Review econometrics

Lazy

December 10, 2006

Foreword

Although I have spent more than 30 hrs to prepare this review notes (average 3hrs per chapter), I cannot promise you this note is error-free. It would be best to question anything seems unusual or unfamilar.

The first thing I want to share with you is not the technical knowledge of how to calculate but how to study econometrics efficiently?

1. Learn the assumption of models
2. Try to understand the proofs (but don't overdo it!)
3. Memorize the main results
4. Remember the limitation, testing, forecasting of the models
5. Do all exercises available! (including past papers!)

Instead of putting the material reference at the back, I decide to put it in the front cover:

1. Econometric Analysis (5th edition) by William H. Greene, Prentice Hall
2. Introduction to Linear Regression Analysis (3rd edition) by Douglas C. Montgomery, Elizabeth A. Peck and G. Geoffrey Vining, Wiley Interscience Publication
3. Statisical Inference (2nd edition) by Geogre Casella and Roger L. Berger, Duxbury

Finally, let me wish you good luck in exam!

.

Contents

1 A Review of Probability and Statistics

1.1 Sample space, random experiment, events and probability

$$\text{Sample Space} \overset{\text{Random Experiment}}{\Longrightarrow} \text{Elementary Event} \overset{\text{combination}}{\Longrightarrow} \text{Event} \overset{\text{assignment}}{\Longrightarrow} \text{Probability}$$

1. Sample Space: all possible outcomes of random experiment
2. Random Experiment: (i) all outcome known (ii) particular trial not known (iii) can be repeated
3. Elementary Event: particular outocme of random experiment
4. Event: elementary event(s), subset of sample space
5. Probability: numbers from 0 to 1 assigned to events

1.2 Kolmogorov's definition of probability

This is a technical definition to eliminate any possible paradox and ambiguity.

1. $0 \leq \Pr(E) \leq 1$
2. $\Pr(S) = 1$
3. $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ if $A \cap B = \varnothing$.

where S is sample space, $E,\ A,\ B$ are events. $\cup$ and $\cap$ means union and intersection. $\varnothing$ means null set.

1.3 Conditional Probability and Statistical Independence

This can be thought as reduction in sample space. If we know something has happened, we know something has not happened. Therefore, the original probability should be revised. The revision is done by using the concept of conditional probability.

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

From conditional probility, we can derive the independence concept. If realization of event A has no effect on event B, we call the two event statistically independent. That is,

$$\begin{aligned}&\Pr(A \mid B) = \Pr(A) \text{ and } \Pr(B \mid A) = \Pr(B)\\&[\Rightarrow \Pr(A \cap B) = \Pr(A) \times \Pr(B)]\end{aligned}$$

1.4 Random Variable

This is to abstract the concept of events. Here, we usually assume events to be numbers rather than heads, tails, diamond, club, heart and spade.

Formally, random variable is the method(function) of assigning non-numbers (events) to numbers numbers (value of random variable).

$$\text{Elementary Event} \overset{\text{random variable}}{\Longrightarrow} \text{Numbers}$$

If numbers representing events are discrete, it is discrete random variable.

If numbers representing events are continuous, it is continuous random variable.

How can we describe a random variable? Distribution function.

$$F(a) = \Pr(X \leq a)$$

How distribution function correlates probability?

Probability mass function for discrete random variable.

$$\begin{aligned}\Pr(X = a) &= f(a)\\F(a) &= \sum_{x \leq a} f(x)\end{aligned}$$

Probability density function for continuous random variable.

$$\begin{aligned}\Pr(a \leq X \leq b) &= \int_a^b f(x)\,dx\\F(a) &= \int_{-\infty}^{a} f(x)\,dx\end{aligned}$$

How to complicate a random variable? Add more random variables. Then you have to use joint distribution to describe.

$$F(a,b) = \Pr(X \le a, Y \le b)$$

Of course, you could get joint density function if both random variables are continuous where you will need to compute double integration or partial differentiation to play around with them.

1.5 Functions on random variable

1. Mean (1st moment, mathematical expectation)

   Central Tendency: A single number to represent data.

   $$E(X) = \begin{cases} \sum\limits_{x \in S} x f(x) & \text{if } X \text{ is discrete} \\ \int_S x f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

2. Variance (2nd moment about mean)

   Dispersion: A number to describe the spread of data. Accuracy of mean.

   $$Var(X) = E(X - E(X))^2 = \begin{cases} \sum\limits_{x \in S} (x - E(X))^2 f(x) & \text{if } X \text{ is discrete} \\ \int_S (x - E(X))^2 f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

3. Covariance

   Relationship between two variables. Positive means moves in the same direction and negative means move in opposite direction. Zero means no (linear) relationship.

   $$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

4. Correlation coefficient

   Normalized covariance. Only assumes value from $[0,1]$.The rescale is done by dividing standard deviation of both variables.

   $$\rho_{xy} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

2 Special probability distribution

This chapter only covers continuous random variables. Bascially, most of all distributions needed are all variation of normal distribution.

$$\begin{tikzcd}
 & & & \text{Student's t} \\
\text{Normal} \arrow[r, "Z=\frac{X-\mu}{\sigma}"] & \text{Standard Normal} \arrow[r, "\chi_k^2=\sum_{i=1}^{k} Z_i"] & \text{Chi-square} \arrow[ur, "t_k=\frac{Z}{\sqrt{\chi_k^2/k}}"] \arrow[dr, "F=\frac{\chi_m^2/m}{\chi_n^2/n}"'] & \\
 & & & \text{F}
\end{tikzcd}$$

2.1 Uniform distribution

By the name, we know its density should be uniform in the sample space. So, if $X \sim U(a)$, the density function should be

$$f(x)=\begin{cases} \frac{1}{a} & \text{if } 0 \le x \le a \\ 0 & \text{otherwise} \end{cases}$$

Use: to model event that might happen in equally likely manner

Exponential distribution

Again, by the name, we know its density should be in exponential form. So, if $X \sim$exponential(θ), the density function should be

$$f(x)=\begin{cases} \frac{1}{\theta}e^{-x/\theta} & \text{if } 0 \le x < \infty \\ 0 & \text{otherwise} \end{cases}$$

Use: to model radioactive decay, time for light bulb to burn out

2.2 Normal distribution

First used by de Moivre to approximate binomial distribution for large number of trial (large n). Later, Laplace and Gauss used this to model the errors of experiment. Therefore, it is also sometimes referred as Gaussian distribution. The density function might look intimidating at the first hand, but it will be much friendly amd ‘normal’ if we look at it more.

If $X \sim N\left(u, \sigma^2\right)$,the density function will be

$$f(x) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} & \text{if } -\infty < x < \infty \\ 0 & \text{otherwise} \end{cases}$$

Uses: IQ, height, wegiht and many many.... that's why we call it 'normal'. By the way, remember it can be used to model error. So in many model, error is usually modelled by normal distribution.

2.3 Standard normal distribution

One interesting and convenient property of normal distribution is that it is completely characterized by its mean μ and variance σ^2. So, is there any basic form? Yes, set $\mu = 0$ and $\sigma^2 = 1$.In the other words, standard normal is $N(0, 1)$.

Hence, if $X \sim N(0, 1)$, the density function would then be

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} & \text{if } -\infty < x < \infty \\ 0 & \text{otherwise} \end{cases}$$

Uses: Save the Earth. Honestly, it has no special use in statistics. But it serves a great job to environmental protection! Since the density function could not be handled by pencils, we need to check tables for calculation. Given any normal distribution can be standardize by transformation ($Z = \frac{X-\mu}{\sigma}$), only one normal table is need.

2.4 Log-normal distribution

A natural extension of normal distribution. It limits the sample space to non-negative real numbers and changes it from symmetric to left-skewed.

Mathematically, if X is normal, $Y = \ln X$ is lognormal. The density function of Y is so complicated that I dont remember so I think you might wish not to memorize it.

$$f(y) = \begin{cases} \frac{1}{y\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\ln(y)-\mu}{\sigma}\right)^2\right\} & \text{if } 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Uses: stock variation, personal incomes...things that cannot assume negative value.

2.5 Chi-square distribution

It should be X^2 distribution rather than χ^2 distribution. This error is due to the printing error in taking sorts. Actually, the X is actually the standard normal distribution. So, χ^2 is actually Z^2 where Z is standard normal.

To generalize this, we add a parameter degree of freedom k to represent the number of Z^2 added together. Therefore, $\chi_1^2 = Z_1^2$, $\chi_2^2 = Z_1^2 + Z_1^2, ...,$ $\chi_k^2 = Z_1^2 + Z_2^2 + ... + Z_k^2$.

Its distribution function is so complicated that I dont write it down.

Uses: famous χ^2 test

Remark: $\chi_1^2 = Z^2$ and $\chi_2^2 =$ exponential$(\theta = 2)$

2.6 Student's t-distribution

Student is a fictitious name becasue the inventor, Gosset, was not allowed to use his own name. Anyway, all you need to know is to remember is that if it is formed by having a standard normal random variable over square root of Chi-squared random variable divided by its degree of freedom. That is,

$$t_k = \frac{Z}{\sqrt{\chi_k^2/k}}$$

One more thing to remember it is something similar to normal distribution though its tail is much thinner. Forget about the density function. Don't ask me on that stuff!

Uses: famous t-test

2.7 Cauchy distribution

Special case for t-distribution. It is when the k of χ_k^2 equal to one. In the other words, it is also the ratio of two standard normal random variables. Again, forget about the density function.

Mathematically,

$$R = \frac{Z}{\sqrt{\chi_1^2/1}} = \frac{Z_1}{Z_2}.$$

Use: an example to show mean and variance need not be finite

2.8 F distribution

Generalized case for t-distribution or ratio of two chi-square random variables. Is there anyone able to write down the density function without looking at the book? I dout it.

Mathemathically,

$$F_{m,n} = \frac{\chi_m^2/m}{\chi_n^2/n}$$

Uses: famous F test

Remark: $F_{1,n} = (t_n)^2$

3 Estimation and Hypothesis testing

3.1 Estimation: Estimator and Estimate

Why we need estimate? It is not possible to know it or it is too costly to know it.

$$\text{population} \overset{\text{Sampling}}{\Longrightarrow} \text{sample} \overset{\text{estimator}}{\Longrightarrow} \text{estimate}$$

Estimator is the rule we do the estimation. That is, the way how we guess it.
Estimate is the result of estimation. It is our guessed result.
Technically, estimator is function of sample data and estimate is value of the function.

3.2 Property of Estimator

1. Unbiasedness

   Average of our estimates from many samples equal to the true value.

   $$E\left(\hat{X}\right) = \mu$$

   Remark 1 *Asymptotically unbiased: when sample size is large, it becomes unbiased.*

   $$\lim_{n\to\infty} E\left(\hat{X}\right) = \mu$$

2. Consistency

   Estimate from very large sample equal to the true value.

   $$\lim_{n\to\infty} \Pr\left(\left|\hat{X} - \mu\right| < \varepsilon\right) = 1$$

3. Efficiency

   Variance of estimator is low. A is more efficient than B if

   $$Var\left(X\right) < Var\left(B\right)$$

3.3 Law of Large Numbers and Central Limit Theorem

Assume random variables are independent and identically distributed, if the mean and variance of random variables are finite, then we have LLN and CLT.

1. Law of large Numbers

   Average of sample mean would be close to true mean when the sample size is large.

2. Central limit theorem

   Distribution of sample mean would be normally distributed when sample size is large. The mean of distribution is of course the true mean while the variance is equal to original variance of random variable but divided by sample size. This means if the sample size become larger, our estimate would likely to be more accurate.

3.4 Hypothesis Testing

1. Null hypothesis H_0 v.s. alternative hypothesis H_1

   Null hypothesis is the initial assumption. Usually from theory or custom belief. This is what we want to test

   Alternative hypothesis is another assumption which we would like to believe (from theory or custom belief)it is true if the null hypothesis is rejected. It doesn't need to be the complement of the null.

2. Process of test

   (a) Assume the null hypothesis is correct, we can derive a few of conclusions. (run models, do the computation and get distribution of parameters)

   (b) Check whether the conclusions fit the evidence we have based on a prescribed rule. (compare whether data deviates too much from our model too much or not)

      i. If it fits the rule, we dont reject the null.

      ii. If it dont fit the rule, we reject the null.

   Remark 2 *The procedure is similar to legal process. We assume the defendant is innocent at first. Evidences are presented in the courtroom. If the proof is sufficient for judge to believe it is highly improbable for the defendant to be innocent, the defendant is convicted to be guilty of the charge.*

   Remark 3 *This kind of procedure is quite similar to mathematical proof by contradiction. First we assume the conclusion is incorrect. If we can derive a contradiction, then we have proof the theorem. In hypothesis testing, since we are dealing* *with sampling and data, there are bound to be some stochastic elements in it, so, the evidence usually cannot fully support or reject the claim. To make decision, we must have some rules to decide as judge in courtroom. For rules to work efficiently, we are bound to commit some kind of errors.*

3. Type I error, Type II error and significance level

   Type I error is the mistake we make when we have rejected the null hypothesis which is actually correct.

   $$\Pr(\text{Type I error}) = \Pr(\text{reject null} \mid H_0 \text{ is true}) = \alpha$$

   Probability of committing a type I error is also called significance level.

   Type II error is the mistake we make when we have not rejected the null hypothesis which is actually false.

   $$\Pr(\text{Type II error}) = \Pr(\text{not reject null} \mid H_0 \text{ is false}) = \beta$$

4. General testing principle

   If we want to test whether a random variable follows a particular distribution, what we can do to test it? To think it in more concrete term, you want to judge whether a coin is biased, you would like to toss it for a number of times to test whether the number of tails and number of heads is equal or not. When you find out the tails is significantly more than heads, you can draw the conclusion. However, how large the difference would be 'significant' enough to lead you to the conclusion if you have tossed 100 times? Five, ten, twenty or thirty? Will you reject the unbiasedness assumption if there are 60 heads coming out?

   For a student who has learnt elementary probability course, he might come up with the rule of calculating the probability of getting that result. He reasons that if the coing is unbiased, the number of heads X coming out is a binomial distribution with $n = 100$ and $p = 1/2$. He knows that the mean number of head is 50 so that the deviation from mean is 10. To answer the given question, he simply figures out what is the probability that the deviation is more than 10. Using normal approximation, he find out that the probability is less than 0.05. Therefore, he concludes that given this low probability, the coin should not be unbiased.

5. Normal test, t-test, chi-square test, F test....

   Now, if we change the underlying distribution of random variable of above example from binomial distribution to normal distribution. Then we will have the normal test. If we change to student's t distribution, it is T test. If we change to F distribution, it is F test. If we change to chi-square distribution, it is chi-square test.

   Well, am I telling you all those stuff that have troubled you so long are actually the same? Unfortunately, yes. All we need to do is to find out the probability under different distributions.

4 Simple Linear Regression Model

4.1 Fitting a line

We want to using x to estimate y. Mathematically, we wish to find out the functional form of this equation:

$$y = f(x)$$

To simplify the case, we assume the relationship is linear in x.

$$y = a + bx$$

That is we want to fit a line in the scatter graph of x and y. In the other words, we want to find out the best a and b.

In primary school, we all learn about direct proportionality. For instance, if apple costs two each, the relationship between cost of apple y and quantity of apples x is taking $a = 0$ and $b = 2$.

$$y = 2x$$

Of course, the world is not always perfect as the above cases, there are so many small factors which may be negligble alone but significant in sum (principle of integration!). Collectivelly, they are called luck or error. Therefore, the model would change from deterministic model to stochastic one by adding an error term.

$$y = a + bx + \varepsilon$$

The error term ε represent factors that is not capture by x.

Now, we might understand the term 'simple linear regression model'.

Simple : only one variable x is used in explaining y

Linear: the relationship is linear (no exponential, no power, no logarithm)

Regression model: with error terms in mean zero

4.2 How to fit?

What is the principle to estimate the value of a and b so that our relationship is most fitted? What do we mean by the most fit?

Since we want to estimate $y = f(x)$, we wish to minimize

$$error = |y - f(x)|$$

However, manipulation of absolute value is difficult. We try to minimize the sqaure of error.

$$error^2 = (y - f(x))^2 = (y - a - bx)^2$$

This is so-called ordinary least square(OLS) estimation method. To visualize this method, one can think that we are trying to fit a line such that the vertical distances between the point and line is minimized.

4.3 Assumptions

1. The underlying relationship between y and x is assumed to be

   $$y_t = \beta_0 + \beta_1 x_t + u_t$$

   This is what we have said in above using only x to explain y in linear fashion with error term to capture non-modelled factor. Of course, we could use more variables to explain y. Therefore, you will have learn that in the next chapter.

2. Error has zero mean.

   $$E(u_t) = 0 \text{ for all } t.$$

   We have to assume our model in the average to be correct. Otherwise, what is the use to estimate it?

3. Value of explanatory variable x_t cannot be all same. Well, without variation of x, what is used to explain the variation in y.

4. Explanatory variable x_t is given and non-random. This implies

   $$E(x_t) = x_t \text{ for all } t.$$

   Together with zero-mean error assumption, we have

   $$Cov(x_t, u_t) = 0.$$

This is to simplify the calculation. Of course, this assumption will be relaxed a little bit by allowing the explanatory variable to interact with the explained variable.

5. Error has constant variance. Its technical term is homoskedasticity.

$$Var\left(u_t\right) = \sigma^2 \text{ for all } t.$$

   This is also to simplify the calculation. Since it usually does not satisfy in economics, you will have to relax this assumption after the next chapter.

6. Errors are not serially correlated. Its technical term is serial Independence.

$$Cov\left(u_t, u_s\right) = 0 \text{ for all } t \neq s.$$

   This is also to simplify the calculation. As for data collected across time (time series data), this assumption usually fails to hold. Again, we will revisit it later.

4.4 Results

OLS means that we are going to get the value of β_0 and β_1 by minimizing the squared error:

$$\min_{\beta_0, \beta_1} \sum \left(y_t - \beta_0 - \beta_1 x_t\right)^2$$

Using simple differentiation technique, we will get

$$\begin{cases} \hat{\beta}_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

The estimator β_0 and β_1 have the following properties:

1. Unbiasedness

2. Consistency

3. Efficiency among linear estimators

4. Passing through $(\bar{x}, \bar{y})$

4.5 Goodness of fit

Given the regression, how could we interpret the relationship? Is the fitted model suitable? If x is directly proportional to y, then the regression result is no doubt a prefect fit. If both x and y are actually independent of each other, we would say the regression failed.

To measure the linear fittness of two variables, we want to find out the portion of variation of y is explained by variation of x. However, one should note that even if we don't do any regression, we could still estimate y by simply using the mean of y, $\bar{y}$. Therefore, the comparison should be set to see how much the improvement is by including x in our reasoning. One way to do it is to take the ratio of explained portion to portion to be explained.

$$R^2 = \frac{\text{explained portion}}{\text{portion wished to be explained}} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

where $\hat{y}$, y and $\bar{y}$ are respectively the y's fitted value, actual value and mean. The reason of using the square instead of absolute one is same as the reason using OLS method.

There are three technical terms about R^2. We define

$$ESS = \sum (\hat{y} - y)^2 ; RSS = \sum (\hat{y} - \bar{y})^2 ; TSS = \sum (y - \bar{y})^2$$

Therefore, we have

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

One can easily prove that

$$TSS = RSS + ESS$$

Therefore, we have

$$0 \leq R^2 \leq 1$$

where $R^2 = 1$ implies perfect fit while $R^2 = 0$ implies no linear relationship.

Graphically, RSS is sum of sqaured distance between fitted point and the mean of y, which is the portion we have explained. ESS is the sum of squared distance between the actual data point and the fitted point, which is the portion we haven't explained. TSS is the sum of squared distance between actual data point and the mean of y,which is the portion we wish to explain. Now, it is immediately clear why we define R^2 to be ratio of

RSS to TSS.

4.6 Hypothesis Testing

Although R^2 provides a good measure on the fittness of the model, it doesn't provide us the criteria to say this model is acceptable or not. Therefore, we might want to perform some hypothesis testing to ensure the model is useful before we use it in estimation and prediction.

As I have said in previous section, all we need to know before hypothesis testing is to know what is distribution of test statistic. Since we have not say anything about distribution, we cannot do any testing yet. So now let us assume the error follows normal distribution.

$$u_t \sim N\left(0, \sigma^2\right)$$

The first and foremost test is to find out whether our regression is actually more useful than using only mean of y. Therefore, the first test is to test whether $\beta_1 = 0$ or not. That is,

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

To test this hypotheis, we need to know the distribution of $\hat{\beta}_1$. Given the assumption of normality of error, if we know the value of σ^2, $\hat{\beta}_1$ follows normal distribution with zero mean (by null hypothesis) and variance equal to

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum (x - \bar{x})^2}$$

Then, we can do the normal test with the test statistic equal to

$$W = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var\left(\hat{\beta}_1\right)}}.$$

However, it is usually we don't know the value of σ^2, we need to use the sample estimate of σ^2 to estimate $Var\left(\hat{\beta}_1\right)$.

$$\widehat{Var\left(\hat{\beta}_1\right)} = \widehat{\sigma}^2 = \frac{\sum u_t^2}{T-2}$$

Then this test statistic

$$W = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var\left(\hat{\beta}_1\right)}}}$$

would follow student's t distribution with $(T-2)$ degree of freedom.

Of course, we could also test whether $\beta_0 = 0$. The procedure would be nearly the same except the variance of $\hat{\beta}_0$.

4.7 Prediction and Forecasting

Now we come to the application of the regression model. It is used to do prediction and forecasting. This is why we need to estimate the formula of getting the value of y based on the value of x. The interpretation of coefficient of intercepts and slope would be the same as the usual equation.

Of course, using the equation alone would not be fruitful enough given that we have not specify the accuracy of our estimation. Therefore, more often than not, we would like to given interval estimate rather than point estimate.

5 Multiple Regression

5.1 Fitting a plane: More explanatory variables

Simple linear regression uses one variable x_t to explain another variable y_t.

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

The underlying assumption is that other factors are included in error terms ε_t and the error term is insignificant and random $(E(u_t) = 0)$. However, this assumption may not be good enough to model situation where there are more than one explanatory factors affecting the explained variable. So, we are using more than one variable to estiamte y_t.

Mathematically, we are trying to estimate y_t using k variables $x_{1t}, x_{2t}, x_{3t}, ..., x_{kt}$.We wish to estimate the functional form of:

$$y_t = f(x_{1t}, x_{2t}, x_{3t}, ..., x_{kt})$$

Particularly, we use the function form to estiamte:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + ... + \beta_{kt} x_{kt} + u_t$$

Graphically, we are fitting a $k-$dimensional hyperplane by changing the values of $\beta_0, \beta_1, \beta_2, ...\beta_k$.

5.2 How to fit?

The estimation method is again ordinary least square (OLS), which tries to minimize sum of squared errors.

$$\begin{aligned} \min \sum error^2 &= \min \sum (y_t - f(x))^2 \\ &= \min \sum (y - \beta_0 - \beta_1 x_{1t} - \beta_2 x_{2t} - \beta_3 x_{3t} - ... - \beta_{kt} x_{kt})^2 \end{aligned}$$

5.3 Assumptions

Similar to linear regression. One important additioal assumption is no perfect multicollinearity which we would discuss at the end of this section.

5.4 Results

Using simple calculus, we are solve $k-$simultaneous equations:

$$\min_{\beta_0,\beta_1,\beta_2,...\beta_k} \sum_{t=1}^{T} \left(y_t - \beta_0 - \beta_1 x_{1t} - \beta_2 x_{2t} - \beta_3 x_{3t} - ... - \beta_{kt} x_{kt}\right)^2$$

However, it is rather tedious to solve so many equation each time. Therefore, to minimize the calculatioin procedure, let us introduce matrix.

Let $Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ y_T \end{bmatrix}$, $X = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} & .. & x_{k1} \\ 1 & x_{12} & .. & .. & .. & .. \\ 1 & .. & .. & .. & .. & .. \\ 1 & .. & .. & .. & .. & .. \\ 1 & x_{1T} & x_{2T} & .. & .. & x_{KT} \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_0 \\ \beta_2 \\ \beta_3 \\ . \\ \beta_K \end{bmatrix}$ and $u = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ . \\ u_T \end{bmatrix}$

Note that Y and X are data and β is the coefficient vector to be estimate.

Then the regression model becomes

$$Y = \beta X + u$$

The minimization of sum of squared errors becomes

$$\min_{\beta} u'u = \min_{\beta} \left(Y - \beta X\right)' \left(Y - \beta X\right)$$

Using matrix calculus, the solution would then

$$\hat{\beta} = \left(X'X\right)^{-1} X'y$$

One can show the variance of $\hat{\beta}$ would be

$$Var\left(\hat{\beta}\right) = \sigma \left(X'X\right)^{-1}$$

The estimator properties would be still unbiased, consistent and most efficient linear estimator.

5.5 Goodness of fit

It is tempting to use R^2 to measure the goodness of fit. However, it is not good as R^2 increases automatically with the number of regressor (explanatory variable). We cannot judge whether adding another explanatory variable is better fit or not by looking at R^2. We need to do some adjustement.

$$\overline{R}^2 = 1 - \frac{T-1}{T-k-1}\left(1-R^2\right)$$

The adjustment to make penalty for more variable. If k increases without changing the value of R^2, the adjusted $\overline{R}^2$ would decrease.

5.6 Hypothesis Testing

In simple regression, the first test is to do the test on whether β_1 is zero or not. This intends to figure out whether regression is useful than mean or not.

Now, in multiple regression, if we wish to do the same test, we need to test whether $\beta_1 = \beta_2 = ... = \beta_K = 0$ or not. Thus, instead of testing one parameter alone, we need to have a joint-test.

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = ... = \beta_K = 0 \\ H_1 &: \text{at least one of } \beta_1, \beta_2, ..., \beta_K \text{ not equal to zero} \end{aligned}$$

Instead of t-test, we need to use F-test.

$$F = \frac{(ESS_0 - ESS_1)\,/\,(df_0 - df_1)}{ESS_1/df_1}$$

where ESS_0 and df_0 are respectively the residual sum of square and degree of freedom under null hypothesis while ESS_1 and df_1 are respectively the residual sum of square and degree of freedom under alternative hypothesis.

The resonale behind is that if the null is true then the error under null should be very closed to the alternative hypothesis.

Of course, if we simply want to test whether a particular one is zero or not, we could still use the t-test.

5.7 Partial Regression: Bivariate and Trivariate Model

Cofficients of regression with two variables could be obtained by using four regression with one variables. In the other words, we could directly estimate this:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t$$

Or we could first obtain τ_t and ν_t from the following system:

$$\begin{aligned} x_{1t} &= \alpha_0 + \alpha_2 x_{2t} + \tau_t \\ x_{2t} &= \gamma_0 + \gamma_2 x_{1t} + \nu_t \end{aligned}$$

Then regress them on y_t :

$$\begin{aligned} y_t &= \rho_0 + \rho_2 \tau_t + \omega_t \\ y_t &= \phi_0 + \phi_2 \nu_t + \psi_t \end{aligned}$$

where

$$\beta_1 = \rho_2 \text{ and } \beta_2 = \phi_2$$

and β_0 can be obtained through solving the taking the mean of regression equation.

$$\beta_0 = \overline{y}_t - \beta_1 \overline{x}_{1t} + \beta_2 \overline{x}_{2t}$$

5.8 Model misspecification

Therer are two common errors:

1. Inclusion of irrelvant variables (omit-variable bias)

   OLS estimate is still unbiased.

2. Exclusion of pertinent variables

   OLS estimat is biased.

5.9 Multicollinearity

Multicollinearity means that correlation among explanatory variables. No doubt, we wish to have explanatory variables to be as uncorrelated as possible. If their variations are so similar, why just use of one them?

If they correlate perfectly, the estimation would fail. In partial regression, if two variables are perfectly correlated, there the residuals are zero and we cannot estimate the coefficient.

If they correlate too much, we would face the following problems:

1. Large variance of OLS estimates

2. Wider confidence Intervals

3. Insignificant t-ratio

When we know there is problem of multicollinearity?

1. examine the correlation of the raw data

2. High R^2 but low t

How to remedy multicollinearity?

1. adjust the model from economic theory

2. use first differences or take ratios

3. drop variables

4. increase sample size

5. neglect it

6 Dummy variable

6.1 Cardinal data v.s. Ordinal data v.s. Nominal data

Discrete or continuous data such as height of man, income and age could be easily fitted into regression model. Their numerical values and intervals are well-defined since they are cardinal measurement. However, for ordinal(or rank) measurement such as job rank or preference, and nominal(or categorial) measurement such as gender or race, their numerical values and the meaning of interval is not so well-defined. To cope with the difficulty of assignment of numerical value, more often than not, dummy variable is to deal with categorial explanatory variable while ordered logit and probit model is applied to deal with ordinal explained variable.

6.2 Indicator variable

We would still use number to represent nominal data since how can we build a numerical model without using numbers? The number would be assigned by using indicator variable or dummy variable.

For a categorial variable representing two values such as gender, we would assign the variable to be zero if it assumes one value and assigned to be one if it assumes another value.

In terms of formula, for example, if we want to have a sex dummy, we could use the following indicator variable I:

$$I = \begin{cases} 1 & \text{if sex=Male} \\ 0 & \text{if sex=Female} \end{cases}$$

If our nominal variable have more than two categories, then we need more categorial variable to represent the value. We cannot use the same variable again as the interval is not defined.

If we want to have seasonal dummy, we could use the indicators I_1, I_2, I_3:

	spring	summer	autumn	winter
I_1	0	1	0	0
I_2	0	0	1	0
I_3	0	0	0	1

6.3 Dummy variable trap

However, it should be noted that if we have N categories, we only need to have $N-1$ indicator variables only. Cautious readers should have observed this fact from our using only one sex dummy to represent male and female, and only three dummy to represent four seasons . If we use more than that, we would fall into dummy variable trap which is actually the problem of perfect multicollinearity.

Take sex dummy as example, if we use I_1 and I_2 to represent male and female:

	male	female
I_1	1	0
I_2	0	1

Then the two variables value are perfectly correlated with correlation coefficient of -1.

6.4 Intercept dummy

This kind of model assumes the effect of categorial data only exerts on the intercept. This assumes the effect of the categorial variable is one-off and independent of other explanatory variable.

Take the example of sex dummy,

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_t + \beta_2 I + u_t \\ I_t &= \begin{cases} 1 & \text{if sex=Male} \\ 0 & \text{if sex=Female} \end{cases} \end{aligned}$$

Then this model actually estimate this:

$$\begin{gathered} y_t = (\beta_0 + \beta_2) + \beta_1 x_t + u_t \qquad \text{if sex=Male} \\ y_t = \beta_0 + \beta_1 x_t + u_t \qquad \text{if sex=Female} \end{gathered}$$

Therefore, the gender only exerts its effect on intercept.

Hypothesis testing $H_0 : \beta_2 = 0$ could be used to test whether there is significant gender effect.

6.5 Interaction term: slope dummy

It is sometimes unrealistic to assume the categorial variable would have no effect with other variable. To model this, we would add interaction term to the system. This kind of model assumes the effect of categorial data would exerts on both slope and intercept.

Using the example of sex dummy again,

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_t + \beta_2 I + \beta_3 x_t I + u_t \\ I_t &= \begin{cases} 1 & \text{if sex=Male} \\ 0 & \text{if sex=Female} \end{cases} \end{aligned}$$

Then this model actually estimate this:

$$\begin{array}{cc} y_t = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, x_t + u_t & \text{if sex=Male} \\ y_t = \beta_0 + \beta_1 x_t + u_t & \text{if sex=Female} \end{array}$$

Hypothesis testing $H_0 : \beta_2 = 0 \;\; and \;\beta_3 = 0$ could be used to test whether there is significant gender effect at all while $H_0 : \beta_2 = 0$ and $H_0 : \beta_3 = 0$ are respectively used to test whether there is intercept gender effect and whether there is slope gender effect.

7 Heteroskedasticity

In regression model we assume the variance of error to be a constant (homeskedasticity).

$$Var\left(u_t\right) = \sigma^2 \text{ for all } t.$$

However, it has been said that this assumption may not be plausible for many ecnomomic applications. Therefore, in this section, we would examine heteroskedasticity, that is,

$$Var\left(u_t\right) \neq \sigma^2 \text{ for all } t.$$

7.1 Consequence of heteroskedasticity

1. OLS still unbiased

2. but inefficient

7.2 Solution: WLS, GLS

If we know how the formula of variance, we could correct the heteroskedasticity. For example, if the variance of error is function of another Z_t which is independent of u_t, we can run a corrected regression by first dividing all the data by Z_t. Mathematically, if

$$Var\left(u_t\right) = \sigma^2 Z_t^2$$

We can change the heteroskedasticity to homoskedasticity by dividing all y_t and x_t by Z_t and run the regression:

$$\frac{y_t}{Z_t} = \beta_0 \frac{1}{Z_t} + \beta_1 \frac{x_{1t}}{Z_t} + \beta_2 \frac{x_{2t}}{Z_t} + \ldots + \beta_k \frac{x_{kt}}{Z_t} + \frac{u_t}{Z_t}$$

Hence, the variance of error term is constant:

$$VAR\left(\frac{u_t}{Z_t}\right) = \frac{1}{Z_t^2} VAR\left(u_t\right) = \sigma^2$$

This method is called weighted least square (WLS) or generalized least square (GLS) method.

7.3 Testing

1. G-Q test (Goldefeld-Quandt test)

   We divide the data into two groups and find out the variance of errors of two groups: A and B wherer group A has n_1 observations and group B has n_2 observations. If the homoskedasticity holds, variances are should not differ significantly. Our hypothesis is based on assumption group A has lower variance than group B. Hence, the hypothesis are

   $$\begin{aligned} H_0 &: \sigma_A^2 = \sigma_B^2 \\ H_1 &: \sigma_A^2 < \sigma_B^2 \end{aligned}$$

   The testing statisitic is

   $$F_{n_1-k,n_2-k} = \frac{ESS_B/\left(n_2-k-1\right)}{ESS_A/\left(n_1-k-1\right)}$$

   Of course, if this

   $$Var\left(u_t\right) = \sigma^2 Z_t^2$$

   holds, we could sort the data according to the value of Z_t and divides the data into to two groups.

2. B-P test (Breusch-Pagan test)

   The test assumes variance of errors can be explained by a group of explanatory variables Z_t.

   $$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots + \beta_k x_{kt} + u_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 Z_{1t} + \alpha_2 Z_{2t} + \ldots + \alpha_k Z_{pt} + v_t \end{aligned}$$

   If error is homoskedasticity, we should expect $\alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$. Hence, the hypothesis to be test is

   $$\begin{aligned} H_0 &: \alpha_1 = \alpha_2 = \ldots = \alpha_p = 0 \\ H_1 &: \text{at least one of } \alpha_1, \ldots \alpha_p \text{ is not zero} \end{aligned}$$

So the procedure is:

(a) regress y_t on $x_{1t}, x_{2t}, ..., x_{kt}$ can get OLS estimates and calculate

$$\hat{\sigma}_t^2 = \sum \hat{u}_t^2$$

(b) regress $\frac{\hat{u}_t^2}{\hat{\sigma}_t^2}$ on $Z_{1t}, Z_{2t}, ..., Z_{pt}$ (auxiliary regression)

$$\frac{\hat{u}_t^2}{\hat{\sigma}_t^2} = \alpha_0^* + \alpha_1^* Z_{1t} + \alpha_2^* Z_{2t} + ... + \alpha_k^* Z_{pt} + v_t$$

(c) perfrom chi-square test with p degree of freedom, the testing statistic is $RSS/2$. That is to reject if

$$\frac{RSS}{2} > \chi_p^2(\alpha)$$

3. White test

This test is similar to B-P test except that we don't need to have the knowledge on what is the formula on variance of error. We assume the error to be function of explanatory variables. That is, our modelling on variance of error would be square or multiples of explanatory variables.

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + ... + \beta_k x_{kt} + u_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 x_{1t} + \alpha_2 x_{2t} + ... + \alpha_k x_{kt} + \alpha_{k+1} x_{1t}^2 + ... + \alpha_{2k} x_{2kt}^2 \\ &\quad + \alpha_{2k+1} x_{1t} x_{2t} + ... + \alpha_{\frac{k(k+5)}{2}} x_{(k-1)t} x_{kt} + v_t \end{aligned}$$

The hypothesis to be tested is

$$\begin{aligned} H_0 &: \alpha_1 = \alpha_2 = ... = \alpha_{\frac{k(k+5)}{2}} = 0 \\ H_1 &: \text{at least one of } \alpha_1, ...\alpha_{\frac{k(k+5)}{2}} \text{ is not zero} \end{aligned}$$

So the procedure is:

(a) regress y_t on $x_{1t}, x_{2t}, ..., x_{kt}$ can get OLS estimates and calculate

$$\hat{\sigma}_t^2 = \sum u_t^2$$

(b) regress $\frac{\hat{u}_t^2}{\hat{\sigma}_t^2}$ on any combinations of $x_1, ..., x_k$,for example, we regress $\frac{\hat{u}_t^2}{\hat{\sigma}_t^2}$ on explanatory variables $x_{1t}, x_{2t}, ..., x_{kt}$ squares of them $x_{1t}^2, ..., x_{2kt}^2$ and cross multiplication of them $x_{1t}x_{2t}, ..., x_{(k-1)t}x_{kt}$(auxiliary regression)

(c) The test is again chi square test with $\frac{k(k+5)}{2}$ degree of freedom (d.f. depends on number of regressor used in the auxiliary regression). The testing statistic is equal to unadjusted R-square R^2 times sample size T, that is, TR^2. That is to reject if

$$TR^2 > \chi^2_{\frac{k(k+5)}{2}}(\alpha)$$

8 Serial Correlation

Very often, for time series data (data collected on the same object across time), errors of different period are correlated. This kind of phenomenon is called autocorrelation. Take stock price as an example, the over-optimism or irrational exuberance would stimulate the stock price for a certain period of time but this kind of exogenous psychological factors are usually not included in the regression model. Hence, the excluded explanatory variables, implicitly hidden in error terms, are not random across time which lead to autocorrelation in errors.

8.1 Consequence

1. still unbiased

2. but inconsistent

3. and may not efficient

8.2 Estimation

1. Cochraine-Orcutt Iterative procedure (COIP)

   Suppose we know the model is

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{1t} + ... + \beta_k x_{kt} + u_t \\ u_t &= \rho u_{t-1} + \varepsilon_t, \ \ -1 < \rho < 1 \end{aligned}$$

   After we know the value of ρ, then we could take the quasi-differencing on data directly so as to remove the autocorrelation of the error term. It should be note that the sample size of data would reduce by 1 as there is nothing to difference for data at initial time. Formally, the procedure is outline as follows:

   (a) Get the value of ρ.

      i. If we already know it, we are done.

      ii. If we don't know it, we are going to estimate it.
      We will have to get the estimate $\hat{\rho}$ by run two regressions which is very similar to what we have done in last section.

A. Estimate the original regression and get the estimate $\hat{u}_t$

$$\hat{u}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{1t} - ... - \hat{\beta}_k x_{kt}$$

B. Run auxiliary regression on error and lag of error to get the estimate $\hat{\rho}_t$

$$\hat{u}_t = \rho \hat{u}_{t-1} + \varepsilon_t$$

(b) lag the regression equation by one period

$$y_{t-1} = \beta_0 + \beta_1 x_{1(t-1)} + ... + \beta_k x_{k(t-1)} + u_{t-1}$$

(c) multiple by ρ $(or\ \hat{\rho})$

$$\rho y_{t-1} = \rho\beta_0 + \rho\beta_1 x_{1(t-1)} + ... + \rho\beta_k x_{k(t-1)} + \rho u_{t-1}$$

(d) subract it from original regression equation

$$y_t - \rho y_{t-1} = \beta_0 (1-\rho) + \beta_1 \left(x_{1t} - \rho x_{1(t-1)}\right) + ... + \beta_k \left(x_{kt} - \rho x_{k(t-1)}\right) + (u_t - \rho u_{t-1})$$

(e) estimate this new equation

$$y_t^* = \alpha_0 + \alpha_1 x_{1t}^* + ... + \alpha_k x_{kt}^* + u_t^*$$

where quasi-differenced data are $y_t^* = y_t - \rho y_{t-1}$, $x_{1t}^* = x_{1t} - \rho x_{1(t-1)}, ..., x_{kt}^* = x_{kt} - \rho x_{k(t-1)}$

and new coefficients are $\alpha_0 = \beta_0 (1-\rho)\,,\ \ \alpha_1 = \beta_1, ..., \ \alpha_k = \beta_k, \ u_t^* = \varepsilon_t = u_t - \rho u_{t-1}$.

2. Hidreth-lu Search procedure

Nothing special, just search and search. This numerical method is not using regression to estimate the value of ρ but directly changing the value of ρ to minimize the ESS of the original regression equation.

8.3 Testing

1. Durbin-Watson Test (DW Test)

This is a test to find out whether there is first-order autocorrelation. One feature (or drawback) of this test is we will have inconclusive region which means we cannot draw any conclusion about the null and alternative hypothesis.

The test statistic is

$$d = \frac{\sum\limits_{t=2}^{T}(u_t - u_{t-1})^2}{\sum\limits_{t=1}^{T} u_t^2}$$

which converge to 2 if there is no autocorrelation.

There is two types of D-W test where the difference is that whether the alternative hypothesis is residuals are positively correlated or residuals are negatively correlated. The decision to select the alternative hypothesis is based on theory or preconception.

(a) Alternative hypothesis is residuals are positively correlated

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho < 0 \end{aligned}$$

After we find out d_L and d_U from table(both depends on the sample size T and the number of explanatory variable k). Our decision rule on H_0 is

$0 \le d \le d_L$	$d_L < d < d_U$	$d_U \le d \le 4 - d_U$	$4 - d_U < d < 4 - d_L$	$4 - d_L \le d \le 4$
not reject	inconclusive	reject	reject	reject

(b) Alternative hypothesis is residuals are positively correlated

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho > 0 \end{aligned}$$

$0 \le d \le d_L$	$d_L < d < d_U$	$d_U \le d \le 4 - d_U$	$4 - d_U < d < 4 - d_L$	$4 - d_L \le d \le 4$
reject	reject	reject	inconclusive	not reject

2. Lagrange Multiplier Test (LM test)

Although this test does not have inconclusive region, it requires large sample size to work.

This method is directly test the value of ρ by putting the residual equation into the original equation and test whether it is zero. Suppose our model is

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{1t} + ... + \beta_k x_{kt} + u_t \\ u_t &= \rho u_{t-1} + \varepsilon_t, \quad -1 < \rho < 1 \end{aligned}$$

That is we are going to estimate whether $\rho = 0$ in the combined regression model.

$$y_t = \beta_0 + \beta_1 x_{1t} + ... + \beta_k x_{kt} + \rho u_{t-1} + \varepsilon_t$$

Our procedure would be as follows:

(a) Do the regression of the original equation and get estimate $\hat{u}_t$

$$\hat{u}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{1t} - ... - \hat{\beta}_k x_{kt}$$

(b) Do the combined regression

$$y_t = \beta_0 + \beta_1 x_{1t} + ... + \beta_k x_{kt} + \rho \hat{u}_{t-1} + \varepsilon_t$$

(c) Compute $(T-1) R^2$ from the combined equation. Reject null if

$$(T-1) R^2 > \chi_1^2 (\alpha)$$

9 Discrete and Limited dependent variable

Dummy is forn categorial and nominal explanatory variable. Now, we want to study how to model the case when the limitation or restriction is on the explained variable. The discrete means that the value of explained variable could only assume few possible values such as number of children and limited means that the value of the explained varible could lie on within some range numbers such as probability.

9.1 Problem of using linear model

1. out of the acceptable range estimation
2. heteroskedasticity
3. inefficient estimate
4. non-normality of errors
5. problematic explanation of R^2

9.2 Probit and logit model

Since linear model would have so many problems, we have to drop this simplification. Remember why we do regression? Yes, we would to estimate the functional form f such that $y = f(x_1, x_2, ..., x_k)$. In the previous chapters, we use the most basic linear form. However, linear form places no restriction on the range of the function. (range = set of all possible value of y).

If we put the restriction on the linear function, it is no longer linear. Hence we have no choice but to adopt other functional forms which would allow the restriction on possible value of y. For example if y is probability which could only assume value from zero to one, we then restrict the family of functions to be set of functions which maps number to $[0, 1]$ interval. In particular, the most readily used functional form we have learnt is the probability function.

If the probability function used is normal distribution function, it is called probit model.

If the probability function used is logistic distribution function, it is called logit model.

9.3 Least square Estimation

In linear regression model, our estimation method is ordinary least square (OLS) or weighted least square (WLS). Least square(LS) estimation method is actually an unconstrained minimization.

$$\min_{\beta_0,..\beta_k} \sum_{t=1}^{T} \left[y_t - f\left(x_{1t}, x_{2t}, ..., x_{kt}\right)\right]^2$$

However, as I have said above, when the explanatory variable can only assume discrete or limited values, the unconstrained optimization would give us undesirable outcome. Therefore, if we continue to use least square estimate, we have to do the minimization under constrains like $0 \leq f\left(x_1, x_2, ..., x_k\right) \leq 1$ in probability estimation. That is we are going to do:

$$\begin{aligned} &\min_{\beta_0,..\beta_k} \sum_{t=1}^{T} \left[y_t - f\left(x_{1t}, x_{2t}, ..., x_{kt}\right)\right]^2 \\ &\text{s.t} \;\; 0 \leq f\left(x_1, x_2, ..., x_k\right) \leq 1 \end{aligned}$$

Remember $f\left(x_1, x_2, ..., x_k\right)$ is not longer linear and then we would not have this

$$f\left(x_1, x_2, ..., x_k\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

and we might have functional form like this (remember this is uniform distribution?)

$$f\left(x_1, x_2, ..., x_k\right) = \frac{1}{\beta_0 + \beta_1 x_1 + ... + \beta_k x_k}.$$

or even more complicated functions such as (remmeber this is normal distribution?)

$$f\left(x_1, x_2, ..., x_k\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k\right)^2}{2\sigma^2}\right]$$

The problem of using differentiation with such complicated function to obtain a close form solution would lead us into the regime of numerical optimization. (remember HKCEE bisection method?)

9.4 Maximium Likelihood Estimate

Although LS would still provide us a resonable estimation, it is usually not a good way to deal with probability function estimation. A much better way to do this kind of estimation is maximum likelihood estimate (MLE).

The principle of MLE is just opposite what we do to calculate the probability given parameter of probability density function. We are going to estimate the parameter based on "probability". Now, given the data, we already know which outcome is true. In the other words, if we travel back through time machine, we actually know which event has actually happened. Then we could maximize the parameter of the density function so that the probability function would give the greatest "probability" to the event realized.

Under the classical definition of probability, outcome of event which has realized does not fit into the criteria of random experiment. Hence, we could not call "probability" but to change to new term "likelihood" L, which we try to maximize.

9.5 Truncated Data

When we do sampling, it is quite often that some group of population could not be reached or observed easily. For example, it might be difficult to observed profit of triad from Inland revenue department data.

Particularly, if variable below certain valueis unobserved, data is called lower-truncated or truncation from below. On the other hand, if the variable above certain is unobserved, data is called upper-truncated or truncation from above.

If we know that certain data has suffer from truncation problem, we have to change the likelihood function by using conditional probability density function instead of unconditional one.

9.6 Censored Data

Besides truncation, another common data problem is censoring. It means that we know the value is greater than or smaller than some values but we do not know its exact value. In the other words, we observe the inequality rather than equality. For example, if we observe the wage of unemployed, we would find his wage would be zero. However, if there is no welfare system, he would still work for living. This means his market wage is below the welfare payment and strictly larger than zero but we could only observe zero. This means wages below welfare payment would be censored from the data. As the censoring point is from below, we call this censor from below. If the point is from above, then we calll censor from above.

Of course, similar to the truncated case, the likelihood functions would change if we know the data has been censored. However, we do not need conditional probability but we only need to put more likelihood (probability) on the censored point since it has represented a range of value below it.

10 Simultaneous equations

10.1 Interaction between explained and explanatory variable

Very often, the explaned variable is not completely independent of explanatory variables. For example, in simple demand-supply framework, price and quantity can affect each other. It would be violate our assumptions of independence of x_t $(Cov(x_t, u_t) = 0)$ if we are to estimate the following regression equation following the usual OLS procedure.

$$p_t = \beta_0 + \beta_1 q_t + u_t$$

One of the obvious problem is that the above regressio equation ignore the effect of price on the quantity. That is we assume $\alpha_0 = \alpha_1 = 0$ in the following equation.

$$q_t = \alpha_0 + \alpha_1 q_t + v_t.$$

Therefore, generally, if we want to estimate variables which would simultaneously affect value of each other, we need to estimate the both equations at the same time. For example, if we know that y would affect x and vice versa, we need to estimate the following system of simultaneous equations at one time:

$$\begin{cases} y_t = \beta_0 + \beta_1 x_t + u_t \\ x_t = \alpha_0 + \alpha_1 y_t + v_t \end{cases}$$

10.2 Structural Form v.s. Reduced Form

How can we estimate two equations at the same time? Remember how we solve simultaneous equations in alegbra course in secondary school? Yes, we use substitution method which we try to eliminate one variable by plugging in equations. Then, why not we repeat the same procedure to eliminate the simultaneous equation to single equation?

Putting second equation $x_t = \alpha_0 + \alpha_1 y_t + v_t$ into the first equation $y_t = \beta_0 + \beta_1 x_t + u_t$, we have

$$y_t = \beta_0 + \beta_1 (\alpha_0 + \alpha_1 x_t + v_t) + u_t$$

which can be further written as

$$y_t = (\beta_0 + \beta_1 \alpha_0) + \alpha_1 x_t + (v_t + u_t).$$

Then, we can handle this by our ordinary least squeare estimation method.

However, we should have known that the substitution can be done in another way. We now put first equation $y_t = \beta_0 + \beta_1 x_t + u_t$ into the second equation $x_t = \alpha_0 + \alpha_1 y_t + v_t$, we have

$$x_t = \alpha_0 + \alpha_1 (\beta_0 + \beta_1 y_t + u_t) + v_t$$

and, after rearranging of terms,

$$x_t = (\alpha_0 + \alpha_1 \beta_0) + \beta_1 y_t + (u_t + v_t).$$

Again, we use our regression technique to recover the above coefficients.

To sum up, we can estiamte the following equations separately,

$$\begin{cases} y_t = (\beta_0 + \beta_1 \alpha_0) + \alpha_1 x_t + (v_t + u_t) \\ x_t = (\alpha_0 + \alpha_1 \beta_0) + \beta_1 y_t + (u_t + v_t) \end{cases}$$

using the least square method.

We would call the original equations in structural form since it shows the structure of relationship between variables while the new equations in reduced form since it combines all equations

10.3 Indirect least-squares method

Remember our target? We can going to estimate structural form equations

$$\begin{cases} y_t = \beta_0 + \beta_1 x_t + u_t \\ x_t = \alpha_0 + \alpha_1 y_t + v_t \end{cases}$$

but our method above is going to estimate reduced form equations

$$\begin{cases} y_t = (\beta_0 + \beta_1 \alpha_0) + \alpha_1 x_t + (v_t + u_t) \\ x_t = (\alpha_0 + \alpha_1 \beta_0) + \beta_1 y_t + (u_t + v_t) \end{cases}$$

This estimation method called indirect least-squares method (ILS) as we do not using least square method to estimate the coefficient directly. OLS calculations would not tell us the value of each coefficients, instead we could only have the following

$$\begin{cases} y_t = \gamma_0 + \gamma_1 x_t + \lambda_t \\ x_t = \delta_0 + \delta_1 y_t + \omega_t \end{cases}$$

Then we have

$$\begin{cases} \gamma_0 = \beta_0 + \beta_1 \alpha_0 \\ \gamma_1 = \alpha_1 \\ \delta_0 = \alpha_0 + \alpha_1 \beta_0 \\ \delta_1 = \beta_1 \end{cases}$$

which we could recover the value of α_1 and β_1 but not β_0 nor α_0. Hence, ILS might not able to recover all you want. This kind of problem is called identification problem which refers to our inability to recover all coefficients in structural form.

10.4 Identification

How to identifiy all cofficients? It seems like our information is not enough as the number of unknowns exceeds number of equations. What information does we need? More variables. In particular, we need variables specific in each equation. Remember our problem is intercepts of structural form equations could not be recovered. If we have specific variables, we then can trace back the intercepts as they remain unchanged even when specific variables vary.

Now, we know that ILS might suffer from identification problem depending on the existence of specific variables, which is actually the result of more unknowns than equations. In theory, it is possible that number of equations may equal or more than unknowns. So, we have the following identification condition:

1. Exact-identification

   All parameters can be found and solution is unique.

2. Under-identification

   Not all parameter can be found.

3. Over-identification

   All parameters can be found but solution is not unique.

10.5 Order condition v.s. Rank condition

There are two methods to determine the system is exact-identified,under-identified or over-identified, though the underlying principle is the similar.

1. Order condition

   Remember our identification would require specific variables? Hence, the condition for an equation to be fully identified would require that it cannot have all variables and must have some variables omitted so that other equations would have specific variables. So, the necessary (not sufficient) condition is

$$K \geq G - 1$$

   where K is the number of excluded variables and G is number of equations.

2. Rank condition

   Although this condition is much more complicated, it still worth studying it as it is necessary and sufficient condition for identification.

   For matrix A, Rank(A) is the number of of independent rows or number of independent columns. In fact, they are always the same. To check the identification condition, we need to first write the system in matrix form. That is, if the structural form is

$$\begin{cases} x_t = \alpha_0 + \alpha_1 y_t + \alpha_2 a_t + v_t \\ y_t = \beta_0 + \beta_1 x_t + \beta_2 a_t + \beta_3 b_t + u_t \\ z_t = \gamma_0 + \gamma_1 y_t + \gamma_2 b_t + \gamma_3 c_t + \omega_t \end{cases}$$

   The matrix form would then be

$$\begin{pmatrix} 1 & -\alpha_1 & 0 \\ -\beta_1 & 1 & 0 \\ 0 & -\gamma_1 & 1 \end{pmatrix} \begin{pmatrix} x_t \\ y_t \\ z_t \end{pmatrix} = \begin{pmatrix} \alpha_2 & 0 & 0 \\ \beta_2 & \beta_3 & 0 \\ 0 & \gamma_2 & \gamma_3 \end{pmatrix} \begin{pmatrix} a_t \\ b_t \\ c_t \end{pmatrix} + \begin{pmatrix} v_t \\ u_t \\ \omega_t \end{pmatrix}$$

   We have to check rank of the combined cofficients matrix:

$$\begin{pmatrix} 1 & -\alpha_1 & 0 & \alpha_2 & 0 & 0 \\ -\beta_1 & 1 & 0 & \beta_2 & \beta_3 & 0 \\ 0 & -\gamma_1 & 1 & 0 & \gamma_2 & \gamma_3 \end{pmatrix}$$

   To check whether the first equation if identified or not, we need to:

(a) look at the first row and locate the columns of zero.

$$\begin{pmatrix} 1 & -\alpha_1 & 0 & \alpha_2 & 0 & 0 \end{pmatrix}$$

which has zero at 3rd, 5th and 6th columns

(b) take out those columns

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & \beta_3 & 0 \\ 1 & \gamma_2 & \gamma_3 \end{pmatrix}$$

(c) remove the first row

$$\begin{pmatrix} 0 & \beta_3 & 0 \\ 1 & \gamma_2 & \gamma_3 \end{pmatrix}$$

(d) calculate the rank, which is equal to 2 in the above case.

(e) If rank $= G - 1$, then the equation is identified, otherwise not.

In the above case, $G - 1 = 3 - 1 = 2$ and so it is identified.

Now, if you want to check second or third equation, just follow the same procedure but change the target row in step a and step c.

10.6 Two-stage least-squares estimation

What can we do if the system is under-identified or over-identified? Fortunately (or might be unfortunate for you), we have another way to do estimation besides ILS.

Remember our violation of OLS is $Cov\left(x_t, u_t\right) = 0$. So, we could still use OLS if we substitute x_t by another proxy variables. In two-stage least-squares estimation, we use the reduced form predicted $\hat{x}_t$ to replace x_t. Then we perform OLS on the structural equations directly.

Formally, if we wish to estimate this system

$$\begin{cases} y_t = \beta_0 + \beta_1 x_t + u_t \\ x_t = \alpha_0 + \alpha_1 y_t + v_t \end{cases}$$

we could first estimate this reduced equation

$$x_t = \delta_0 + \delta_1 y_t + \omega_t$$

where $\delta_0 = \alpha_0 + \alpha_1\beta_0$, $\delta_1 = \beta_1$ and $\omega_t = u_t + v_t$.

Then, we could obtain $\hat{x}_t$ by

$$\hat{x}_t = \hat{\delta}_0 + \hat{\delta}_1 y_t + \omega_t$$

Since we know our estimation of coefficients would have not changed if order of variables are reversed (as long as those assumptions are still satisfied), we can estimate this system

$$\begin{cases} y_t = \beta_0 + \beta_1 x_t + u_t \\ y_t = -\frac{\alpha_0}{\alpha_1} + \frac{1}{\alpha_1} x_t + \omega_t \end{cases}$$

and do the OLS estimation by replacing x_t by $\hat{x}_t$.

$$\begin{cases} y_t = \beta_0 + \beta_1 \hat{x}_t + u_t \\ y_t = -\frac{\alpha_0}{\alpha_1} + \frac{1}{\alpha_1} \hat{x}_t + \omega_t \end{cases}$$

Then we could recover all coefficients $\alpha_0, \alpha_1, \beta_0$ and β_1.

Congratulation!

You have reached the last page!

Thank you for watching!

`Lazy Production @ 2006`