1} \text{ where } \mathbb{E}[\epsilon|X] = 0$$

($\mathbb{E}[\epsilon|X] = 0$ means that the error, ϵ, conditioned on the complete set of covariates, X, has mean zero). That is,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1K-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2K-1} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nK-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{K-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

OLS: matricial estimation

In this case, capital letters mean that we are working with the matricial form. The error would be defined by

$$e = Y - X\beta$$

Then,

$$\begin{aligned} e'e &= (Y - XB)'(Y - XB) \\ &= Y'Y - B'X'Y - Y'XB + B'X'XB \\ &= Y'Y - 2B'X'Y + B'X'XB \end{aligned}$$

OLS: matricial estimation

You can verify that $e'e = \sum_{i=1}^{n} e_i^2$. Then, we can represent the problem as

$$\min_B e'e = \min_B Y'Y - 2B'X'Y + B'X'XB$$

Taking first order conditions wrt B

$$-2X'Y + 2X'XB = 0$$

then, the minimizer would be given by

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Goodness of fit

We want to know how well my regression line fits the data.

For simplicity, let's consider the case with only one regressor. That is, for the data $(x_i, y_i)_{i=1}^n$, consider the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ with the error having conditioned (on x_i) mean zero. Also, take $\hat{\beta}_0$ and $\hat{\beta}_1$ as the estimates for β_0 and β_1, respectively.

Goodness of fit

Define the following terms

- y_i: the observed value.
- $\bar{y}_i$: (sample) mean of y_i.
- $\hat{y}_i$: predicted y_i from the regression model. That is, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
- e_i: residuals for "in-sample" observations. id est, $e_i := y_i - \hat{y}_i$

R^2

Now define the following sums of squares

$$\text{Residual sum of squares } \mathrm{SS}_{\text{residual}} := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

$$\text{Explained sum of squares } \mathrm{SS}_{\text{explained}} := \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2$$

$$\text{Total sum of squares } \mathrm{SS}_{\text{total}} := \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

It can be shown that these three sums satisfy the equation

$$\mathrm{SS}_{\text{total}} = \mathrm{SS}_{\text{explained}} + \mathrm{SS}_{\text{residual}}$$

R^2

The "coefficient of determination" R^2 is defined by

$$R^2 := 1 - \frac{\mathrm{SS}_{\mathrm{residual}}}{\mathrm{SS}_{\mathrm{total}}} = \frac{\mathrm{SS}_{\mathrm{explained}}}{\mathrm{SS}_{\mathrm{total}}}$$

Interpretation:

- R^2 measures how close regression predictions are to the data points in your data set $(x_i, y_i)_{i=1}^n$.
- $R^2 = 0$ means that the explanatory variables have no effect. In other words, the regression line is not better than just the average.
- $R^2 = 1$ means that the explanatory variables perfectly predict y_i. This is not necessarily a good thing.

R^2

Notice that R^2 is defined for the dataset you are working with. If we add irrelevant (garbage) variables, the R^2 will increase even if they are independently generated random variables. This is just a mechanical (mathematical) result from the definition of R^2.

But this does not mean that the predictive power of your model is better. On the contrary, it can screw the model, and make it useless for working with new data.

Maximizing prediction is not the focus of this class. Rather it is that the predictions (even if they are just slightly increasing the R^2) are well-estimated.

$\bar{R}^2$

The adjusted R^2 is defined as:

$$\bar{R}^2 = 1 - \frac{SS_{\text{residual}}/(n-q-1)}{SS_{\text{total}}/(n-1)}$$

where:

- n: sample size.
- q: number of regressors, excluding the intercept.

The logic is simple: $\bar{R}^2$ compares the residual variance of the model with the total variance of y, but corrects for the number of regressors.

Unlike R^2, adjusted R^2 can decrease when we add variables that do not improve the model enough.

Properties of the OLS estimator

Some immediate implications of the OLS estimates are the following

1. The sum of the residuals is zero.

$$\sum_{i=1}^{n} e_i = 0$$

2. The sample covariance between the regressors and the OLS residuals is zero.

$$\sum_{i=1}^{n} x_i e_i = 0$$

3. $(\bar{x}, \bar{y})$ is always in the OLS regression line.

Expectation and variance of the OLS

Assumptions

We first need to discuss a list of assumptions to take the expectation of $\hat{\beta}_1$ (from the bivariate model). The first assumption is obvious and we have been using it from the beginning.

- **Assumption 1.** Linearity in parameters. That is, the true relationship between y_i and x_i is

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

Assumptions

- **Assumption 2.** There is random sampling. That is, the observations are i.i.d.
- **Assumption 3.** There is variation in the explanatory variable. This means that x_i is not the same value for all the observations.
- **Assumption 4.** Zero conditional mean. The error has an expected value of zero given any value of the explanatory variable.
$$\mathbb{E}[\epsilon_i|x] = 0 \quad \forall x \in \text{Supp}(X)$$

Unbiasedness

Theorem. Under assumptions 1 to 4, the OLS estimator is unbiased.

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

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_1 | x_i] &= \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \Big| x_i\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \Big| x_i\right] \\
&= \beta_1 + \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i - \bar{x})\epsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\
&= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \mathbb{E}[\epsilon_i | x_i] = \beta_1
\end{aligned}$$

Unbiasenedness

Finally, applying the LIE on $\mathbb{E}[\hat{\beta}_1|x_i] = \beta_1$ implies $\mathbb{E}[\hat{\beta}_1] = \beta_1$.

Therefore, $\hat{\beta}_1$ is an unbiased estimator of β_1. Also,

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_0|x] &= \mathbb{E}[\bar{y} - \hat{\beta}_1\bar{x}|x] \\
&= \mathbb{E}[\beta_0 + \beta_1\bar{x} + \epsilon_i - \hat{\beta}_1\bar{x}|x] \\
&= \beta_0 + \mathbb{E}[(\beta_1 - \hat{\beta}_1)\bar{x}|x] \\
&= \beta_0 + \mathbb{E}[(\beta_1 - \hat{\beta}_1)|x]\bar{x} = \beta_0
\end{aligned}$$

Again, applying LIE we get $\mathbb{E}[\hat{\beta}_0] = \beta_0$, and hence $\hat{\beta}_0$ is unbiased for β_0.

Variance

To find the variance of the estimator, we add a new assumption.

- **Assumption 5.** Homoskedasticity. That is, the error has the same variance given any value of the explanatory variable.

$$\text{Var}(\epsilon_i|x_i) = \sigma^2 \quad \forall i, x_i$$

Notice $\sigma^2 = \text{var}(\epsilon_i|x_i) = E[\epsilon_i^2|x_i] - (E[\epsilon_i|x_i])^2 = E[\epsilon_i^2|x_i]$. And applying the LIE (integrating over all the values of x_i), we get $\sigma^2 = \text{Var}(\epsilon_i)$. σ^2 is often called the error variance. Notice that this also implies $\text{Var}(y_i|x_i) = \sigma^2$.

Variance

Theorem. The variance of the estimator $\hat{\beta}_1$ is of the form

$$\mathrm{Var}(\hat{\beta}_1|x_i) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and for $\hat{\beta}_0$ is

$$\mathrm{Var}(\hat{\beta}_0|x_i) = \frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Notice that $\mathrm{Var}(\hat{\beta}_1|x_i) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ implica $\mathrm{Var}(\hat{\beta}_1) = \sigma^2 \mathbb{E}\left[\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$ (you should be able to prove this).

Error variance

The unbiased estimator of σ^2 will be of the form

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

The reason the unbiased estimator is divided by $n-2$ is that we have two restrictions on these errors ($\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} x_i e_i = 0$) and therefore, there are only $n-2$ degrees of freedom.

Multivariate regression

For the case when we have more than one x_i, the procedure is very similar. Assumptions 1 and 2 are the same, and assumptions 3 and 4 are slightly different.

- **Assumptio 3':** The matrix $\mathbb{E}\left[\frac{X'X}{n}\right]$ has complete rank. (i.e. it's invertible)
- **Assumption 4':** Zero conditional mean, $E[\epsilon|X] = 0$
- **Assumption 5':** Homoskedasticity. $Var(\epsilon|X) = \sigma^2 I_n$

Under those assumptions, $E[\hat{\beta}] = \beta$.
It can also be proven that $\text{Var}(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$. Notice that this implies $\text{Var}(\hat{\beta}) = \sigma^2\mathbb{E}[(X'X)]^{-1}$ (this is something you should be able to do).

The intercept

Should we always include the β_0 intercept in our regression model? The short answer is yes. Notice that we proved that $E[\hat{\beta}_0] = \beta_0$, which tells us that our estimator is still unbiased even if the real value is zero.

On the contrary, if we don't include it and $\beta_0 \neq 0$, then $\hat{\beta}_1$ is biased.

Analyzing the results

Interpretation of coefficients

In general, when working in absolute terms (or "in levels"), the interpretation is straightforward. Consider wage_i as the salary of the person i and educ_i as the years of education of i. The typical regression model would be given by

$$\text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \epsilon_i$$

In this case, we say "if a person has 1 more year of education, her salary is β_1 higher on average".

What is the interpretation of β_0?

Interpretation of coefficients

But what if the variables are not in levels? A typical case is that one of the variables is in logarithm or both are.

Consider l_i as the employment level of a company i, and k_i as the number of machines in i.

Let's analyze the following four cases:

1. $l_i = \beta_0 + \beta_1 k_i$
2. $\log(l_i) = \beta_0 + \beta_1 k_i$
3. $l_i = \beta_0 + \beta_1 \log(k_i)$
4. $\log(l_i) = \beta_0 + \beta_1 \log(k_i)$

Interpretation of coefficients

1: $l_i = \beta_0 + \beta_1 k_i$
The interpretation is analogous to our previous example: "if there is 1 additional machine, there is β additional employees on average".

2: $\log(l_i) = \beta_0 + \beta_1 k_i$
If we take the total derivative on both sides

$$\frac{\Delta l_i}{l_i} = \beta_1 \Delta k_i \rightarrow \frac{\Delta l_i}{l_i} \times 100\% = \beta_1 \Delta k_i \times 100\%$$

Notice $\frac{\Delta l_i}{l_i} \times 100\%$ is the percentage of change in employment. Then, setting $\Delta k_i = 1$, we say "if there is 1 additional machine, the employment changes by $100\beta_1\%$ on average"

Interpretation of coefficients

3: $l_i = \beta_0 + \beta_1 \log(k_i)$
Doing the same,

$$\Delta l_i = \beta_1 \frac{\Delta k_i}{k_i} \rightarrow \Delta l_i = \beta_1 \frac{\Delta k_i}{k_i} \times 100\%$$

The interpretation is that "if there are 1% more machines, there is $0.01\beta_1$ more employment".

4: $\log(l_i) = \beta_0 + \beta_1 \log(k_i)$
In this case, we can see that β_1 represents the elasticity between l_i and k_i. That is, "If there are 1% more machines, there is β% more employment on average".

Categorical variables

A **categorical variable** is a variable that can only take a fixed number of values. It is used to characterize unordered and qualitative data.

An example may be the country where a person was born.

- Colombia
- Italy
- United States
- $\vdots$

Notice that there is no predetermined order in this variable.

Categorical variables

Since the variable follows no order, just including it like it would be an error. Imagine that we have a linear regression between the income of a person y_i and the country of birth as the explanatory variable x_i. For simplicity, assume

$$x_i = \begin{cases} 1 \text{ if Colombia} \\ 2 \text{ if Italy} \\ 3 \text{ if United states} \end{cases}$$

In the linear model $y_i = \beta_0 + \beta_1 x_i$, β_1 would have no correct interpretation since there is no ordering in x_i. Then, how should we proceed?

Categorical variables

We should transform this variable into several dummy variables. That is, create a variable

$$x_{i1} = \begin{cases} 0 \text{ if Italy or U.S.A.} \\ 1 \text{ if Colombia} \end{cases}$$

$$x_{i2} = \begin{cases} 0 \text{ if Colombia or U.S.A.} \\ 1 \text{ if Italy} \end{cases}$$

$$x_{i3} = \begin{cases} 0 \text{ if Colombia or Italy} \\ 1 \text{ if U.S.A.} \end{cases}$$

Categorical variables

Then, choose a base country. Let's say, we choose Colombia. Finally, run the following linear regression

$$y_i = \beta_0 + \beta_1 x_{i2} + \beta_2 x_{i3}$$

Notice that we did not include one of the countries in the regression. The reason is that including x_{i1} will generate multicollineality. In this case, $x_{i1} + x_{i2} + x_{i3} = 1$, which is the value multiplying the intercept.

Also, when analyzing β_1 and β_2, the interpretation is made with respect to Colombia's average income. Notice that we could also have dropped another country or the intercept to avoid multicollinearity.

Interacting Variables

Sometimes the effect of one variable depends on the value of another variable.

A simple way to model this is by including an interaction term:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 (X_i \times Z_i) + u_i$$

The coefficient β_3 captures whether the effect of X_i changes when Z_i changes.

Main idea: an interaction allows the slope, or the difference between groups, to vary across another variable.

Interaction Between Two Dummies

Suppose we estimate:

$$\text{wage}_i = \beta_0 + \beta_1 \text{Female}_i + \beta_2 \text{College}_i + \beta_3 (\text{Female}_i \times \text{College}_i) + u_i$$

where Female_i and College_i are dummy variables.

- β_0: average wage of non-college men.
- β_1: difference between women and men among non-college workers.
- β_2: difference between college and non-college men.
- β_3: additional difference for college women.

So β_3 tells us whether the college wage premium is different for women than for men.

Interaction Between a Dummy and a Cont. Var.

Now suppose we estimate:

$$\text{wage}_i = \beta_0 + \beta_1 \text{Female}_i + \beta_2 \text{Educ}_i + \beta_3 (\text{Female}_i \times \text{Educ}_i) + u_i$$

where Educ_i is years of schooling.
For men:

$$\frac{\partial wage_i}{\partial Education_i} = \beta_2$$

For women:

$$\frac{\partial wage_i}{\partial Education_i} = \beta_2 + \beta_3$$

Therefore, β_3 tells us whether the return to one more year of education is different for women than for men.

Convexity or concavity

Now imagine two variables whose relationship is concave. This means $y_i(\lambda x_{i1} + (1-\lambda)x_{i2}) \geq \lambda y_i(x_{i1}) + (1-\lambda)y(x_{i2})$ where x_{i1} and x_{i2} are values in the domain of x_i. How would we model this?

For instance, consider y_i as the income and age$_i$ as the age. We should expect a concave relationship since the income of a persona increases during the adulthood and then starts falling in the later stages. To analyze this, we can propose the following model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

Analyzing the value of β_2 would confirm if the relationship is concave or not. Should it be positive or negative?

Conclusion

As you see, we may run a regression among two or more variables even if there is no linearity. Of course, we still need this non-linear relationship to be "linearizable" like in the examples we covered in the previous slides.

There are other methods for solving non-linear relationships (the most common is the method of moments), but we will not cover them.