EFR summary

Introduction to Econometrics, FEB12012X

2024-2025

Lectures 1 to 8

Weeks 1 to 3

Deloitte.

DeNederlandscheBank
EUROSYSTEEM

financials
LAUDAME

Details

Subject: Introduction to Econometrics IBEB 2024-2025

Teacher: T. Bago d'Uva, T. Van Ourti, P. Hans Franses

Date of publication: 21.03.2025

Introduction to Econometrics – IBEB – Lecture 1, week 1

Methods

Everyday vs Scientific learning

Everyday learning
Learning through tradition or from experts requires little effort, but these sources can be wrong. On the other hand, via one's own experience, one may learn the causal relationships between things, but there's a possibility of overgeneralization due to selective observations and drawing inaccurate observations.

Scientific learning answers the question "Is something true or not?" This involves

1. Extending existing knowledge, which can be tested by collecting data and performing analysis (scientific methods).

Association versus Causal Effect

When there is an association between two variables, it does not necessarily imply causation.

Association can provide useful fact descriptions, while causal effects indicate the relations between variables and, thus, can be used to understand the effectiveness of policy intervention.

Types of data and unit of analysis

Types of data

Experimental data: used to estimate the causal effects (e.g. treatment and control group)

Observational data: collected for general purposes and not designed to estimate causal effects

Time dimensional

Time series information on a set of indicators over time (e.g. GDP over several years)

Cross-section is when a sample is observed and data collected at a specific point of time

Panel data set combines the last two types mentioned before, when cross-sectional study is carried out over time

Unit of analysis

For different purposes the analysis will be built on different units:

- Individuals
- Firms
- Regions
- Countries

Operationalization and conceptualization

Conceptualization: means specifying what is meant by the specific terms used in research.

Operationalization: the process of developing specific procedures to empirically represent the concepts defined during conceptualization. In other words, it is about measuring theoretical concepts.

Quality of operationalization

Reliability: Measurement methods are reliable and if the concept was to be measured repeatedly the findings of the research would be the same.

Validity means that a measure accurately reflects the concept it is intended to measure.

Introduction to Econometrics – IBEB – Lecture 2, week 1

OLS: simple linear regression model

Relationships between variables

One of the ways to find out about the relationship between variables can be by constructing a **scatter plot**. The relationship can be negative or positive, or there can be no relationship.

Covariance and correlation

It is not always enough to just observe the relationship, thus when we want to quantify it we can make relevant computations.

Sample covariance: $s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$

Where n is sample size, X_i is the value of X for observation i (similarly Y_i) and $\overline{X}$ is the sample average of X (similarly $\overline{Y}$). The units of sample covariance = units X * units Y.

Sample correlation: $r_{XY} = \frac{s_{XY}}{s_X * s_Y}$

Where s_{XY} is the sample covariance and s_X is the sample standard deviation of X (similarly s_Y). A correlation of 0 reflects no correlation, a correlation of +1 reflects perfectly positive correlation and -1 reflects a perfectly negative correlation. The correlation coefficient is unitless. It shows the strength of the relationship between X&Y.

The linear regression model

Linear regression attempts to formulate a causal effect of one variable (x) over another (y) which is unlike a mere two-sided association of correlation.

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

The error term u_i represents all other factors influencing Y and measures a vertical distance between the population regression line and an observation. β_1 is the slope of the regression line and β_0 is the intercept.

The line of best fit

The best fit line that fittingly depicts the relationship between the X and Y variables has to minimize the sum of the squared differences between observed data points and the population regression line. The differences are squared in order to prevent the positive and negative residuals from negating each other.

To minimize $\sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)^2$

- OLS estimator $\widehat{\beta_0}$: $\widehat{\beta_0} = \overline{Y} - \beta_1 \overline{X}$
- OLS estimator $\widehat{\beta_1}$: $\widehat{\beta_1} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{s_{XY}}{s_X^2}$

Note: unit of the coefficient of X can be stated as unit of Y by unit of X.

- OLS predicted/fitted values: $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i$
- Residuals: $\widehat{u_i} = Y_i - \widehat{Y_i}$

Comparing correlation with linear regression model

The linear regression model is a very flexible framework that allows several directions for extensions, such as:

- Multiple X variables: having more than one independent variable simultaneously influencing Y. (multiple regression)
- Nonlinear relationships
- Discrete or binary variable

While correlation coefficient is unitless, the OLS estimator of β_1 is measured in $\frac{units\ Y}{units\ X}$. Linear regression model coefficient shows causality only under OLS assumptions, and if those do not hold it shows association and should not be used for policy design.

Goodness of fit measures

The R^2

Observed value equal: $Y_i = \widehat{Y_i} + \widehat{u_i}$, in which $\widehat{Y_i}$ is explained by the model fitted value and $\widehat{u_i}$ is unexplained residuals.

- **Total sum of squares (TSS**) is the total variation in the data: $TSS = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$
- **Explained sum of squares (ESS)**: $ESS = \sum_{i=1}^{n} \left(\widehat{Y_i} - \overline{Y}\right)^2$

It shows the variation in the data explained by the model

- Finally, the R^2 is the proportion of sample variance of Y_i that is explained by X_i

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}\left(\widehat{Y_i} - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2};\ R^2 = corr\left(Y_i, \widehat{Y_i}\right)^2$$

In the case of single explanatory variable $R^2 = corr\left(Y_i, X_i\right)^2$

Range of R^2: $0 \leq R^2 \leq 1$

R^2=1: model predicts Y_i perfectly

R^2=0: model (X_i) does not predict any variance in Y_i

Standard Error of the Regression (SER)

The SER shows the spread of the data points around the population regression line. Larger values indicate stronger deviation from predicted values.

$$SER = S_{\hat{u}} = \sqrt{S^2_{\hat{u}}}$$

Where $S^2_{\hat{u}}$ is the sample variance of the residuals $\hat{u}_i$

$$S^2_{\hat{u}} = \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2 = \frac{SSR}{n-2}$$

Introduction to Econometrics – IBEB – Lecture 3, week 1

OLS assumptions

1. Zero conditional mean
2. The observations are independently and identically distributed
3. Large outliers are unlikely

Assumption 1: Zero conditional mean

The **zero conditional mean assumption** implies that the expected values of the residual value given a value of X is zero.

$$E(u_i|X_i) = 0$$

The expected value of the residual is independent of X. This means that the correlation between the residual and X is zero. Thus the explanatory variable is uncorrelated with other factors that influence Y.

To know and assert whether this condition holds, we must be sure of the random assignment of the variable X. If X is not randomly assigned it is difficult to confirm the validity of the zero-conditional mean assumption.

If there is no random assignment to satisfy the OLS model, we need to suppose that X is uncorrelated with other factors, that is when X is 'as good as random'. In order to measure the pure causal effect of X on Y, the uncorrelated assumption is important. Otherwise, there would be an omitted variable bias and the pure effect would not be accounted for.

With **simultaneous causality** that is when variables influence each other, the zero-conditional mean will not hold.

Assumption 2: Independently and identically distributed observations

Independent and identical distribution holds in the case of simple random sampling from the same population. The distribution will be identical when the observations are obtained from the same population, and the observations are uncorrelated and thus independent.

When the sample is not representative as well as for time series and panel data this assumption does not hold.

Assumption 3: Large outliers in X and Y are unlikely

OLS is very susceptible to the influence of outliers, and thus "finite kurtosis" is an essential assumption. Mathematically, this is defined as:

$$0 < E(X_i^4) < \infty; 0 < E(Y_i^4) < \infty$$

If there are data errors it is best to eliminate large outliers by fixing or removing those data points. Fixing or dropping the data should only happen if it is suspected to be an error.

Sometimes, certain outliers can have dramatic effects on the population regression line hence it is desirable to be skeptical of extreme points.

Sampling distribution of OLS estimators

The estimators of the constant ($\widehat{\beta_0}$) and coefficient of X ($\widehat{\beta_1}$) of the linear regression models are computed from random samples and thus are random variables themselves with a probability distribution.

As different samples can lead to different estimates, the estimators are just some points in the sampling distribution of the estimator.
If one uses all possible samples of size n from a population and applies OLS to estimate the coefficients, one will realize that large samples of the β_1 estimator ($\widehat{\beta_1}$) approximate to a normal distribution.

This comes directly from the central limit theorem. As the coefficient of X is independent and identically distributed, the expected value is the true value of the coefficient of X. This implies that the OLS estimator is unbiased:

$$E(\widehat{\beta_1}) = \beta_1$$
$$E(\widehat{\beta_0}) = \beta_0$$

Note: With large samples OLS estimators follow approximately normal distribution. Therefore, normality assumption is not needed.

Property of OLS estimators

Unbiasedness

When the estimators are unbiased. Therefore, the mean sampling distribution $\widehat{\beta_1}$ equals β_1 and similar for $\widehat{\beta_0}$

Unbiasedness of $\widehat{\beta_1}$ is satisfied if assumptions 1 and 2 hold. When the number of observations (sample size) increases the estimator of the coefficient of X becomes more consistent and converges towards the true value.

Variance of estimators and consistency

Because of the central limit theorem in large samples, $\widehat{\beta_0}$ and $\widehat{\beta_1}$ approximately follow a normal distribution $\widehat{\beta_1} \sim N\left(\beta_0; \sigma^2_{\widehat{\beta_1}}\right)$, and jointly they follow a bivariate normal distribution.

Variance of $\widehat{\beta_1}$: $\sigma^2_{\widehat{\beta_1}} = \frac{1}{n} \times \frac{var\left[\left(X_i - \mu_i\right)u_i\right]}{\left[var\left(X_i\right)\right]^2}$

The variance of $\widehat{\beta_1}$ decreases when the number of observations increases, when the variance of residual factors decreases, and when the variance of the explanatory variable X increases.

=> OLS estimator unbiased and consistent
=> The sampling distribution used are hypotheses tests and confidence intervals

Interpretation

Conditional expectation of Y, given X for the population model $Y_i = \beta_0 + \beta_1 X_i + u_i$: $E\left(Y_i \mid X_i\right) = E\left(\beta_0 + \beta_1 X_i + u_i \mid X_i\right) = \beta_0 + \beta_1 X_i + E\left(u_i \mid X_i\right)$, which under Assumption 1 further simplifies to $E\left(Y_i \mid X_i\right) = \beta_0 + \beta_1 X_i$

Most common interpretation of β_1: when X goes up by 1, the expected value of Y given X goes up by $\beta_1 E(Y_i|X_i + \Delta X) = E(Y_i|X_i) + \Delta X \beta_1$

Interpretation of the intercept

Generally β_0 indicates an average Y when $X_i = 0$

The intercept may not always be interpretable and it will depend on the data whether the interpretation will be meaningful.

Example with binary regressors

Take on only two values (Male/Female, Yes/No, Agree/Disagree)
Dummy variable: D = 0,1

Population model: $Y_i = \beta_0 + \beta_1 D_i + u$

Conditional expectation: $E(Y_i|D_i) = \beta_0 + \beta_1 D_i$

Average Y when D = 0: $E(Y_i|D_i) = \beta_0$

Average Y when D = 1: $E(Y_i|D_i) = \beta_0 + \beta_1$

Interpretation: β_1 is the difference between the average when D=1 and the average when D=0. β_1 is the change in average Y when D=1 compared to D=0.

Introduction to Econometrics – Introduction to Econometrics – IBEB – Lecture 4, week 2

Hypothesis tests and confidence intervals in linear regression

Two-sided hypothesis test

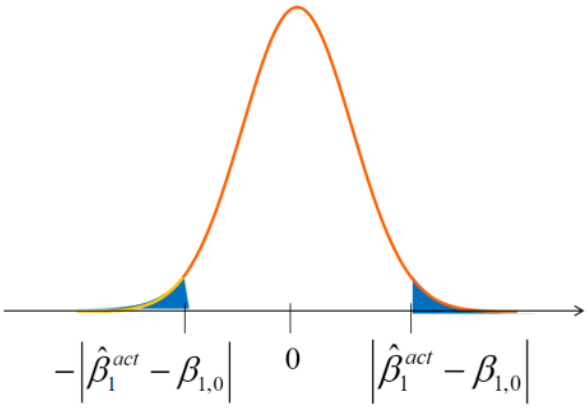
$$H_0: \beta_1 = \beta_{1,0} \quad vs \quad H_1: \beta_1 \neq \beta_{1,0}$$

Rejects null hypothesis if the estimated value $\hat{\beta}_1^{act}$ deviates substantially from the given hypothesized value $\beta_{1,0}$.

In other words, the null hypothesis is rejected if the probability of getting at least a value as extreme as the estimate $\hat{\beta}_1^{act}$ is very small (p-value)

t statistic and p-value

P-value: probability of obtaining $\hat{\beta}_1$ which is even further away from hypothesized value $\beta_{1,0}$ than he obtained $\hat{\beta}_1^{act}$ (shown by the blue area).



Source: Lecture 4

t-statistic: $t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$

Decision rule and Rejection region

Using the significance level of 5%:

Reject H_0 if

1. $P - value < 0.05$
2. $\left|t^{act}\right| > 1.96$ (critical value for a two-sided test)

Confidence intervals

E.g: 95% confidence means that from all samples that can be drawn, the interval contains the true value of β_1 in 95% of the cases.

Confidence interval: $\left[\widehat{\beta_1} - 1.96 \times SE\left(\widehat{\beta_1}\right); \widehat{\beta_1} + 1.96 \times SE\left(\widehat{\beta_1}\right)\right]$

Variance of error terms

When the variance of the error term is constant for all given values of X, then there is **homoskedasticity**. Otherwise, it is **heteroskedasticity**. When homoskedasticity holds, the formula of standard errors of $\widehat{\beta_1}$ can be simplified and the OLS estimator has minimal variance amongst all unbiased linear estimators (efficient).

However, homoskedasticity does not always hold. Therefore, it is important to use heteroskedasticity-robust standard errors.

Significance

Statistical significance is decisive in whether to reject or not to reject the null hypothesis. Economic significance involves not only statistical significance, but also the economic effect implied by the data analysis and testing's result. Some statistical results may be significant but not economically meaningful. In this discussion of hypothesis testing for the linear regression coefficient, the key warning is that the size (the magnitude of the effect, i.e., $\widehat{\beta_1}$) matters.

Introduction to Econometrics – IBEB – Lecture 5, week 2

OLS: OVB, multiple linear regression, assumptions

To measure the causal effect of variable X on Y one would want the OLS estimator to be unbiased. If, however, another variable Z is correlated with X and it has a causal effect on Y too, then the coefficient estimated for X would not purely reflect the

causal effect of X on Y and would be a combination of effects, thus leading to a biased estimator.

If the correlation between the error term and variable X is not equal to zero (the zero-conditional mean assumption is violated), then the variable Z as mentioned previously would affect Y. In this case the error term is function Z and other factors: $u_i = \beta_2 \times Z + v_i$, where v_i is the remaining of the error term, with everything else influencing Y except for X and Z.

When zero-conditional mean assumption holds: $corr(u_i, X_i) = 0 \Leftrightarrow \beta_2 \times corr(Z_i, X_i) = 0$ (assuming $corr(v_i, X_i) = 0$).

When the estimator of β_1 is biased we have that the expected value of $\widehat{\beta_1}$ equals β_1 plus the bias. This is summarized in the formula: $E(\widehat{\beta_1}) = \beta_1 + \beta_2 \frac{corr(X_1, X_2)}{s_{X1}} \times s_{X2}$

Direction of bias

The bias, when a simple model includes X1 and omits X2, can be positive or negative depending on the sign of β_2 and correlation between X1 and X2.

	$corr(X_1,X_2)>0$	$corr(X_1,X_2)<0$
$\beta_2>0$	**Positive** bias	**Negative** bias
$\beta_2<0$	**Negative** bias	**Positive** bias

Source: Lecture 5

Multiple regression model

By including the omitted variable into the model we try to satisfy the zero-conditional mean assumption. The omitted variable will no longer cause a correlation of error term with X1. The regression model thus can be finally expanded as follows:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + v_i$$

where the **main variable** of interest is X1, and X2 can be considered as the **control variable**. Since there is more than one coefficient that explains Y, it is a **multiple regression model**.

Interpretation of multiple regression model

Population multiple regression model is similar to the model with a single regressor and represents the average relationship between the independent variables and Y. The interpretation is the change in Y due to the change in X1 when X2 held constant. For example, if X2 is held constant, and X1 goes up by 1 then Y on average goes up by β_1. The OLS estimators, predicted values and residuals are obtained similarly to a model with a single regressor.

Assumptions for the multiple regression model

Similar assumptions to the one discussed for linear regression.

1. **Zero-conditional mean**: $E\left(u_i|X_{1i}, X_{2i},..., X_{ki}\right) = 0$

If we are interested in causal effect of all X1, X2...Xk, we use a weaker assumption

1. **Conditional mean independence**: $E\left(u_i|X_{1i}, X_{2i},..., X_{ki}\right) = E\left(u_i|X_{2i},..., X_{ki}\right)$

If interested in the causal effect of X1, we can make use of a weaker assumption that implies that the error term is independent of X1 and not all other variables. When conditional mean independence holds, one can interpret the effect of X1 on Y as a *causal effect*, and one can only interpret the effect of X2 on Y as a *partial association*. If the variable of interest is only X this is not a problem, we call X2 a control variable and X1 the variable of interest.

2. **Observations being independent and identically distributed**

3. **Large outliers of the variables are unlikely**

4. **No perfect multicollinearity**

Perfect collinearity between X1, X2 and X3 if there is a perfect linear relationship between 3 variables, such that $X_1 = a + bX_2 + cX_3$ with $b \neq 0 \; and \; c \neq 0$. Perfect

collinearity between explanatory variables happens in such cases as having the same variable in different units, or a dummy variable trap.

In situations when there is linear conversion of variables like change of units, it makes no reasonable sense to include both the variables (as you would essentially include the same variable twice in different measurement units).

In the case of dummy variables, it is always advisable to drop out a dummy for one category. When interpreting models with one dummy dropped out, the coefficients are always interpreted relative to the dropped-out dummy (the base/reference category).

Sampling distribution

We need the sampling distribution for both confidence intervals and hypothesis tests. Under the 3 assumptions of OLS and no perfect multicollinearity, the estimator coefficients of the independent variables individually follow a normal distribution and collectively follow a multivariate normal distribution.

Variance of $\hat{\beta}_j$ decreases with sample size n; decreases with variance of Xj; increases with variance of error term u_i; increases with correlation between X's (imperfect multicollinearity), however if assumption 1 holds then the model is still unbiased.

Measures of fit

Standard Error of the Regression (SER)

The SER, similarly to the simple regression model, shows the spread of the data points around the population regression line. Larger values indicate stronger deviation from predicted values.

$$SER = S_{\hat{u}} = \sqrt{S^2_{\hat{u}}}$$

Where $S^2_{\hat{u}}$ is the sample variance of the residuals $\hat{u}_i$

$$S^2_{\hat{u}} = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-k-1}$$

However the SSR (sum squared residuals) is now divided by n-k-1 (where k stands for the number of independent variables that influence Y) to derive variance.

The R^2

Similarly to the simple regression model

$$TSS = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

$$ESS = \sum_{i=1}^{n} \left(\widehat{Y_i} - \overline{Y}\right)^2$$

It shows the variation in the data explained by the model

Finally, the R^2 is the proportion of sample variance of Y_i that is explained by X_i

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

A special characteristic of R^2 is that it always increases when a regressor is added to the model. In order to deflate this sensitivity the adjusted R-squared formula is as follows: $\overline{R}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SSR}{TSS}$

Finally, it is noteworthy that the measure of fit cannot be compared and used if the dependent variables differ in the ways they are defined. Additionally, the measure of fit only represents the explained variation, but does not account for biases and whether the assumptions even hold.

Introduction to Econometrics – IBEB – Lecture 6, week 2

OLS: hypothesis tests, confidence intervals and model specification

When analyzing regression models, one of the worst problems encountered is when one of the three assumptions is violated, as that makes the estimator biased. If one variable is stipulated as being correlated with the variable X1, then this variable

should be included in the model as a control variable since otherwise, the model could be subject to omitted variable bias.

After introduction of this variable X2 (control variable), the zero-conditional mean must hold such that the conditional mean of other factors given variable X1 and variable X2 is 0.

Hypothesis test for a single coefficient

Hypothesis: $H_0: \beta_j = \beta_{j,0} \; vs \; H_1: \beta_j \neq \beta_{j,0}$

T-statistic: $t = \frac{\hat{\beta}_j - \beta_{j,0}}{SE(\hat{\beta}_j)}$

P-value: $p - value = 2\phi(-|t^{act}|)$

Reject H0 at 5% sig.level: $|t^{act}| > 1.96$

Test of joint hypotheses

Test of joint hypotheses can be used to specify the null hypothesis that the coefficients of various variables equal to a hypothesized value, which are q restriction and the alternative hypothesis that one or more of these q restrictions does not hold. In general form the hypotheses are formulated as follows:

$H_0: \beta_j = \beta_{j,0}, \beta_m = \beta_{m,0} \ldots (total\ of\ q\ restrictions)$

$H_1: one\ or\ more\ of\ the\ q\ restrictions\ does\ not\ hold$

One must use joint hypotheses testing instead of individual one because under the assumption, the coefficients have an approximate **bivariate normal distribution** in sufficiently large samples.

In case if one only knows the individual t test and not the F test then the **Bonferroni method** can be incorporated which uses special critical values to account appropriately for the significance level.

F-statistic

$H_0: \beta_1 = 0 \; and \; \beta_2 = 0 \; vs \; H_1: \beta_1 \neq 0 \; and/or \; \beta_2 \neq 0$

The formula for F statistics can look different from previous courses as we do not assume homoskedasticity here.

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t1,t2} t_1 t_2}{1 - \hat{\rho}^2_{t1,t2}}\right)$$

where t_1 and t_2 are the t-statistic of separate tests and $\hat{\rho}_{t1,t2}$ is the estimator of the correlation of the two t-statistics (it will happen to be 0 when there is no correlation between X1 and X2).

The distribution of the F statistics in large samples follows F distribution with degrees of freedom q (number of restrictions) in the numerator and ∞ in the denominator: $F - statistic \sim F_{q,\infty}$

Reject H0: if $F > critical\ value\ of\ F_{q,\infty}$

Common critical values for $F_{2,\infty}$: 10% sig. level =2.30, 5% =3.00, 1%=4.61

When testing whether the coefficients have no effect on Y, that is when all coefficients except the constant are zero, the hypotheses can be stated as follows:

$$H_0: \beta_1 = 0, \beta_2 = 0 \ldots \beta_k = 0 \ vs \ H_1: \beta_j \neq 0, \ at\ least\ one\ j$$

When such a null hypothesis is rejected at a given significance level, it means that coefficients are jointly significant or jointly significantly different from zero.

Omitted variable bias

Despite incorporating another variable X2 to prevent any bias, it is still very plausible that variables X1 and X2 (explanatory and control variables) do not satisfy the zero conditional mean assumption. In this case, one can adopt the weaker assumption of conditional independence.

This implies the correlation between variable X1 and other factors is 0. This will result in the effect of variable X1 to be purely causal, but the effect of variable X2 will display partial association, thus a mixture of effects of variable X2 and other factors.

If, however, even the conditional independence does not hold then the model has an omitted variable bias and more control variables can be introduced to the model. It is called a **robustness check** when one introduces changes into the model (like including new control variables) to see if the results would differ.

Introduction to Econometrics – IBEB – Lecture 7, week 3

Nonlinear regression functions

If the effect measured by the slope of the regression function depends on the value of the independent variable(s), we should have a nonlinear relationship.

It is always advisable to check whether a non-linear model improves the linear model by testing whether an additional regressors are significantly different from 0, furthermore, the graph can be used to observe the evenness of the spread of points and whether there is an improvement in the fit too.

There are a number of forms of non-linear models we can employ. Here we will cover:

Form 1: Polynomials
Form 2: Natural Logarithmic Transformation of the dependent and/or independent variable(s)
Form 3: Interaction Effects

Polynomial regression models

Polynomials use a linear function of a variable, where the linear function contains the variable taken to the power.

For quadratic polynomials, when the coefficient in front of squared variable is positive it represents the increasing returns to scale and when that coefficient is negative we can see decreasing returns to scale.

Testing

If the population regression function is considered linear, then the quadratic and higher-degree coefficients would not be useful in the regression functions. To test this, we can perform an F-test where the null hypothesis is the regression being linear and the additional regressors are equal to 0; the alternative hypothesis is that at least one of the additional regressors is not equal to 0.

Natural logarithmic transformation of the variable

This method employs the same regression model but with a logarithmic transformation of variable Y.

2 reasons:

- Outliers in the right tail can be dealt with using this method. Large outliers lead to a violation of the third OLS assumption, and they are less likely to affect the model after this transformation when the large outliers are compressed.
- Used if one is interested in percentage changes.

Log-linear model

Logarithmic transformation of dependent variable (Y) only

- Interpretation of β_1: a 1-unit change in X corresponds to $\left(\beta_1 \times 100\,\%\right)$ change in Y (semi-elasticity).

Linear-log model

Logarithmic transformation of variable X only.

- Interpretation: $1\,\%$ change in X corresponds to $0.01 \times \beta_1$ units of Y .

Log-log model

Both independent (X) and dependent (Y) variables are transformed logarithmically.
- Interpretation: a $1\,\%$ change in X corresponds to a $\beta_1\,\%$ change in Y. In this case, β_1 is called elasticity.

Interaction effect

The example of a model that includes interaction effect:
$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u_i$

The inclusion of $\beta_3 X_1 X_2$ term accounts for the interaction effect. It is useful to add when we believe that the effect of a variable depends on another variable.

When OLS fails

OLS fails when there's nonlinearities in the parameters. The previous models are nonlinear in X but are linear functions of the coefficients (parameters)

Introduction to Econometrics – IBEB – Lecture 8, week 3

Internal and external validity

Association is not causation and when there are any policy recommendations only causal effects should hold an important value.

A study is **internally valid** if the statistical inferences about the causal relationship are valid for the population and setting studied.

There are two sets of population and setting: one that is studied and one to which inferences can be generalized upon. The **population studied** is the one from which the sample was derived. The **population of interest** is one to which the inferences are generalized on. The **setting** is the institutional, legal, social and economic background of the study.

Threats to internal validity if these do not hold:

1. The estimator of the causal effect should be unbiased and consistent.

2. The hypothesis test should have the required significance level

A study is **externally valid** if the inferences can be used to make generic inferences to other populations and settings too.

Threat 1: Omitted variable bias (OVB)

If there is a variable that is omitted and is correlated with the variable of interest, as well as being a determinant of the dependent variable, then there is an omitted variable bias.

If the correlation between the variable of interest and omitted variable has the same sign as the effect of omitted variable on dependent variable, then there is an **upward bias**. If however, the signs are opposite then there is a **downward bias.**

Good and bad control variables

Control variables' values should always be generated before the variable of interest's are. Stated simply, if the variable of interest has a causal effect on the new (control) variable then the new variable is not a good control. This is not applicable the other way around (that is, if the control variable affects the variable of interest).

Threat 2: Errors-in-variables

Independent variable

- Random measurement error (classical measurement error): Bias towards 0
- Non-random: downward or upward sloping bias

Dependent variable

- Random: No bias but reduced precision
- Non-random: downward or upward sloping bias

Solutions: instrumental variables regression, and developing a mathematical model of the measurement error and using the resulting formula for correction.

Threat 3: Sample selection

Missing data at random leads to no bias. Missing data for the regressor also leads to no bias, however, the interpretation of the coefficient would then only hold for a subset of the population for which the observations are not missing.
The exception is when there is missing data on the dependent variable, then there is a bias. The solution to this issue is the use of appropriate sampling.

Threat 4: Simultaneous causality

There is no problem in the case of having causality that runs from the regressor to the dependent variable. However, if the reverse also holds true, then there is a bias as OLS will include both directions of causality. The potential solutions are instrumental variables regression and the design of research (randomized control trial).

Threat 5: Functional form misspecification

If there is a non-linear relationship but we adopt a linear model in some sense, there is an omitted variable bias. Therefore, it is best to test whether a significantly different from zero non-linear coefficient exists.

Threat 6: Inconsistency in the standard error

To avoid inconsistency in the standard error, always adopt a heteroskedasticity robust standard error and ensure independent and identically distributed observations.

Forecasting vs causal relationship

The goal of developing a causal model is deriving the best description of behavior. The first concern is internal validity, and the second concern is external validity of the model. In contrast, for forecasting models, the models' external validity is of greater importance than internal validity. To have the best forecast for the future, the requirements are good explanatory power, stability of results, and precision.

References

Van Ourti, T. (2025). Lecture 1: *Methods 1* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119869?wrap=1

Bago d'Uva, T. (2025). Lecture 2: *Methods 2* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119870?wrap=1

Bago d'Uva, T. (2025). Lecture 3: *Methods 3* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119871?wrap=1

Bago d'Uva, T. (2025). Lecture 4: *Methods 4* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119904?wrap=1

Bago d'Uva, T. (2025). Lecture 5: *Methods 5* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119594?wrap=1

Bago d'Uva, T. (2025). Lecture 6: *Methods 6* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119890?wrap=1

Van Ourti, T. (2025). Lecture 7: *Methods 7* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119930?wrap=1

Van Ourti, T. (2025). Lecture 8: *Methods 8* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/47709/files/100119913?wrap=1