ECON 326 - INTRODUCTION TO ECONOMETRICS II

计量经济学入门2

Author
Wenyou (Tobias) Tian
田文友
University of British Columbia
英属哥伦比亚大学
2024

Contents

1 Fundamentals

Econometrics is a discipline of economics that uses natural economic data to draw conclusions that describe the economic state.

A general economic model is a formal, but simplified description of the real world. It is an abstraction and focuses on the forces and questions at interest. The theoretical framework does not rely on real-world data, while empirical models, or econometric models do.

An ***econometric model*** first maintains hypothesis about the world (as a model does). There are mainly two types of econometric models:

1. Reduced Form: This type answers a narrow question of interest, and only merely describes a data relationship (e.g. minimum wage vs. unemployment).
2. Structural: This type uses more information to measure specific points from theoretical models. That is identifying the specific components to answer complex questions. It is more generalized, and it is often used as a policy tool.

An econometric model involves the use of data, which are a snapshot of economic process. Data reflects the behaviour of the participants in the economy. Some usual kinds include survey data, cross-sectional data, and time series data.

The econometric model is built due to the interest of research questions, which is transforming an observation into a hypothesis. Some common types of research questions include classification of observations, predictions, explorations, correlations, clustering, fitting, identification etc.

An econometric model can serve mainly two different purposes, predictive or inferential. Consider a general model: $y = f(x|\beta)$, where y is the outcome variable, x is the explanatory variable, β is the parameter that describes f, and f is the relationship between x and y, we can first come up with the optimal β^* where we calculate fitted values $y^* = f(x|\beta^*)$ and calculate the error to be $y - y^*$,

1. Predictive: A predictive model cares about y^*, and how accurately the model can predict. We generally do not care about how the predictions are come up.
2. Inferential: An inferential model cares about β^* that describes the relationship between x and y. We do not care about how accurate the predictions can be.

The reason why we use structural models is that it connects to a theory such that if we estimate the properties of this model, we can make **counterfactual** predictions which are not based on pre-existing patterns in the pre-existing data. This leads out the ***Lucas critique*** where he argues an econometric model includes agents' decisions based on both past outcomes and **expectations** of future outcomes, thus a change in the policy will also systematically alter the structure of econometric models. Thus, structural models can identify **invariant** properties and estimating them which escape the Lucas critique.

Other questions we wish to answer include "does A cause B", which are causal questions in need of causal models. They relate specific parts of the model to cause-and-effect, but we need to consider the possibility of reverse causality or other common factors. Note that causal models are nearly exclusively *inferential* as they are inferential in nature.

2 Data

Definition 2.1. ***Data or datasets*** are information used to answer econometric questions in interest. Specifically, econometric data are data generated by collective or individual choices and decisions.

Most data in econometrics are **non-experimental**, or specifically observational, where researchers are not actively manipulating the population, in contrast to **experimental** data which are generated with an *intervention.* **Quasi-experimental** data are non-experimental but can be *imagined* as experimental.

Data can be good or bad, but bad data does not mean erroneous data. Bad data usually refers to data that are not suitable for the research, which means the data is misrepresentative of the population, the size may be too small, or it mismatches the structure of the population. We need **meta-analyses** to ensure validity externally.

2.1 Structure of Data

Data are usually of three structures: cross-sectional, time series, or panel.

Cross-section data are a **snapshot of population**, where time is usually fixed, and there are a number of independencies. The observations are focused on the independencies (i).

Time series data are generated by observations on **time**, with only one independency. We observe how the independencies change with respect to time.

Combined together, they form panel data or longitudinal data, where there are both a number of independencies and a period of time. They can be balanced or unbalanced.

2.2 Data and Models

Models are designed to fit certain data.

Predictions from models are statements about the ***population***. However, we cannot directly observe the population. Therefore, we rely on ***samples*** collected from the population (**sampling process**) to make estimates about the population. Tandem to other properties of the data, we refer this as the **data-generating process**.

Observations are particular instances of the population we collected: specific instantiations of the population which form our sample. This is sometimes referred as ***statistical population***.

A general guideline for a sample is that the sample should be ***representative*** of the population, where sampling theory itself is another branch worth investigating. The key assumption we make here that the sample we use comes from random sampling such that the properties of a given observation do not affect the probability of it being sampled (randomness) and the properties of one observation do not affect other observations being sampled (independence). A violation is usually the presence of panel data, where time is involved in the scheme.

2.3 Variables

Definition 2.2. ***Variables*** are mathematical notation of the properties of an observation.

We phrase the population as ***random variable*** to be specific.

Generally, the types of variables are either **qualitative vs. quantitative** or **discrete vs. continuous**.

1. Qualitative: represent the qualities of an observation, cannot compare numerically. The only operation is to compare whether two variables are the same or not. These variables can be coded into whole numbers but they still represent qualitative variables.
2. Quantitative: represent quantities of an observation, they are **rankable**. They have a comparison operation defined for their values. They can come in either *ordinal* or *cardinal* values.

    1. Cardinal: data that have an order **and** a *relative size*
    2. Ordinal: data where only the order matters

3. Discrete: random variables that take on values from within delineated categories, usually represented with integers. They can either finite or infinite.
4. Continuous: random variables that take on an uncountable number of values, between any two values, there is always another value possible.

Notation-wise, we use capital letters like X to denote random variables. We use lowercase letters x to denote the value of a random variable, **which in most cases, is a vector of values**. Subscripts i represent each independency, and t represent time. Superscripts means exponentiation.

Another special case of variables is ***dummy/indicator variables***.

> ***Dummy or indicator variables*** are variables that are discrete and qualitative that take on two levels 0 and 1 to indicate the presence of a certain attribute.

If a qualitative variable has k levels, then it can be expressed in $k-1$ dummy variables, **NOT** k dummy variables which falls into the dummy variable trap that suffers from perfect multicollinearity.

For a dummy variable,

$$D_i = \begin{cases} 1 \\ 0 \end{cases}$$

We have,

$$E(D_i) = \sum_d d_i \cdot P(D_i = d_i) = P(D_i = 1)$$

They are important because:

1. flexibly describe qualitative variables
2. treated similarly to cardinal quantitative variables
3. most common type of data in real-world datasets

2.4 Parameters and Estimators

We want to use the **distribution** of the variables, specifically the **parameters** of such distributions to answer questions regarding key values or relationships. In this case,

Definition 2.3. A ***sample statistic*** which is used to estimate a specific parameter is called an ***estimator***.

The properties of an estimator usually include:

1. Unbiasedness: the expected value of the estimator is the parameter being estimated, $E(\hat{\mu}) = \mu$
2. Consistency: As the sample size become larger and closer to the population size, the estimator's value is the parameter being estimated, $\lim\limits_{n \to \infty} \hat{\mu} = \mu$
3. Efficiency: the estimator has the smallest possible error relative to a particular criterion

We use a **Greek letter** to represent a parameter of interest, μ, σ, ρ, where μ is the population mean, σ is the population standard deviation, and ρ is the population correlation.
We usually use $\hat{}$ to denote an estimator of a certain parameter, like $\hat{\mu}$ for μ. The exceptions include:

1. $\bar{X} = \hat{\mu}_X$
2. $s = \hat{\sigma}$
3. $r = \hat{\rho}$

Subscripts are used when there is ambiguity about the parameter or estimator.

2.5 Expectation

Nearly all parameters can be phrased in terms of expectations, specifically:

$$E(X) = \int_x x f(x) \mathrm{d}x \iff \sum_x x p(x)$$

By definition $E(X) = \mu_X$, if the population mean exists.

A way to form estimators of expected values is via the ***sample analogue principle***. The sample analogue of $E(X)$ is:

$$\hat{E}(X) = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Other important parameters include (co)variance and correlation:

$$C(X,Y) = E((X - E(X))(Y - E(Y)))$$

$$\sigma_X^2 = V(X) = C(X,X) = E((X - E(X))^2)$$

$$\rho_{X,Y} = \frac{C(X,Y)}{\sqrt{V(X)V(Y)}}$$

For example, the sample analogue of the variance of X is:

$$\hat{V}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Definition 2.4. The ***k-th moment*** of the distribution of X is $E(X^k)$

2.6 Inference, Sampling Distribution, and Bootstrapping

The reason why include all these relevant concepts is for statistical inference, where we

1. have an economic theory we want to test or examine
2. our theory predicts a relationship in the data
3. relate that prediction to statement about a parameter
4. select an estimator of the parameter of interest (e.g. $\hat{\theta}$)
5. perform a test of the statement (e.g. $\hat{\theta} > 0$)

We want to focus on the ***sampling distribution*** of an estimator to describe the parameter we are estimating well. The sampling distribution usually has these two properties to adhere to: ***Law of Large Numbers, and Central Limit Theorem***.

A technique that can used on samples with small size is ***bootstrapping***, where we simulate a large number of possible samples from the original samples. We assume the original sample is the best possible estimator of the population, so we compute our estimator of interest to create an ***empiral distribution*** analogous to the sampling distribution.

Finally, we can utilize hypothesis tests to see the relationship between the estimator and the parameter of interest. These tests give us robust statistical answers to our research question of interest.

3 Condition Expectation Functions

3.1 Conditional Expectation

The expectation of Y_i conditional on $X_i = x$ is,

$$E(Y_i|X_i = x) = \int_y y f_{Y|X}(y|x)\mathrm{d}y \iff \sum_y y f_{Y|X}(y|x)$$

and

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

The sample conditional expectation is thus defined to be

$$\hat{E}(Y_i|X_i = x) = \frac{1}{n_X}\sum_{i=1}^{n} y_i I(X_i = x) = \frac{1}{n_X}\sum_{i\in S_X} y_i$$

where S_X is the set of i, and $X_i = x$ and $|S_X| = n_X$

It is important because it links two variables so that the **conditional expectation** is the typical value of one variable **given another variable**, which is what econometric models are trying to do.

If our model predicts that X_i typically affect Y_i, this is a statement about the conditional expectation of Y_i given X_i. If we want to perform a test like a comparison of means, this is also a statement about conditional expectations.

3.2 Properties of Conditional Expectations

Some common properties of CEs include:

1. $E(aX_i + bY_i) = aE(X_i) + bE(Y_i)$
2. $C(X_i, Y_i) = E(X_iY_i) - E(X_i)E(Y_i)$
3. $C(X_i, a) = 0$
4. $C(aX_i, bY_i) = abC(X_i, Y_i)$
5. $C(X_i + Z_i, Y_i) = C(X_i, Y_i) + C(Z_i, Y_i)$
6. $V(X_i) = C(X_i, X_i) = E(X_i^2) - E(X_i)^2$
7. $V(aX_i) = a^2V(X_i)$

If X_i and Y_i are two independent variables, then

$$E(Y_i|X_i) = E(Y_i)$$

this is equivalent to saying

$$\forall x, \frac{\partial}{\partial x}E(Y_i|X_i = x) = 0$$

Theorem 3.1. The ***Law of Iterated Expectation (LIE)*** states that:

$$E(E(Y_i|X_i = x)) = E(Y_i)$$

3.3 Conditional Expectation as a Function

Consider $E(Y_i|X_i = x)$, as x changes, how does the level of expectation change? Thus,

Definition 3.1. A ***conditional expectation function (CEF)*** is defined as:

$$m(x) = E(Y_i|X_i = x)$$

This describes the average relationship between X_i and Y_i.

Some general tests include:

1. $\frac{\mathrm{d}m}{\mathrm{d}x} = 0$
2. $m(ax) = am(x)$
3. $m(0) = 0$

If X_i and Y_i are independent, then $m(x) = k \in R$, a constant.
The LIE states that $m(x) = \int_x m(x, z)\mathrm{d}z$

The decomposition property of CEFs include that, if we define $\epsilon_i = Y_i - E(Y_i|X_i)$, then we can decompose Y_i into $m(X_i)$ and ϵ_i, where the CEF describes part of Y_i completely determined by X_i and ϵ_i describes part of Y_i entirely independent of X_i. This also means that:

1. $E(\epsilon_i|X_i) = 0 \iff E(\epsilon_i X_i) = 0$
2. $C(\epsilon_i, f(X_i)) = 0$

The above CEF describes the population, but we only have the sample, so the **sample CEF** is

$$\hat{m}(x) = \hat{E}(Y_i|X_i = x) = \frac{1}{n_X}\sum_{i=1}^{n} Y_i I(X_i = x)$$

This is a naive estimator, as it is **non-parametric**. This is an example of a category of estimators called ***kernel density*** models, where the values of X_i are called ***bins or windows***.
The problem is the X_i can have lots of possible values, or it may be continuous. Some bins might be empty, or only have a small number of estimations.

The solution is to make assumptions about the CEF, or specifying the model. This brings out the technique of regression analysis. Some common regression models include:

1. Logistic regression: $m(x) = \frac{1}{1+e^{-\beta_0+\beta_1 x}}$
2. Poisson regression: $m(x) = \exp(\beta_0 + \beta_1 x)$

3. Polynomial regression: $m(x) = \beta_0 + \beta_1 x + \beta_2 x^2$
4. **Linear regression**: $m(x) = \beta_0 + \beta_1 x$

If we consider a linear regression model as CEF, then if $m(x) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$, then the linear regression would be

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

where k is the number of variables, and K is the number of parameters. $K = k + 1$. We can think of linear regression as an **approximation** to some non-linear CEFs.

4 Regressions

Key terminologies surrounding regressions first starts with the regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

where Y_i is the outcome, X_{ji} is an explanatory variable, ϵ_i is the residual, and β_j is the coefficient of X_{ji}.

If $k = 1$, then it is a simple regression; if $k > 1$, then it is a multiple regression. Multivariate regression refers to a vector of outcomes. The most important term we are interested in is β.

We should ask the question if the model is well-defined, what are the meaning of βs, do we know β is unique.

4.1 Simple Regression

We start with CEFs:
The CEF is defined by the condition that $E(\upsilon_i|X_i) = 0$. If our regression is a good approximation of the CEF, then it needs to meet these conditions, $E(\epsilon_i) = 0$ and $E(\epsilon_i X_i) = 0$

Thus, for a linear regression equation

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

we thus yield the moment equation conditions:

$$\begin{cases} E(\epsilon_i) = 0 \\ E(\epsilon_i X_{1i}) = 0 \\ \vdots \\ E(\epsilon_i X_{ki}) = 0 \end{cases}$$

Specifically, we say that the linear regression coefficients βs are defined as:

$$(\beta_0, \beta_1, \ldots, \beta_k) \equiv \arg \min_{b_0, b_1, \ldots, b_k} E((Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2) = \arg \min_{b_0, b_1, \ldots, b_k} E(\epsilon_i^2)$$

where $E(\epsilon_i^2)$ is the mean squared error.
This is equivalent to saying

$$\begin{cases} \frac{\partial E}{\partial b_0} = E(\epsilon_i) = 0 \\ \frac{\partial E}{\partial b_1} = E(\epsilon_i X_{1i}) = 0 \\ \vdots \\ \frac{\partial E}{\partial b_k} = E(\epsilon_i X_{ki}) = 0 \end{cases}$$

The necessary condition for a unique solution is that X_{ji} must not be constant, and must not be redundant (not collinear). This is the condition of ***no perfect collinearity***.

After solving for a simple linear regression, we have

$$\beta_1 = \frac{C(X_i, Y_i)}{V(X_i)}$$

and we also know $\frac{\partial Y_i}{\partial X_i} = \beta_1$, meaning it is the ***marginal effect*** of X_i. Algebra also shows that

$$\beta_1 = \rho_{X,Y} \cdot \frac{\sigma_Y}{\sigma_X}$$

Notice that "linear" refers to the coefficients **NOT** the variables, thus, we can also have a polynomial regression with terms exponentiated.

$$g(Y_i) = \beta_0 + \beta_1 f_1(X_{1i}) + \cdots + \beta_k f_k(X_{ki}) + \epsilon_i$$

These can be models of log-log models, log-linear models, or linear-log models. In this case, remember when calculating coefficients, apply **chain rule**. Some useful approximations for smooth continuous functions include Taylor series or Fourier series.

4.2 Multiple Regression

We consider the **Frisch-Waugh** anatomy theorem. Essentially, we want to set up another regression like this:

$$X_{1i} = \alpha_0 + \alpha_2 X_{2i} + \cdots + \alpha_k X_{ki} + \tilde{X_{1i}}$$

from the original multiple regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

This eventually gives the ***regression anatomy equation***:

$$\beta_j = \frac{C(\tilde{X_{ji}}, Y_i)}{V(\tilde{X_{ji}})}$$

This means, it is proportional to the correlation of Y_i and the part of X_{ji} **not explained by other variables**. This is referred to as **controlling** for other variables.

The simple regression coefficient is a special form of the regression anatomy equation where $X_i = \theta_i + \tilde{X_i}$

4.3 Relationship between CEFs and Regressions

If CEF of Y_i given X_i is linear, it is linear regression, where all X_{ji} are jointly normal.

Linear regression is the best linear approximation in Mean Squared Error (MSE) sense, meaning that if CEF is not linear, the regression is the best possible approximation.

Linear regression is the best **linear predictor** of the CEF, in the MSE sense.

4.4 Regression as Model

When constructing a linear regression, it is not simply a variable selection problem. It can be very flexible as the polynomial regression above, or as another important category, where dummy terms are involved in the regression.

They can be directly included in the model. However, notice if D_i has k levels, one must use $k-1$ dummies, as including all of them leads to perfect collinearity. This is referred to as the ***dummy variable trap***.

Dummies can also be outcome variables, which are referred to as a ***linear probability model*** which will be discussed later. One needs to be careful with **heteroskedasticity** and there are **no limitations on** $\hat{Y}_i$

5 Estimation

After constructing our regression model, we hope to create **estimators** for key parameters of interest. We then can establish properties of these estimators giving us the opportunity to perform statistical inference.

Monte Carlo simulation can be used to test whether novel models are accurate or not.

There are essentially three types of estimators:

1. Sample Analogue
2. Method of Moments
3. Ordinary Least Squares

5.1 Sample Analogue Estimators

The sample analogue estimator simply adds a ˆ on top of the statistics involved. For example,

$$\hat{\beta}_1 = \frac{\hat{C}(X_i, Y_i)}{\hat{V}(X_i)}$$

However, for multiple regression, we need to use iterated regression, that is, the Frisch-Waugh anatomy theorem. The problem with this estimator is that

1. It is difficult to calculate
2. Hard to find an analytical expression for coefficients
3. Hard to analyze properties of the estimator $\hat{\beta}_j$
4. Matrix algebra can make the iterations go away

5.2 Method of Moments

The variable coefficients in the sample analogue estimator comes from moment equations. In this case, we use **sample moments**, that is finding estimators that satisfy

$$\begin{cases} E(\epsilon_i) = 0 \\ \vdots \\ E(\epsilon_i X_{ki}) = 0 \end{cases}$$

For a simple regression, it may look like this

$$\begin{cases} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ \frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \end{cases}$$

This can be applied to multiple regressions more easily. We can use linear algebra to solve those sample moment equations.

5.3 Ordinary Least Square

First of all, OLS is a method not a model.

In a regression model, there are data we can see, and data we cannot see, where β_j is predictive and the residual is everything else. Since the CEF is the most predictive model possible in squared deviation sense, then we approximate the CEF.

This means, we hope to choose $\hat{\beta_j}$ to make the total ϵ_i^2 as small as possible, mathematically,

$$\min_{b_j} \sum_{i=1}^{n} \epsilon_i^2$$

5.4 Which Estimator?

All of these methods have the same properties and the same assumption, but we prefer Method of Moments as it is fast to compute relying on matrix algebra.

After estimating the coefficients, we can have the **estimated or sample residuals** from

$$\hat{\epsilon_i} = Y_i - \hat{\beta_0} - \hat{\beta_1} X_{1i} - \cdots - \hat{\beta_k} X_{ki}$$

The **fitted or predicted values** are that

$$\hat{Y_i} = \hat{\beta_0} + \hat{\beta_1} X_{1i} + \cdots + \hat{\beta_k} X_{ki} \approx \hat{m}(X_i)$$

and $\hat{\epsilon_i} = Y_i - \hat{Y_i}$

6 Evaluation

The model's performance depends on the model itself and the research question. If the model is well-defined and it can explain the research question well, it does not matter if the model does not perform well in other dimensions.

6.1 Model vs. Non-model

Since the relationship between Y_i and X_i are explained by the CEF and the residual, we can think about "how much do the explanatory variables actually explain"? If our model is well, then we should rely on the residual as little as possible.

This is a predictive type of statement, where a model does well in predicting Y_i does a good job answering predictive questions. It can also help inferential questions, but not essential. However, what is objectively "small"?

Once we have estimated the model, we would have sample residuals and fitted values. They estimate the true residual and fitted values, meaning there is a difference in the **theoretical** and **empirical** performance. ˆ can simply removed to represent the population.

6.2 Loss Functions

Mean Squared Error (MSE)

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon_i}^2$$

Mean Absolute Deviation (MAD)

$$\frac{1}{n}\sum_{i=1}^{n}|\hat{\epsilon_i}|$$

Supremum Style Deviation (SSD)

$$\max_i |\hat{\epsilon_i}|$$

Other loss functions include AIKE, or KLS

Theorem 6.1. The ***Gauss-Markov*** Theorem states that the OLS coefficients provide the **smallest variance** among all possible unbiased linear estimators.

The MSE approach:

1. Treats positive and negative variance in the same way
2. Penalize many small errors less than a few large errors
3. Differentiable and smooth
4. An absolute number

But it does not provide a way of differentiating variations:

1. Model does not explain the data well
2. Data naturally has lot of variation

6.3 R-squared

We introduce R-squared to differentiate between errors caused by the model and error natural to data:

$$\begin{cases} SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \text{ total} \\ SSR = \sum_{i=1}^{n}(Y_i - \hat{Y})^2 \text{ residuals} \\ SSE = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \text{ errors} \end{cases}$$

and R-squared is defined to begin

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

that is the ratio of variation explained by the model and the total variation.

1. SST: sum of squared total
2. SSR: sum of squared residuals
3. SSE: sum of squared estimates

The properties include that:

1. $R^2 = 0$: model is non-explanatory
2. $R^2 = 1$: model is a perfect-fit
3. $R^2 \in [0, 1]$, and is unitless

When we increase model sizes, R^2 is guaranteed to increase and become closer to 1, thus, it should not be used to select between models of different number of variables.

If we try to penalize a model for having more parameters, we can use the **adjusted R-squared**.

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-K}$$

But it loses interpretability and other useful properties, and it can be less than 0 or greater than 1.

6.4 Overfitting

Mathematically, it is always true if the "prediction" we are trying to make is an explanation of the existing sample, but it may not always be true predicting out of sample. Thus, a good in-sample fit and a bad out-of-sample fit is referred to as ***overfitting***.

We can often reserve 10% of the data so that, we use these data to validate the model.

We may also do cross-validation, where the data is divided to k equal folds. For each fold, we estimate the model using the other $k - 1$ folds, and test on the un-used fold.

Then we consider performance across folds to see how the model does.
Bad controls (or over-controlling) can be case for model reduction.

6.5 Interpreting Coefficients

Note that,

$$\beta_j = \frac{C(Y_i, \tilde{X_{ji}})}{V(\tilde{X_{ji}})} = \frac{\partial m(X_i)}{\partial X_{ji}}$$

we ask about the impact of changing an explanatory variable on the CEF.
X_{ji} as a variable does not need to enter directly into the regression, it can enter more than once, or interacts with another variable.

Similarly, for a dummy variable, we have

$$\beta_j = \frac{\Delta m(X_i)}{\Delta X_{ji}}$$

meaning the change in typical value of Y_i when X_i is present versus when it's not present.

The qualitative variables can be expressed as a set of dummies, thus, a similar tactic can be used to interpret qualitative variables when using $k-1$ dummies to represent it.

Furthermore, what if we believe two different groups have different effects on the outcome? We can use interactions like,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 G_i + \beta_3 X_i \times G_i + \epsilon_i$$

Then,

$$\frac{\partial m}{\partial X_i} = \beta_1 + \beta_3 G_i$$

given different values of G_i, the impact will also be different.

6.6 Over-controlling

Over-controlling refers to including controls that restrict the purpose of the study. This is usually the result of not having a clear conception of what the situation being modeled means economically. In econometrics, this can mean adding structure or justifying variable choices. In causal analysis, this means thinking about the cause-and-effect relationships which create the data.

6.7 Interpreting Estimates and Precision of Estimates

Different samples will lead to slightly different estimates (maybe even very different) of the coefficients, which leads to different interpretations. This also applies to predictions, as it will lead to different fitted values.

Fundamentally, coefficients also have **sampling distributions**. Thus, we need to answer how much the coefficients vary to determine the uncertainty of the estimates.

This depends on how big our sample is and what we assume about the structure of the model.

7 Small Sample Properties

Right now, we can compute **point estimates** from our regression, but we do not know how accurate our estimates are. Thus, we need to know the sampling distribution of the relevant estimates, $\hat{\beta}_j, \hat{Y}_i, \hat{\epsilon}_i$.

Small sample properties are those properties that are true regardless of sample size, and large sample properties are asymptotic properties obtained as $n \to \infty$.

The assumptions we have made about a regression is that X_i and Y_i are randomly sampled, and for a linear regression equation, $E(\epsilon_i|X_i) = 0$ (strict exogeneity assumption). However, we only need the weak exogeneity condition, $E(\epsilon_i X_i) = 0, E(\epsilon_i) = 0$ for a regression to work. We also assume no perfect multicollinearity.

7.1 Unbiasedness

The bias in an estimator is described as

$$B = E(\hat{\theta}) - \theta$$

such that the estimator is unbiased if $B = 0$.

The first property is that $\hat{\beta}_j$ is **unbiased**. That is

$$E(\hat{\beta}_j) = \beta_j$$

The precision of this estimate is then mathematically expressed as:

$$V(\hat{\beta}_j|X_i) = \frac{\sum_{i=1}^n \hat{\tilde{X}}_{ji}^2 \sigma^2(X_i)}{(SST_j(1 - R_j^2))^2}$$

Essentially, to reduce the standard error, we hope to have X_{ji} to have more variation relative to Y_i, but also avoid being explained by other explanatory variables.
Notice that if X_{ji} is perfectly collinear with other X variables, the denominator has $R_j^2 = 1$, blowing up the standard error. A measure of collinearity is ***variance inflation factor***, which is simply $\frac{1}{1-R_j^2}$

We also need to estimate the numerator, and the assumption we have made is ***homoskedasticity***, or specifically $\sigma^2(X_i) = \sigma^2$, a constant. In this case, $V(\hat{\beta}_j|X_i) = \frac{\sigma^2}{SST_j(1-R_j^2)}$. We estimate σ^2 by having $E(\epsilon^2)$, so the sample analogue for σ^2 is

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

and correcting by the degrees of freedom, the fraction should be $\frac{1}{n-(n-k-1)}$, or specifically, this estimator is also unbiased.

7.2 Distribution of Estimates

By the unbiasedness and the Gauss-Markov theorem, we know $\hat{\beta}_j$ is unbiased and efficient, now we focus on the sampling distribution.

If X_i is homoskedastic and normally distributed, then $\hat{\beta}_j \sim N(\beta_j, \sqrt{V(\beta_j)})$. This is derived from the assumption that the residual is normally distributed that $\epsilon_i|X_i \sim N(0, \sigma)$ (may not be a good one).

Thus, the inferential formla are either

$$Z \equiv \frac{\hat{\beta}_j - \beta_j}{\sqrt{V(\hat{\beta}_j)}} \sim N(0,1)$$

$$t \equiv \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}(\hat{\beta}_j)}} \sim t_{n-k-1}$$

8 Large Sample Properties

Occasions we apply the large sample properties include the normality assumption on residuals may be too strong, there is a large amount of data and need to use other estimators and properties, or using techniques that rely on large sample theory.

It relies on two core theories: ***Law of Large Numbers, Central Limit Theorem***.

Definition 8.1. Consider an estimator $\hat{\theta}$, it ***converges in probability*** to θm, when as $n \to \infty$, the probability of any difference between the estimator and the actual value goes to 0.

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| > 0) = 0 \iff \hat{\theta} \to_p \theta$$

This represents consistency, a weaker condition.

Theorem 8.1. The ***Law of Large Numbers*** states that for a random variable X_i, we have

$$\hat{E}(X_i) \to_p E(X_i)$$

By this theorem, we can also demonstrate that $\hat{\beta}_j$ is consistent, the weak exogeneity condition.

Definition 8.2. Consider an estimator $\hat{\theta}$, and F being a CDF, we say it ***converges in distribution*** to F when

$$\forall x, \lim_{n\to\infty} P(\hat{\theta} \leq x) = F(x)$$

we write $\hat{\theta} \sim_a F$

This brings us to

Theorem 8.2. The ***Central Limit Theorem*** states that if a random variable X_i has mean of $E(X_i)$ and variance σ_X^2, then

$$\frac{\hat{E}(X_i) - E(X_i)}{\frac{1}{\sqrt{n}}\sqrt{E(X_i^2) - E(X_i)^2}} \sim N(0,1)$$

as long as $E(\epsilon_i | X_i) = 0$, with no need to assume it being normal. This is equivalent to

$$\frac{\sqrt{n}(\hat{E}(\cdot) - E(\cdot))}{\sqrt{V(\cdot)}} \sim_a N(0,1)$$

Thus, if $n \to \infty$, as long as $E(\epsilon_i | X_i) = 0$, then the estimates are consistent;

furthermore, if the residuals are also normally distributed, the estimates remain efficient (error is homoskedastic).

9 Inference

Now we know how to build, estimate the model with sampling distribution, we now can test the relevant hypotheses.

9.1 Testing Single Coefficient

This is a hypothesis test that looks like:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

The distribution follows:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{SE}_1} \sim_a t_{n-k-1}$$

The test statistic here would be

$$t^* = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

9.2 Test Combination of Coefficients

In this case, we use the delta methods to test their combination. That is, given a linear regression, our null hypothesis is

$$H_0 : \sum_j a_j \beta_j = a$$

Note that all the $\hat{\beta}_j$s are jointly normal, so their linear combination is also normal. The test statistic is thus

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{SE(\hat{\beta}_1 - \hat{\beta}_2)}$$

for a hypothesis $H_0 : \beta_1 = \beta_2$

An alternative approach is to create new variable like $X_{1i} + X_{2i}$ and transform the previous regression model to estimate one specific coefficient.

9.3 Multiple Restrictions

Notice that performing multiple comparisons is not equivalent to perform each comparison independently. We wish to introduce the joint size with Bonferroni's correction

$$\frac{\text{p-value}}{\text{\# of tests}}$$

An example null hypothesis is

$$H_0 : \sum_m a_m \beta_m = a, \text{ and } \sum_n b_n \beta_n = b$$

The correlated test is the F test. The intuition is that if the hypothesis is true, then there are certain restrictions on the model such that the fit for an unrestricted model is almost the same as the fit for a restricted model. If it is not true, then the fit for the unrestricted model is better than that for the restricted one. The test statistic is thus

$$F \equiv \frac{(SSR_r - SSR_u)/m}{SSR_u/(n-k-1)} \sim F_{m,n-k-1}$$

The last thing is that we can test whether a coefficient is 0 or not such that the variable can be excluded from the model. We test this for each $j = 1, \ldots, k$, and yield model significance.

10 Issues with Regression

There are mainly two technical issues in a regression model that are fixable. One is multicollinearity and the other is heteroskedasticity.

10.1 Multicollinearity

Given an explanatory variable written like:

$$X_{1i} = \theta_0 + \theta_2 X_{2i} + \cdots + \theta_k X_{ki} + \tilde{X_{1i}}$$

If $\tilde{X_{1i}} = 0$ for all i, then this implies perfect collinearity. $R_1^2 = 1$. In some cases, it is almost collinear, where $\tilde{X_{1i}} \approx 0$ for all i, here $R_1^2 \approx 1$.
All variables are to some extent collinear to each other, problems occur when multiple variables measures the same property, and they are highly collinear. The variance inflation factor measures the extent of collinearity. Recall that

$$VIF = \frac{1}{1 - R_j^2}$$

given that the data we are using is homoskedastic.

10.2 Heteroskedasticity

To "solve" heteroskedasticity, one needs to be careful about what variables to choose. The principle is to choose variables that keep as much as explanatory power as possible, and keep non-collinearity.

One may use the **Principal Components Analysis**, or specifically decomposition to isolate the "latent" aspects of a dataset. We can use the line of best fit as an axis and examine whether the data are heteroskedastic or not. However the new axes would have no economic interpretation.

Returning to the meaning of heteroskedasticity, it means variance in variances. That is,

$$V(\epsilon_i | X_i) = \sigma^2(X_i)$$

Two notable tests to detect heteroskedasticity is to use **White's test** or **Breusch-Pagan test**.

White's test focuses on the relationship between the squared residual and the fitted values, where

$$\hat{\epsilon_i}^2 = \beta_0 + \beta_1 \hat{Y_i} + \beta_2 \hat{Y_i}^2 + \eta_i$$

If β_1 and β_2 are significant, there is likely heteroskedasticity. We use the test statistic $t = nR^2$ from this regression, and turns out it converges in distribution to a χ^2 distribution, where the degrees of freedom is the number of variables interacting.
If we fail the White's test, we no longer assume homoskedasticity and turn to robust

standard errors. That is, finding a way to estimate:

$$\sum_{i=1}^{n} \hat{\tilde{X}}_{ji}^{2} \sigma^2(X_i)$$

10.3 Linear Probability Models

Linear probability models are linear regressions where the outcome variable is a dummy variable. It is heteroskedastic by construction. One way is to use weighted OLS to correct for the standard errors such that, we find w_i for each ϵ_i that

$$w_i \epsilon_i \sim_a N(0, w_i b_i)$$

and $b_i = \sqrt{V(\epsilon_i|X_i)}$

10.4 Miscellaneous Things

Many assumptions of OLS can be relaxed or violated with additional work.

One issue is the representativeness of the sample chosen. The first assumption is that the data are independent of each other, but many datasets deliberately are non-random. These approaches record **weights** to "reverse" the non-randomness. Usually, frequency weights or sample weights are used to determine the proportion of a certain sample in the population.

Another issue is clustering, where the data may come from a common group, that may be subject to idiosyncratic effects. Another version of the sandwich (White) estimator can be used to address this issue.

11 Omitted Variable Bias

A regression specification is an **assumption** about the structure of the CEF. Misspecification occurs when we get the assumptions wrong. One way is to not include certain variables as we have many variables to choose from when constructing the model.

In predictive models, we covered this problem by considering the fit of the regression. But in inferential models, we want our estimates to be correct, which amounts to evaluating whether or not we have specified the CEF correctly.

If we have our CEF correct, we are saying $E(\epsilon_i|X_i) = 0$, the strict exogeneity assumption, which further implies $E(\epsilon_i) = 0, E(\epsilon_i X_i) = 0$ to solve for the regression. However, what if $E(\epsilon_i|X_i) \neq 0$, this is ***endogeneity*** as both Y_i and X_i are jointly affected by the part of the model. Particularly, it arises in causal models.

11.1 Omitted Variables

We can first consider a "true" model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \gamma W_i + \epsilon_i$$

but we have a model of what we believe to be true:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

Recall that,

$$\gamma = \frac{C(Y_i, \tilde{W_i})}{V(\tilde{W_i})}$$

If $\gamma = 0$, then our model did not misspecify, as W_i is left out.
The other special case is to consider that $E(W_i|X_{1i}, X_{2i}) = \alpha$, which is constant. In this case $e_i = \gamma W_i + \epsilon_i$, so

$$E(e_i|X_{1i}, X_{2i}) = \gamma\alpha$$

In this specific case, $b_1 = \beta_1, b_2 = \beta_2$, but $b_0 = \beta_0 + \gamma\alpha$.

11.2 Omitted Variables Bias

When $E(W_i|X_{1i}, X_{2i})$ is not constant, we refer to this situation as ***omitted variables bias***.

Firt we have the relationship that $E(e_i|X_{1i}, X_{2i}) = \gamma E(W_i|X_{1i}, X_{2i})$

Notice that by the regression anatomy equation, we still have

$$b_1 = \frac{C(Y_i, \tilde{X_{1i}})}{V(\tilde{X_{1i}})}$$

After substituting Y_i in with the true model and the relationship between e_i and

W_i, we eventually use algebraic means to get that

$$b_1 = \beta_1 + \frac{C(Y_i, \tilde{W_i})}{V(\tilde{W_i})} \cdot \frac{C(W_i, \tilde{X_{1i}})}{V(\tilde{X_{1i}})}$$

where if we have a regression

$$W_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \tilde{W_i}$$

then

$$\begin{cases} \gamma = \frac{C(Y_i, \tilde{W_i})}{V(\tilde{W_i})} \\ \alpha_1 = \frac{C(W_i, \tilde{X_{1i}})}{V(\tilde{X_{1i}})} \end{cases}$$

This means that, in b_1, not only does it have the actual β_1, but also the direct impact of W_i on Y_i, and the impact of other variables on the omitted variable (we can use controls to make this as weak as possible).

In practice,

$$\gamma = \rho_{Y_i, \tilde{W_i}} \cdot \frac{\sigma_{\tilde{W_i}}}{\sigma_{Y_i}}$$

$$\alpha_1 = \rho_{W_i, \tilde{X_{1i}}} \cdot \frac{\sigma_{X_{1i}}}{\sigma_{W_i}}$$

If other variables ”eliminate” the effect of W_i on Y_i, OVB gets smaller.

This provides us with two rationale of using controls: they are included to capture the effect of a variable we cannot see which we think is relevant, and they isolate a part of X_i which is not affected by the omitted variables.

11.3 Ramsey RESET Test

If the model is well-specified, then conditional on the explanatory variables, the fitted values should have no predictive power, that is

$$Y_i = \theta_0 + \theta_1 X_{1i} + \cdots + \theta_k X_{ki} + \delta_1 \hat{Y_i} + \delta_2 \hat{Y_i}^2 + \cdots + \epsilon_i$$

We expect δ_j to be small, close to 0. We test this with F_{n-k-3} test.

However, one should not rely on this test as it is a functional form so it does not tell you where the model failed. And it cannot always detect OVB. It can help detect bad models, but cannot determine if the model is good.

Fundamentally, to handle OVB, one should choose to use good controls, or sign the bias to provide context about overestimation or underestimation, or use a model or technical that handle OVB by design.

12 Causal Models

In inferential questions, we often want to answer a cause-and-effect statement. We use causal modelling for this specific purpose.

"Correlation is not causation", so first consider the regression anatomy equation:

$$\beta_1 = \frac{C(Y_i, \tilde{X_{1i}})}{V(\tilde{X_{1i}})} \propto \rho_{Y_i, \tilde{X_{1i}}}$$

This is a correlation, not causation.

The framework for a causal model is the ***Potential Outcomes Model***. In this framework, we define:

1. Y_i: outcome variable
2. D_i: a dummy variable that refers to "treatment"
3. X_i: covariates (controls)

Then $Y_i(1)$ means the potential outcome when treated, and $Y_i(0)$ means the potential outcome when untreated. We call

$$\Delta_i = Y_i(1) - Y_i(0)$$

the **treatment effect**.

The **average treatment effects** include:

1. $ATE = E(\Delta_i) = E(Y_i(1) - Y_i(0))$, this is the whole population.
2. $ATT = E(Y_i(1) - Y_i(0)|D_i = 1)$, this is the subpopulation which was treated.
3. $ATU = E(Y_i(1) - Y_i(0)|D_i = 0)$, this is the subpopulation which was untreated.

This runs into the **Fundamental Problem of Causal Inference**, where each individual is either treated or not, so we cannot observe both potential outcomes, meaning individual treatment effect is **unobservable**. This means ATE and ATT are fundamentally unobservable. This is not a statistical problem but a model problem.
If a person's outcome depends only on their own treatment status, then we have

$$Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$$

This is referred to as the ***stable unit treatment value assumption (SUTVA)***.

If we compare this with a linear regression,

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

then we have

$$\beta_1 = E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$$

This can use the causal decomposition where:

$$\begin{aligned}\beta_1 &= (E(Y_i(1)|D_i = 1) - E(Y_i(0)|D_i = 1)) + (E(Y_i(0)|D_i = 1) - E(Y_i(0)|D_i = 0)) \\ &= ATT + B\end{aligned}$$

The B part refers to the **selection bias**, that is, it occurs when the treated group would have behaved differently if the treatment was not available in the first place. If the treated group and the untreated group behaved same if the treatment was unavailable in the first place, then $B = 0$.

It is highly common to see selection bias, as "treatment" is usually an agent's voluntary choice. To diagnose selection bias, one can ask:

1. Is there any plausible reason why the treatment group would have behaved differently from the control group if the treatment is unavailable?
2. Is this reason not captured or controlled by the other covariates in the models?

Solutions to this is that it can be minimized if the assignment of treatment is random, or they can be signed to provide lower/upper bound of the treatment effect.

Mathematically, the ***conditional independence assumption*** says that

$$E(Y_i(0)|D_i) = E(Y_i) = E(Y_i(1)|D_i)$$

for all values of D_i. In other words, if D_i is **as good as randomly assigned** given X_i, selection bias would be zero.