EE425 Lecture Notes

By Wanwiphang Manachotphong[1]

In this course, we will study applications of statistical and economic theories in analyzing economic data. This includes parameter estimation using the ordinary least squares (OLS) technique and hypothesis testing. The course covers single and multiple regressions. It also discusses problems encountered by OLS, including autocorrelation, heteroskedasticity, multi-collinearity, specification error, and stochastic regressors. Other estimation techniques such as generalized least squares (GLS), maximum likelihood estimation (MLE) and the use of computer software will also be covered.

[1] Email: wanwiphang@econ.tu.ac.th

BACHELOR OF ECONOMICS
THAMMASAT UNIVERSITY

Course Outline

Course ID: EE425 ECONOMETRICS 1

Semester 1/2021 (August 9 – November 27, 2021)

Number of Credit: 3 Credits

Prerequisite: MA216 (or MA211) and ST217 (or ST212)

Course Description: In this course, we will study applications of statistical and economic theories in analyzing economic data. This includes parameter estimation using the ordinary least squares (OLS) technique and hypothesis testing. The course covers single and multiple regressions. It also discusses problems encountered by OLS, including autocorrelation, heteroskedasticity, multi-collinearity and specification error. Other estimation techniques such as generalized least squares (GLS), maximum likelihood estimation (MLE) and the use of computer application in practice will also be covered.

Course Objectives: This course is the more advanced alternative to EE325 (Introductory Econometrics). It will focus on the theory that underpins standard econometric methods. Students will learn to derive simple theoretical results from standard principles. They will also learn to apply the knowledge to analyze actual datasets using a standard software such as STATA.

Class Time and Logistic

Class day: **Tuesdays and Thursdays**

Class time: **11:00 AM – 12:30 PM**

Teaching Materials Platform: BE Moodle

Meeting Platform: [Zoom – meeting ID: 982 606 9955, passcode: 425481]

Instructor:

Name: Asst.Prof.Dr.Wanwiphang Manachotphong

Office Hours: By appointment

Email: wanwiphang@econ.tu.ac.th

Phone: 081-021-5664

Main Text: Wooldridge, J. M. ***Introductory Econometrics: A Modern Approach***. 3rd ed. Thompson: South-Western, 2006.

Recommended Texts & Materials: Gujarati, D.N., and D.C. Porter***, Basic Econometrics***. 5th ed., N.Y., McGraw-Hill, 2009.

Suggested Readings: Gujarati, D.N., and D.C. Porter***, Basic Econometrics***. 5th ed., N.Y., McGraw-Hill, 2009.

Grading Criteria:

Homework	20%	
Paper Replication	15%	**(Due on December 15, 2014 right before the Final Exam)**
Midterm Exam	30%	**(October 3, 2021, 13.00 PM – 15:00 PM)**
Final Exam	35%	**(December 13, 2021, 9.00 AM – 11.00 AM)**

***Late homeworks count as 50% of your actual marks.**

*If there is any handout or additional reading, it will be posted on Moodle prior to class. Students are responsible to review the topic ahead of the class for more effective learning.

Contents

Introduction and Motivation

1 What is econometrics?

Statistical method for estimating economic relationships, testing economic theories, and evaluating an implementing government and business policies (Wooldridge, 2009)

- Estimating economic relationships – How much discount should a hotel give in order to achieve full occupancy? (Price vs. Demand), How many cars would be sold in the car tax is reduced by half? (Tax vs. Consumption), etc.
- Testing economic theories – For example, is it true that the demand curve is always downward sloping? Is it true that firms always maximize profits? (See Levitt, 2006)[1] etc.
- Evaluating and implementing government policies – Does the Universal Health Coverage (UHC) program help decrease mortality rate? Which method is more effective in convincing rural students to come to school, free lunch or subsidy to the parents?
- Evaluating and implementing business policies – Should firm pay the manager a fixed salary or a variable compensation in order to achieve the highest profit? Are part-time workers more productive than full-time workers? etc.

---

[1] Levitt, S. D. (2006). "An Economist Sells Bagels: A Case Study in Profit Maximization." National Bureau of Economic Research (NBER) Working Paper 12152.

- In econometrics, we believe that there are actually "true" answer(s) to the above questions.
- Econometricians collect data (sample data) and use the sample data to answer the questions.
- Econometric methods help justify that the answers that we get from analyzing the sample is comparable to the "true" answers.

1.1 Steps in Empirical Economics Analysis

- – ** Empirical -> data and numbers.

1. Form a question, define the population of interest.
2. In some cases, construct an economic model.

$$y = f(x_1, x_2, x_3, x_4)$$

where y denotes the number of days/week that a student (in rural Thailand) would go to school, x_1 denotes the availability of school lunch, x_2 denotes the provision of subsidies to parents, x_3 denotes parents' income, x_4 denotes parents' occupation.

3. If the analysis is based on a real economic model, one has to adapt it in such a way they can perform econometrics analysis.
4. In many cases, one bypasses 2. and 3., and construct an econometric model right away
5. Let the "Number of days a student goes to school" $= y$

$$y = \beta_0 + \beta_1 \text{lunch} + \beta_2 \text{subsidies} + \beta_3 \text{parents\_inc.} + \beta_4 \, \text{parents\_occ.}$$

---

2 Types of economic data

- Data is usually a "subset" of the population of interest. For example, some students in the village of interest.
- Cross-sectional data – More than 1 individual, households, cities, villages. But 1 time period.
- Let the "Number of days a student goes to school" = y

Student no.	y	lunch	subsidies	parents_inc.	parents_occ.
1	5	1	1	3,000	farmer
2	2	0	1	4,500	hair-dresser
3	3	0	0	6,000	farmer
4	4	0	1	3,500	driver

- Time-series data – 1 individual, households, cities, villages. More than 1 time period.

Student no.	Time	y	lunch	subsidies	parents_inc.	parents_occ.
1	1/02/10	5	1	1	3,000	farmer
1	2/02/10	5	0	1	3,000	farmer
1	3/02/10	3	0	0	3,000	farmer
1	4/02/10	2	0	0	3,000	farmer

- Panel or Longitudinal data – Several individual, households, cities, villages. More than 1 time period.

Student no.	Time	y	lunch	subsidies	parents_inc.	parents_occ.
1	1/02/10	5	1	1	3,000	farmer
1	2/02/10	5	0	1	3,000	farmer
2	1/02/10	2	0	0	4,500	hair-dresser
2	2/02/10	4	0	1	4,500	hair-dresser
...	...	...	...	...	...	...
4	2/02/10	3	0	1	3,500	driver

Review of Some Statistical Concepts

- Suggested readings – Wooldridge, Appendix B and C **OR** Gujarati, Appendix A, pp.869-912

1 Random variables and distributions

- Suppose Nick has been notified that a person he met last week just tested positive for Covid-19. They met at an open-air space and wore a mask at all time. What is Nick's chance of getting Covid-19?
- Here, each incident (getting Covid-19) is a *random variable*. This random variable can have two outcomes – 1) getting Covid-19 and 2) not getting Covid-19.
- The summary of the probability that each outcome could happen is called the *probability distribution*.

1.1 Types of Random Variables

1. Discrete Random Variables – outcomes are only countable numbers, i.e. 0,1,2,3....

   - Bernoulli Random Variable is the one that takes only two values, coin flipping, passengers show up vs. not show up, etc.

   - For example:

$$\begin{aligned} X &= \begin{cases} 1 \text{ if getting Covid-19} \\ 0 \text{ if not getting Covid-19} \end{cases} \\ P(X &= 1) = \theta \\ P(X &= 0) = 1 - \theta. \end{aligned}$$

   where $P(X = 1) = \theta$ is the probability that Nick will get Covid-19. Since there are only two possible outcomes, the probability that Nick will not get Covid-19 is $(1 - \theta)$.

2. Continuous Random Variables – outcomes can be any possible value.

   - For example:

$$\begin{aligned} X &= \text{temperature tomorrow } (c^\circ) \\ P(30 &< X < 35.12) = 1 - P(X \leq 30) - P(X \geq 35.12). \end{aligned}$$

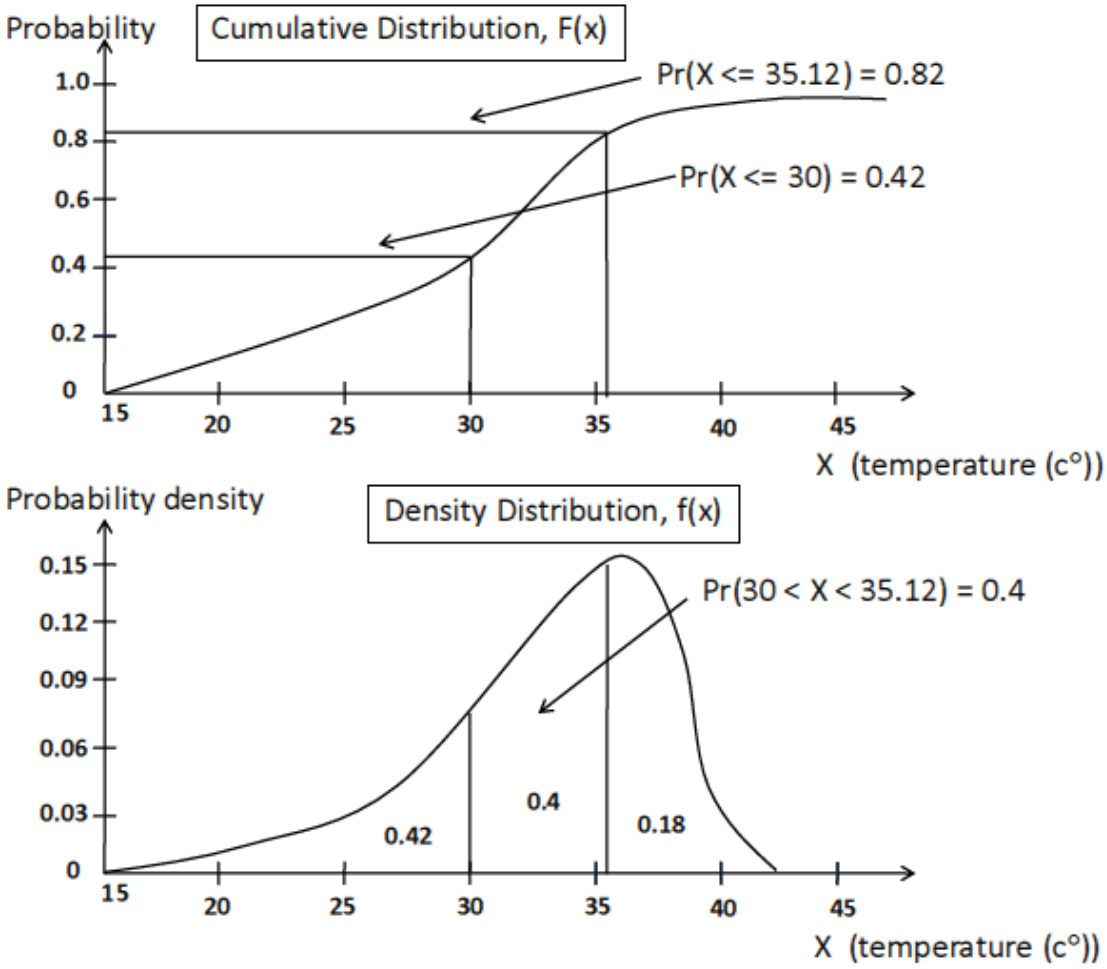


FIGURE 1. Cumulative and Density Distribution Functions.

1.2 Distribution of a Random Variable (X)

- Probability mass function (pmf) of x denoted by $f(x)$ = the summary of the probability that each outcome of x could happen (for discrete variables)
- Probability density function (pdf) of x denoted by $f(x)$ = the summary of the probability that each outcome of x could happen (for continuous variables)
- Cumulative distribution function of (cdf) x denoted by $F(x)$ = the summation of pdf over all values x_j such that $x_j \leq x$.

1.3 Examples of Some Distributions

- Normal Distribution $N(\mu, \sigma^2)$

  - μ = mean ... or the expected value of the random variable X when we draw X repeatedly for many many times (like 1,000 time).

  - σ^2 = variance ... or how far the random variable X is from its mean μ on average.

  - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

  - $F(x) = \int_{-\infty}^{x} f(x)$

- Bernoulli Distribution

- $f(x) = \begin{cases} \theta & \text{if X} = 1 \\ 1-\theta & \text{if X} = 0 \\ 0 & \text{otherwise} \end{cases}$

- - $F(x) =?$

  - $mean(x) =?$

  - $variance(x) =?$

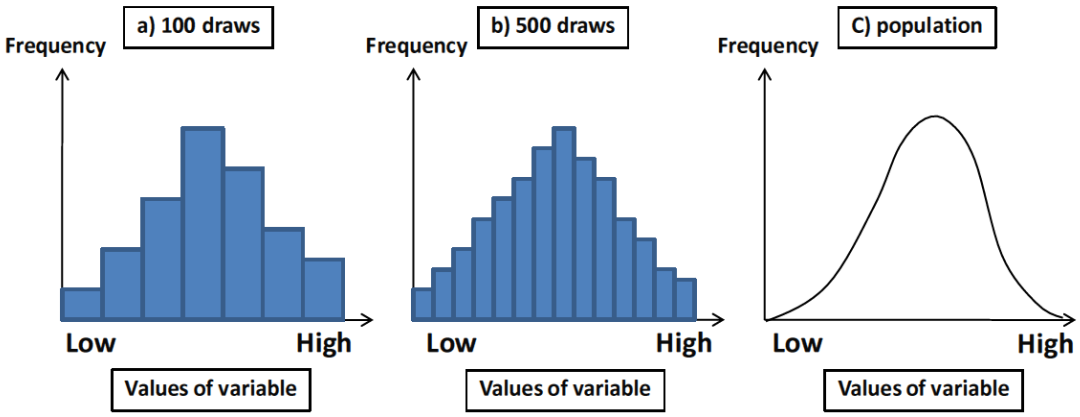


FIGURE 2. Histograms for a Continuous Variable

1.4 Population Distribution vs. Sample Distribution

- In statistics and econometrics, "population" refers to the entire pool of samples we are interested in.
- "Sample" refers to a subset of the population.
- In most cases, the population tends to be too large that we cannot collect data from every sample. Therefore, we collect "samples" of the population instead.
- Thus, we need to be able to distinguish between

  "population distribution" and "sample distribution"

  "population mean" and "sample mean"

  "population variance" and "sample variance"

  etc.
- Sample's statistics are used to infer population's statistics. Can we do this?

2 Joint Distributions, Conditional Distributions and Independence

2.1 Joint Distributions

- It sometimes rains August and I sometimes forget to bring my umbrella. Let X be a discrete random variable which takes the value of 1 when it rains, 0 otherwise. Let Y be a discrete random variable which takes the value of 1 when I bring my umbrella. What is the probability that it rains and I happen to bring my umbrella?
- Variables X and Y have a joint distribution.
- $f_{X,Y}(x, y) = P(X = x, Y = y)$ or $P(X, Y)$ is the joint density function of (X, Y).
- Suppose the probability that it rains $(X = 1)$ on a given day in August is 0.4, and the probability that I bring my umbrella $(Y = 1)$ is 0.7. What is the probability that I bring my umbrella on a rainny day?

  Answer: It depends...

2.2 Conditional Distribution and Marginal Distribution

- Conditional probability $P(Y|X)$, Marginal probability $P(Y)$
- Usually, in economics, we are interested in variables that are *not* independent from each other. Thus, the independence assumption does not usually hold.

  For example

  - the probability that I bring my umbrella may depend on what the weather forecast says.

  - the probability that a basketball player could score may depend on whether he could score in the previous attempt.

  - the probability that a student can pass a university entrace examination could depend on which high school he/she goes to, his/her parents' education level, his/her effort, etc.
- Conditional distribution/probability is the distribution/probability of a random variable given the outcome of another (other) random variable(s).

  - What is the probability that I bring my umbrella given that it rains?

  - What is the probability that a basketball player could score given that he could not score in the previous attempt?

  - What is the probability that a student can pass university entrance exam given her high school, her parent education level and her effort?

- Marginal distribution (probability) is the distribution (probability) of Y regardless of the value of X.

  - Like, the probability that it rains

  - The probability that I bring my umbrella, etc.

$$P(Y = y).$$

- The relationship between conditional, joint and marginal probability.

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

or in the continuous context, we can write

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

2.3 Independence

- If X and Y are independent, then knowing the outcome of X would not change the probabilities of the possible outcomes of Y.
- Like.. if "rain" and "my decision to bring an umbrella" are independent, then, knowing that it's going to rain would not change my probability to bring an umbrella.
- Thus, **if** X and Y are independent:

$$\begin{aligned} P(X &= 1, Y = 1) = P(X = 1) \times P(Y = 1) = 0.4 \times 0.7 = 0.28 \\ \text{or} \quad P(X &= x, Y = y) = P(X = x) \times P(Y = y) \end{aligned}$$

- If X and Y are **not** independent, then:

$$\begin{aligned} P(X &= 1, Y = 1) \neq P(X = 1) \times P(Y = 1) \\ \text{or} \quad P(X &= x, Y = y) \neq P(X = x) \times P(Y = y). \end{aligned}$$

---

3 Expectation, variance, covariance and correlation

3.1 Expected value of a Random Variable – $E(X)$ or μ_X

- An expectation is a measure of central tendency.
- The expected value of a random variable X, denoted $E(X)$ or μ_X, is the average value of the random variable over many repeated draws.

  - Discrete case: $E(X) = \Sigma_{i=1}^{N} x_i P(X = x_i)$

  - Continuous case: $E(X) = \int_{-\infty}^{\infty} x f(x) dx$

  Example: What grade would student A get from EE425?

x_i	$P(X = x_i)$	$x_i P(X = x_i)$
$0(F)$	0.05	0
$1(D)$	0.05	0.05
$2(C)$	0.25	0.50
$3(B)$	0.40	1.20
$4(A)$	0.25	1.00
Total $(\Sigma_{i=1}^{N})$	1.00	$2.75(\approx B-)$

3.2 Properties of Expected Values

1. For any constant c, $E(c) = c$
2. For any constant a and b, $E(aX + b) = aE(X) + b$
3. If $\{a_1, a_2, ..., a_n\}$ are constants and $\{X_1, X_2, ..., X_n\}$ are random variables, then

$$E(a_1X_1 + a_2X_2 + ... + a_nX_n).$$

Or using summation notation

$$E\sum_{i=1}^{n} a_iX_i = \sum_{i=1}^{n} a_iE(X_i).$$

** It is important to be reminded that $E(X^2) \neq E(X)E(X)$!

3.3 Conditional and Marginal Expectations

- Example: What grade would student A get from EE425 given the number of hours/week she spent on studying EE425.

$P(X = x_i, H = h_i)$					
$x_i \backslash h_i$	0	3	6	9	12
$0(F)$	0.0	0.0	0.0	0.0	0.0
$1(D)$	0.1	0.1	0.0	0.0	0.0
$2(C)$	0.0	0.1	0.1	0.0	0.0
$3(B)$	0.0	0.05	0.15	0.1	0.05
$4(A)$	0.0	0.0	0.05	0.1	0.1

where x_i indicates grade, h_i indicates number of hours/week spent on studying EE425.

- What are the expectations of grade (x_i) and hours (h_i)?
- Are (x_i) and (h_i) independent? why or why not?
- What is the conditional expectation of x_i given $h_i = 9$?

3.4 Variance and Standard Deviation of a Random Variable

- The variance and standard deviation measure the "variability", the "dispersion" or the "spread" of a probability distribution

$$\begin{aligned} Var(X) \text{ or } \sigma^2 &= E[(X-\mu_x)^2] \\ &= \\ &= \\ &= \\ &= \\ &= E(X^2) - \mu^2. \end{aligned}$$

- The standard deviation, denoted $sd(X)$, is the positive square root of the variance:

$$sd(X) \text{ or } \sigma = +\sqrt{Var(X)}.$$

Exercise:

- What would happen if variance = 0?

-Depict probability distributions with different values of variance.

3.5 Properties of Variances

1. $Var(X) = 0$ if, and only if, there is a constant c, such that $P(X = c) = 1$, in which case, $E(X) = c$.
2. For any constand a and b, $Var(aX + b) = a^2 Var(X)$

3.6 Properties of Standard Deviations

1. For any constant c, $sd(c) = 0$
2. For any constants a and b,

$$sd(aX + b) = |a| sd(X).$$

In particular, if $a > 0$, then $sd(aX) = a \times sd(X)$.

Exercise: What is $Var(aX + b) =?$, $Var(aX + bY) =?$

- Well, how do we compare and assess one random variable against another? One way to do this is through standardization.
- You may have heard of the Z-score

$$Z = \frac{X - \mu_X}{\sigma}.$$

- If X is normally distributed, then Z would be normally distributed. We can use the well-known Z statistic table for hypothesis-testing.
- In many cases, where the sample size gets very large, $Z = \frac{X-\mu_X}{\sigma}$ is approximately normally distributed regardless of the orginal distribution of X.

  Exercise:

  - What is $E(Z)$ and $Var(Z)$?

3.7 Covariance and Correlation of Two Variables

- We talked about joint distribution and independence earlier. How do these two concepts relate to covariance and correlation?
- When the movement of one variable can give some information about the movement of another variable, these two variables are dependent.
- We can also say that they are "correlated" or their "covariance" is not zero.
- Covariance and correlation measure the amount of *linear* association between variables.
- **It is worth noting that zero correlation does not imply independence.

$$\begin{aligned} \text{covariance or } Cov(X,Y) \text{ or } \sigma_{X,Y} &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= \\ &= \\ &= \\ &= E(XY) - \mu_X\mu_Y. \end{aligned}$$

Exercise: What's the unit of $Cov(X,Y)$?

- Correlation – makes the unit of dependence more standardized.

$$Corr(X,Y) \text{ or } \rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}.$$

- $\rho_{X,Y} = 1$; perfect positive linear relationship
- $\rho_{X,Y} = -1$; perfect negative linear relationship
- $\rho_{X,Y} = 0$; no linear relationship.

---

4 Estimators and desirable properties of estimators

- In statistics and econometrics, we hardly have the complete information of the population.
- Most data that we deal with are from a subset of the population or a "sample".
- We would like to learn about the "population" as we can using the "sample" that we have.

  - It is important to identify the population of interest

  - Once the population is identified, we can specify the model for the population relationship of interest.

- Examples of Estimators...

Population Parameter	Estimator(s)
population mean (μ)	$\bar{X} = \frac{1}{N}\Sigma_{i=1}^{N} X_i$
population variance (σ^2)	$S^2 = \frac{1}{N-1}\Sigma_{i=1}^{N}(X_i - \bar{X})^2$
$\alpha, \beta_1, \beta_2, \beta_N$	$\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_N$

- Miguel and Kremer(2004) "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities."

$$\text{school attendance rate} = \alpha + \beta_1 \times \text{deworming} + \beta_2 x_2 + ... + \beta_N x_N$$

How do we know the true value of $\alpha, \beta_1, \beta_2, \beta_N$? Without having the population and the correct model, it may be impossible to know the true value of $\alpha, \beta_1, \beta_2, \beta_N$. But the econometricians try their best to come up with the estimators, often denoted $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_N$. The better the estimator, the more it satisfies the desirable properties.

4.1 Desirable properties of estimators

- From now on, let's denote the population parameter of interest "θ" and its estimator "W". The desirable properties of estimators are: unbiasedness, efficiency and consistency.

1. Unbiasedness – the expected value of the estimator is equal to the value of the parameter it tries to estimate.

- $E(W) = \theta$
- $Bias(W) \equiv E(W) - \theta$
- Exercise: is $\bar{X}$ a biased estimator? What about X_1 ?

---

2. Efficiency – an estimator with a lower variance is said to be "more efficient" than another estimator with a higher variance

- If $Var(W_1) \leq Var(W_2)$, then W_1 is a more efficient estimator of θ than W_2.
- Exercise: Which estimator is more efficient, $\bar{X}$ or X_1?

---

3. Consistency – when the sample size gets large, the estimater W can do better and better in estimating θ.

- Large sample properties can also be called "asymptotic" properties.

- For W_n, which is an estimator of θ based on a sample $X_1, X_2, ..., X_n$ of sample size n. Then, W_n is a consistent estimator of θ, if for every $\varepsilon > 0$,

$$P(|W_n - \theta| > \varepsilon) \to 0 \text{ as } n \to \infty.$$

if not, then we can say that W_n is inconsistent.

4.2 The concept of Mean Squared Error (MSE)

- is used to compare different biased estimators.

$$MSE(W) = E[(W - \theta)^2]$$

Exercise: show that $MSE(W) = Var(W) + [Bias(W)]^2$.

The Simple Regression Model

- Some questions

  - How often (Y) would a person like to go eat at the Fuji restaurant if the price (X) drops?

  - How many more cars (Y) would be on the street in Bangkok if the government waives car tax (X)?

  - How much would my income (Y) increase if I obtain a master's degree (X)?

- The concept

  - How would you measure the "change in the number of new cars" when the "tax rate" changes.

  - What would be the "change in the number of new cars" when the "tax rate" is reduced by 1 percent?

$$\beta_1 = \frac{\Delta \text{ new cars}}{\Delta \text{ tax rate}}$$

- Or, using calculus, you can get

$$\frac{dY_i}{dX_i} = \beta_1.$$

- ** The linear relationship between Y and X is assumed.

1 Principle, assumptions and derivation of ordinary least squares (OLS) estimators

1.1 Terminology for the Linear Regression

- The simple regression model can be defined as

TABLE 2.1

Terminology for Simple Regression

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

- Y_i and X_i are variables
- β_0 and β_1 are parameters

1.2 Derivation of Ordinary Least Squares (OLS) Estimators

- If we do not have the data point of the entire population, but rather on a subset of samples, what should we do to derive β_0 and β_1?

  Example:

  How often would a TU undergraduate student go eat at Fuji restaurant (Y) if he/she receive an (X) percentage discount in price?

TABLE 3.1. Number if visits to Fuji Restaurant per year and percentage discount

i	Y_i = visit to Fuji restaurant (times/year)	X_i = percentage discount in price
1	95	78
2	42	55
3	56	67
4	83	70
....	...	...
100	32	46

Source: data collected from a random survey of 100 TU undergraduate students

- We surveyed 100 TU undergraduate students (sample)
- We hope that we can use this data to explain the frequency of TU undergraduate students' (population) visit to Fuji restaurant given the price discount.

- The relation of the frequency of TU undergraduate students' (population) visit to Fuji restaurant (Y) given the price discount (X) can be expressed as:

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

- But since we don't have the data of the entire population and we don't know the true value of β_0 and β_1, we need to find estimators of β_0 and β_1.
- The estimators are often called $\hat{\beta}_0$ and $\hat{\beta}_1$. (like $\bar{X}$ is an estimator of μ and S^2 is an estimator of σ^2).

  How do we estimate $\hat{\beta}_0$ and $\hat{\beta}_1$?
- The OLS suggests that we can find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the sum of squared errors (deviation from the regression line).

- Mathematical Derivation of OLS (minimization the sum of squared errors)

$$\underset{\hat{\beta}_0,\hat{\beta}_1}{\arg\min} \sum_{i=1}^{n} (\hat{u}_i^2) = \underset{\hat{\beta}_0,\hat{\beta}_1}{\arg\min} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

First Order Condition (F.O.C):

$$w.r.t.\hat{\beta}_0 \quad \Rightarrow \quad 0 = -2(\Sigma_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)) \tag{3.1}$$

$$w.r.t.\hat{\beta}_1 \quad \Rightarrow \quad 0 = -2(\Sigma_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)) \tag{3.2}$$

Divide 3.1 by -2, we get

2 Properties of OLS estimators

2.1 Algebraic Properties

- Thus far, the OLS estimators are

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 &= \frac{\Sigma_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\Sigma_{i=1}^n (X_i - \bar{X})^2}\end{aligned}$$

  this gives

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i \\ \hat{u}_i &= Y_i - \hat{Y}_i.\end{aligned}$$

  This implies the following algebriac properties

1. $\Sigma_{i=1}^n \hat{u}_i = 0$ – the calculation of OLS $\hat{\beta}_0$ and $\hat{\beta}_1$ is done such that it minimizes $\Sigma_{i=1}^n \hat{u}_i^2$. This is true when $\Sigma_{i=1}^n \hat{u}_i = 0$
2. $\Sigma_{i=1}^n X_i \hat{u}_i = 0$ – we get this from the first order condition deriving $\hat{\beta}_1$. This implies that $Cov(X_i, \hat{u}_i) = 0$. (Why?)
3. The point $\bar{Y}$ and $\bar{X}$ are always on the regression line – we know this from $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$.

2.2 Properties proving BLUE

- The OLS estimator is the Best Linear Unbiased Estimator (BLUE). How?
- The OLS' $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1 respectively

$$\begin{aligned} E(\hat{\beta}_0) &= \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1. \\ \text{or } E(\hat{\beta}_{OLS}) &= \beta. \end{aligned}$$

- The OLS' $\hat{\beta}_0$ and $\hat{\beta}_1$ are the most efficient among all the linear estimators

$$Var(\tilde{\beta}_{non-OLS}|\ X) - Var(\hat{\beta}_{OLS}|\ X) \geq 0.$$

2.3 Assumptions on the simple linear regression (SLR) model

- Some assumptions (or certain conditions) are required for the OLS to be BLUE. This set of assumptions are often called the "Gauss-Markov Assuptions for Simple Regression"

  Assumption SLR1. Linear in Parameter – Y is linear in X.

  Assumption SLR2. Random Sampling – We have a random sample size n, $\{(x_i, y_i) : i = 1, ..., n\}$, This, the model _ _ _ _ _ _ _ becomes:

  Assumption SLR3. Sample Variation in the Explanatory Variable – The sample outcomes are not all the same value

  Assumption SLR4. Zero Conditional Mean –

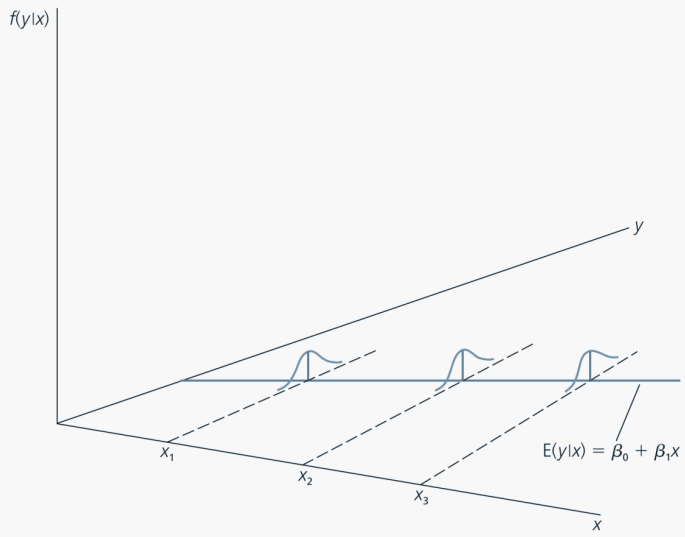
  Assumption SLR5. Homoskedasticity –

- In other words, The OLS estimator of β in the linear model when u_i is $i.i.d.(0, \sigma^2)$ is the best (minimum variance) estimator within the class of linear unbiased estimator.
- The conditional concept, which implies that X_i is predetermined (being conditional upon, or fixed) is very crucial for the OLS estimators to be unbiased.
- We must not forget that there is no reason that all these assumptions should be true!

2.4 *Homoskedasticity VS. Heteroskedasticity*

FIGURE 2.8

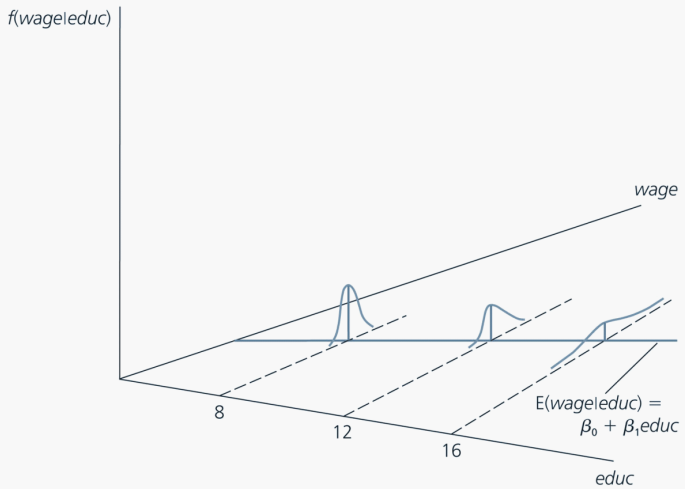
The simple regression model under homoskedasticity.



Homoskedasticity

FIGURE 2.9

Var(*wage*|*educ*) increasing with *educ*.



Heteroskedasticity

2.5 Expectation of Estimators

- Proof for $E(\hat{\beta}_1) = \beta_1$: We need to use assumption SLR 1,2,3,4.

  From

$$\hat{\beta}_1 = \frac{\Sigma_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\Sigma_{i=1}^{n}(X_i - \bar{X})^2} \tag{3.3}$$

  For calculation tractability, let

2.6 Variance of OLS Estimators

- From eq._ _ _ _ _ _ _ _ _ _ _, we can write

$$\begin{aligned} \hat{\beta}_1 &= \beta_1 + (\Sigma_{i=1}^n u_i k_i) \\ Var(\hat{\beta}_1) &= Var(\beta_1) + Var(\Sigma_{i=1}^n u_i k_i) \end{aligned}$$

  Here, β_1 (the true β_1) is a constant. And since we are conditioning on X_i, the values of k_i are also non-random. (This is not to be confused with the random sampling of X_i, $i = 1, ..., n$.) When you are conditioning on some variables, you take those variables as non-random (or as given). In this case, we can write

The prove that $Var(\hat{\beta}_1) = \frac{\sigma^2}{SST}$ is the minimum under the class of "linear unbiased estimator" is complicated without relying on some metrix simplifications. Therefore, we will do this proof using the metrix notation.

2.7 Some Concepts to be exphasized (Population vs. Sample; PRF vs. SRF; error vs. residual)

- Population – is the "truth" and in econometrics, we believe that there is one set of "truth". Therefore, the Population Regression Function (PRF) is fixed. In almost all cases, we do not know "exactly" what PRF is.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{3.4}$$

  where $i = 1, 2, ..., \Psi$. Ψ represents the total number of the "population" we are interested in.

- Sample – is a subset of "truth", a subset of "population". We run, or construct, the Sample Regression Function (SRF) to estimate the PRF.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

  where $i = 1, 2, ..., n$, $n < \Psi$.

- The true "error term" u_i in eq.3.4 is **unobserved** because we never know what β_0 and β_1 is.

- We can, however, observe the "residual" or "predicted residual" $\hat{u}_i$ from the following calculation

$$Y_i - \hat{Y}_i = \hat{u}_i.$$

Examples of some simple linear regressions:

1. $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ where $\hat{\beta}_0 = 120, \hat{\beta}_1 = -9.8$, $Y_i =$ quantity, $X_i =$ price.

2. $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ where $\hat{\beta}_0 = 164, \hat{\beta}_1 = 0.27, Y_i =$ housing expenditure, $X_i =$ monthly income. Here, $\hat{\beta}_1$ is equivalent to the marginal propensity to consume housing out of income.

2.8 Goodness of Fit (R^2)

- We can never measure how well SRF can estimate PRF because PRF is not known to allow us to compare.
- But we can assess how well different SRFs fit with the "sample" data that we collect – this can also be called the "goodness of fit".

Residual Concepts:-

$$\begin{aligned} \text{Total Sum of Squares } (SST) &= \\ \text{Explained Sum of Squares } (SSE) &= \\ \text{Residual Sum of Squares } (SSR) &= \\ SST &= \end{aligned}$$

- The R^2, or the coefficient of determination, is defined as

2.9 Incorporating Nonlinearities in Sample Regression

- So far, we have only mentioned the "linear" relation.
- OLS can actually incorporate non-linearities as long as the non-linearities is in the "variables".
- For examples:

$$\log Y_i = \beta_0 + \beta_1 \log X_i + u_i \quad \text{(This is called a constant elasticity model. Why?)}$$

or

$$\log Y_i = \beta_0 + \beta_1 X_i^2 + u_i$$

or

$$Y_i = \beta_0 + \beta_1 \frac{1}{X_i} + u_i.$$

etc. etc...

---

3 Regression Through the Origin

- If you know that the regression line goes through the origin ($X = 0, Y = 0$) for sure, we can impose this restriction
- This restriction makes sense in some contexts. For example, $X = income$ and $Y = income\ tax$.
- The SRF then becomes

$$Y_i = \tilde{\beta}_1 X_i$$

where $\tilde{\beta}_1$ is an OLS estimator. It can be calculated through finding $\tilde{\beta}_1$ that minimizes the sum of squared residual

Multiple Regression Analysis (Estimation)

1 Motivation

The SLR4. assumption, $E(u_i|X_i) = 0$, is unrealistic. This implies that that u_i is uncorrelated with X_i because no matter what the value of X_i is, the expected value of u_i would still be 0! Thus, when this assumption does not hold, OLS estimates (β_0 and β_1 will be biased). (Why? – note: $Cov(X_i, \hat{u}_i) = 0$ is always true by the OLS calculation. This does not mean that $Cov(X_i, u_i) = 0$ in reality.)

In this case, the multiple regression analysis is introduced in order to achieve the condition $E(u_i|X_i) = 0$. It also enables us to explain the dependent variable better and to conduct the "ceteris paribus" or "holding all other things constant" analysis.

Example: If we want to find the relation between wage and education in the simple linear regression, would our β_0 and β_1 be biased? Probably!

- Consider a simple linear regression

1.1 Assumption SLR 4 ($E(u|X) = 0$) in the Multiple Regression Context

Consider the *wage* equation.

- In the case of simple regression, the assumption SLR4 ($E(u|X) = 0$) has to be satisfied in order to achieve an unbiased estimator of β_0 and β_1.
- In this two-variable regression of assumption SLR4 ($E(u|X) = 0$) becomes $E(u|X_1, X_2) = 0$.
- Therefore, the OLS estimator of β_0, β_1 and β_2 would be unbiased if $E(u|educ, inc) = 0$.
- For example, "*innate ability*" is not included in eq. **??**. Thus, if "*innate ability*" can explain *wage*, it would be in u.
- If it is true that $E(innate\ ability|educ, inc) = 0$, or the expected value of *innate ability* is the same and (equal to zero) for all education and income levels, then the OLS estimators $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ would be unbiased. (Is this likely?)

---

2 The Model with k Independent Variables

- The "population" version of the multiple linear regression model can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + u,$$

where

β_0 is the intercept.

β_1 is the parameter associated with X_1.

β_2 is the parameter associated with X_2, and so on.

u is the error term

- $Y, X_1, X_2, ..., X_k$ are variables
- $\beta_0, \beta_1, \beta_2, ..., \beta_k$ are parameters

TABLE 2.1

Terminology for Simple Regression

y	*x*
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

2.1 Accounting Nonlinearity

- The regression model requires linearity in parameters.
- Similar to the Simple Regression Model, the Multiple Regression Model can also take into account the nonlinear relationships between variables.

Example: In the CEO Salary example, we could write the relations between CEO salary ($salary$), firm sales ($sales$) and CEO age (age) as follows:

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 age + \beta_3 age^2 + u$$

- - This model has $k = 3$ because there are 3 regressors. $Y = \log(salary), X_1 = \log(sales), X_2 = age, X_3 = age^2$.

  - β_1 measures the change in $\log(salary)$ with respect to $\log(sales)$, holding other factors fixed. β_1 is the sales elasticity of CEO salary.

  - How do we measure the change in $\log(salary)$ with respect to age, holding other factors fixed?

  - How do we measure the change in $salary$ with respect to age, holding other factors fixed?

- In any case, the OLS estimates of β would be unbiased if

  $$E(u|X_1, X_2, ..., X_k) = 0.$$

  This is the Multiple Regression version of assumption SLR 4–all factors in the unobserved error term should be uncorrelated with the explanatory variables.

---

3 Estimation of parameters and properties of estimators

3.1 Deriving OLS Estimators

- We begin with a mulitple regression with 2 regressors. Regressions with more regressors can be analyzed in the exact same fashion. Let the population regression model be

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

where the estimated OLS equation (sample version) of the above regression can be written as

- As before, the OLS estimators are the ones that minimizes the sum of residual squared given the observations $i = 1, 2, ..., n$ in the sample.

$$\underset{\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2}{\arg\min} \Sigma_{i=1}^{n} (\hat{u}_i)^2 = \underset{\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2}{\arg\min} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})^2$$

First Order Condition (F.O.C):

$$w.r.t.\hat{\beta}_0 \quad \Rightarrow \quad 0 = -2(\Sigma_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i})) \tag{4.1}$$

$$w.r.t.\hat{\beta}_1 \quad \Rightarrow \quad 0 = \tag{4.2}$$

$$w.r.t.\hat{\beta}_2 \quad \Rightarrow \quad 0 = \tag{4.3}$$

- Solving equations 4.1, 4.2 and 4.3 simultaneously, we can derive the solution for $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ (will come back to this when we start using the matrix notation).

3.2 Algebriac Properties of OLS estimators

1.

2.

3.

4.

3.3 How could the multiple regression analysis enable ceteris paribus analysis?

- Consider a multiple regression function of *wage*

$$wage = \beta_0 + \beta_1 educ + \beta_2 inc + u \tag{4.4}$$

Here,

- β_0 is the intercept.

  β_1 measures the change in *wage* with respect to *educ*, holding other factors fixed.

  β_2 measures the change in *wage* with respect to *inc*, holding other factors fixed.

- What if the function of *wage* is, instead written as

$$wage = \beta_0 + \beta_1 educ + \beta_2 inc + \beta_3 educ^2 + u$$

  Then,

  β_0 is the intercept.

  The change in *wage* with respect to *educ* (holding other factors fixed) is measured by:

  The change in *wage* with respect to *inc* (holding other factors fixed) is measured by:

4 Expected Value of the OLS Estimators

- Under assumptions MLR 1 to 4 (see Wooldridge), $\hat{\beta}_{OLS}$ are unbiased.

- 2 issues should be considered regarding the biasedness of $\hat{\beta}_{OLS}$

4.1 Issue #1: Including Irrelevant Variable (Overspecifying the Model)

- Suppose we specify the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \tag{4.5}$$

and this model satisfies the multiple regression assumptions 1 to 4

4.2 Issue #2: Excluding Relevant Variable (Underspecifying the Model → omitted variable bias. This is a serious problem!)

- Suppose we the **TRUE** model is actually

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

where none of the β is zero **and this model satisfies the multiple regression assumptions 1 to 4.**

- But we omit variable X_2 and estimate the following equation using OLS

5 Variance of the OLS Estimators

- The $\hat{\beta}_{OLS}$ would be the most efficient among the linear unbiased estimators if assumption 5 is satistifed
- Multiple Linear Regression (MLR) assumption 5: Homoskedasticity

  The error term u has the same variance given any values of the explanatory variables.

  $$Var(u|X_1, X_2, ..., X_k) = \sigma^2$$

- Example:

- If the MLR assumption 5 is true, then

6 Estimator of the OLS Variance

- Since we don't know what σ^2 is (population concept), we need to find an estimator of it.

- Thus, STATA's calculation of the std.err. of $\hat{\beta}_j$ is

$$\widehat{std.deviation}.\hat{\beta}_j = std.err.\hat{\beta}_j = \sqrt{\frac{\hat{\sigma}^2}{\Sigma_{i=1}^{N}(X_{ij} - \bar{X}_j)^2(1 - R_j^2)}}.$$

Comments:

Stata Lab 1 – Introduction

1 What is STATA?

- A statistical software package used mostly in economics, sociology, political science and epidemiology.
- Stata can be used to manage database, run regressions, generate graphics, do simulations, etc.
- The user should have their own dataset. The Stata data file is usually saved in the .dta format.
- Data of any other formats (like excel) can be imported and/or converted into .dta format.

1.1 STATA supports

- Stata's own website: http://www.stata.com/support/faqs/
- Princeton stata tutorial: https://data.princeton.edu/stata/
- UCLA Academic Technology Services: https://stats.idre.ucla.edu/stata/
- Stata program's help function: For example, suppose you would like to know more about the "regress" command, then... Open the stata program > in the "command" box > type "help regress" without the " "> press enter.
- Stata's official manual (can be viewed through the program)
- Other Stata's user's manual: My favorite one is "An Introduction to Modern Econometrics Using Stata" by Christopher F. Baum.
- or... simply type your question(s) into a search engine.

1.2 Data files and Do-files

- The Stata's data file keeps all the data points. For example, each Y_i and the corresponding X_i, $\forall\ i = 1, 2, 3..., n$.
- The do-file records all the commands that you use to analyze the data.
- Once the data is cleaned, it is best not to keep on re-saving the original data file. If several steps have to be done before analyzing the data (like running a regression), do it on the do-file.

2 Tutorial 1: Exploring the Data and Running a Simple Regression

- **Download Wooldridge datasets**

1. Download Wooldridge's data – go to thomsonedu's website: "http://www.thomsonedu.com/aise/economics/wooldridge_2e_datasets/".
2. save "statafiles.zip" on your computer. Unzip "statafiles.zip".
3. save "excelfiles.zip" on your computer. Unzip "excelfiles.zip".

- **To open the STATA program**

1. Double click on the STATA icon.
2. Click on the "Do-file" icon on the top panel of the Stata program. "Save As" your Do-file (and name it "EE425") on your computer.

- **Open the data file using Do-file**

1. file -> open -> then, direct the program to the file "CEOSAL2.DTA".
2. On the command window, you will see a command to open this file. If you would like to open the file from your do-file in the future, you can use this command.

- **To explore and understand the data**

1. type: browse
2. type: describe
3. type: summarize
4. type: sum
5. type: codebook
6. type: describe salary
7. type: tabulate college
8. type: tab college
9. to use the "if" command to find conditional mean (average) type: sum if grad == 1
10. type: sum salary if age <= 40
11. type: correlate salary sales profits
12. type: correlate salary sales profits, covariance
13. type: plot salary profits
14. type: twoway scatter salary profits

- **To run a simple (OLS) regression (one explanatory variable)**

1. type: regress salary profits

- **To create the fitted value** ($\hat{Y}_i$) **and the residual** ($\hat{u}_i$)

1. type: predict y_hat, xb
2. type: predict u_hat, residual

- **To see how well we do at finding a Sample Linear Function**

1. type: twoway scatter salary profits || line y_hat profits
2. type: twoway scatter u_hat profits
3. To check if the OLS estimation makes X_i uncorrelated with $\hat{u}_i$ (by the OLS calculation, they should not correlate), type: correlate sales u_hat

- **To execute mathematical operations**

1. type: generate log_salary = log(salary)
2. type: gen log_profit = log(profits)
3. type: gen profit_2 = 3+5*profits
4. type: regress salary profits profit_2
5. type: gen profit_sq = profits^2
6. type: regress salary profits profit_sq

- **To perform a multiple regression analysis**

1. type: regress salary profits sales
2. type: regress salary profits sales ceoten

- **To exit your Stata**

1. Save your do-file
2. file -> exit -> don't save (never ever modify your master dataset!)

- **To find out what all the above commands mean** – (type in the command box) help summarize, help predict, help twoway, etc etc.

2.1 Examples from Wooldridge(2012)

C2.2 The data set in CEOSAL2.dta contains information on chief executive officers for U.S. corporations. The variable *salary* is annual compensation, in thousands of dollar, and *ceoten* is prior number of years as company CEO.

1. Find the average salary and the average tenure in the sample
2. How many CEOs are in their first year as CEO (that is, $ceoten = 0$)? What is the longest tenure as a CEO?
3. Estimate the simple regression model

$$\log(salary) = \beta_0 + \beta_1 ceoten + u.$$

and report your results in the usual form. What is the (approximate) predicted percentage increase in salary given one more year as a CEO?

C2.1 Use the data in WAGE2.dta to estimate a simple regression explaining monthly salary (*wage*) in terms of IQ score (IQ).

1. Find the average salary and average IQ in the sample. What is the sample standard deviation of IQ? (IQ score are standardized so that the average in the population is 100 with a standard deviation equal to 15)
2. Estimate a simple regression model where a one percentage point increase in IQ changes *wage* by a constant dollar amount. Use this model to find the predicted increase in *wage* for an increase in IQ of 15 percentage points. Does IQ explain most of the variation in *wage*?
3. Now, estimate a model where each one percentage point increase in IQ has the same percentage effect on *wage*. If IQ increases by 15 percentage points, what is the approximate percentage increase in predicted *wage*?

Multiple Regression Analysis (Inference)

Objectives

1. Students know how to test hypotheses about the parameter (β)

2. Students can test for the validity of the proposed population model.

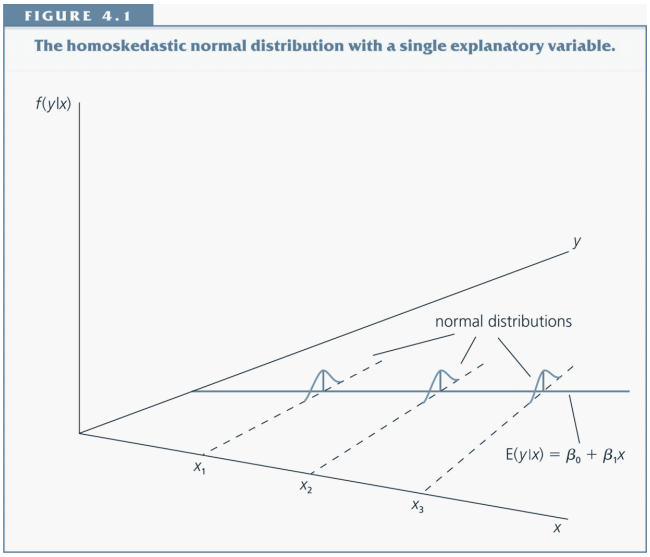
Econometrics Analysis - The Steps

1 Sampling Distribution of the OLS estimators ($\hat{\beta}_{OLS}$)

To be able to test hypotheses about the parameter ($\hat{\beta}$), we need an assumption about the distribution of u.

Assumption MLR 6 - Normality

The population error u is independent of the explanatory variables $X_1, X_2, ...X_k$ and is **normally distributed** with zero mean and variable σ^2.



- By normality of the error term (u), we have nomality of $\hat{\beta}$

Or, using the standardization $\frac{Z-\mu}{\sigma} \sim$ normal (0,1), we have

2 Testing Hypotheses about an individual regression coefficient "the t-test"

- But we do not have $sd.(\hat{\beta}_j)$, we can only calculate the estimator of it :

- for degree of freedom > 30, $t \approx z$.We can use the $Z - score$ table to find the critical value(s).

2.1 Testing Against Two-Sided Alternatives

Consider the wage equation :

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 \exp er + \beta_3 tenure + u.$$

Suppose we want to test whether experience has a partial effect on wage:

$$\begin{aligned} H_0 &: \beta_2 = 0 \quad \text{(experience has no partial effect)} \\ H_a &: \beta_2 \neq 0. \text{ (experience has a partial effect)} \end{aligned}$$

** Note that

Suppose $n = 34$, degree of freedom $= 34 - 3 - 1 = 30$.
From the table t-table, the 5% critical value for a 2-tailed test with 30 d.f. is _ _ _ _ _ _ _ _ _ _

- If we change the significance level to 10%, then the 2-tailed critical value becomes _ _ _ _ _ _ _ _ _ _

- If we cannot reject $H_0 : \beta_2 = 0$ (at a given significance level – e.g. 5%), we say that X_2 is statistically significant at the 5% level.

- In other words X_2 has a partial effect on the expected value of Y.

```
. regress log_wage educ exper tenure
```

Source	SS	df	MS	Number of obs =	526
Model	46.8741776	3	15.6247259	F(3, 522) =	80.39
Residual	101.455574	522	.194359337	Prob > F =	0.0000
				R-squared =	0.3160
				Adj R-squared =	0.3121
Total	148.329751	525	.28253286	Root MSE =	.44086

log_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.092029	.0073299	12.56	0.000	.0776292	.1064288
exper	.0041211	.0017233	2.39	0.017	.0007357	.0075065
tenure	.0220672	.0030936	7.13	0.000	.0159897	.0281448
_cons	.2843595	.1041904	2.73	0.007	.0796756	.4890435

2.2 Testing Against One-Sided Alternatives

Suppose the rule for rejecting H_0 becomes

$$H_a : \beta_j > 0.$$

We could have the H_0 as

$$\begin{aligned} H_0 &: \beta_j \leq 0 \qquad \text{or} \\ H_0 &: \beta_j = 0 \end{aligned}$$

depending on the context. For $H_0 : \beta_j = 0$, it implies that we are ruling out population values of β_j less than zero. For $H_0 : \beta_j \leq 0$,it implies that we are not ruling out population values of β_j less than zero. To given an example of a test against the one-sided alternative hypothesis, consider the wage equation again :

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 \exp er + \beta_3 tenure + u$$

If we want to test whether experience has a positive partial effect on wage

$$\begin{aligned} H_0 &: \beta_2 = 0 \\ H_a &: \beta_2 > 0. \end{aligned}$$

We implicitly rule out the possibility that $\beta_2 < 0$ here.

3 Testing other hypotheses about β_j

- Most common H_0 is $H_0 : \beta_j = 0$.

- However, we can test other types of hypotheses. For example, $H_0 : \beta_j = a_j$ where a_j is a constant number.

If we want to test the H_0 using a 5% significance level, then we reject H_0 if

$$t_{wife\_income} > \_\_\_\_\_ \quad or$$
$$t_{wife\_income} < \_\_\_\_\_$$

4 Testing Hypotheses about a Single Linear Combination of the Parameter

Consider

$$\log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 \exp er + u$$

where jc = number of years attending a two-year college
$univ$ = number of years at a four-year college
$\exp er$ = months in the workforce.
We want to test whether $\beta_1 = \beta_2$.

another possible hypothesis test (one-tailed alternative)

5 Computing p-Values for t-Tests

- What is the significance level given the computed t-statistics?

- p-value : $P(|T| > |t|)$

Example 1: $H_0 : \beta_j \geq 0, \quad H_a : \beta_j < 0, \quad$ d.f.= 140.

suppose the calculated $t_{\hat{\beta}_j} = -2.75$

- From the z-table, the value -2.75 corresponds to area = _ _ _ _ _ _ _ _ _ _ _ _ _.
- Thus, p-value = _ _ _ _ _ _ _ _ _ _ _.
- Would we reject H_0 if we use the significance level = 5%?

Example 2: $H_0 : \beta_j = a_j, \quad H_a : \beta_j \neq a_j, \quad$ d.f.= 18.

suppose the calculated $t_{\hat{\beta}_j} = -2.18$

- From the t-table, the value -2.18 corresponds to area = _ _ _ _ _ _ _ _ _ _ _ _ _.
- Thus, p-value = _ _ _ _ _ _ _ _ _ _ _.
- Would we reject H_0 if we use the significance level = 5%?

6 Confidence Intervals (CI)

- Confidence Intervales for the POPULATION PARAMETER (β_j)
- A 95% CI of β_j is given by

Example 1: **95% CI**

Example 2: **99% CI**

7 Testing Multiple Linear Restrictions: The F-test

Suppose the model is specified by

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u \\ H_0 &: \beta_1 = 0 \text{ and } \beta_2 = 0 \\ H_1 &: H_0 \text{ is not true} \end{aligned}$$

We can use the F-test to test this type of "multiple hypotheses".

1. Our full model is called the "unrestricted" model (ur). Suppose it can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + u$$

2. The model which takes out x (which we think its associated $\beta = 0$) is called the restricted model (r).

3. Some useful facts

4. Other ways to calculate the F-statistics:

Example: Suppose we are interested in understanding the determinant of a baseball player's salary.

$salary$	= season salary
$years$	= years in major leagues
$gamesyr$	= games per year in the league
$bavg$	= career batting average
$hrunsyr$	= homeruns per year
$rbisyr$	= runs batted in per year

- the unrestricted model (ur) is defined by

```
. regress log_salary years gamesyr bavg hrunsyr rbisyr
```

Source	SS	df	MS	Number of obs =	353
				F(5, 347) =	117.06
Model	308.989208	5	61.7978416	Prob > F =	0.0000
Residual	183.186327	347	.527914487	R-squared =	0.6278
				Adj R-squared =	0.6224
Total	492.175535	352	1.39822595	Root MSE =	.72658

log_salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
years	.0688626	.0121145	5.68	0.000	.0450355	.0926898
gamesyr	.0125521	.0026468	4.74	0.000	.0073464	.0177578
bavg	.0009786	.0011035	0.89	0.376	-.0011918	.003149
hrunsyr	.0144295	.016057	0.90	0.369	-.0171518	.0460107
rbisyr	.0107657	.007175	1.50	0.134	-.0033462	.0248776
_cons	11.19242	.2888229	38.75	0.000	10.62435	11.76048

- the restricted model (r) is defined by

```
. regress log_salary years gamesyr
```

Source	SS	df	MS	Number of obs =	353
				F(2, 350) =	259.32
Model	293.864058	2	146.932029	Prob > F =	0.0000
Residual	198.311477	350	.566604221	R-squared =	0.5971
				Adj R-squared =	0.5948
Total	492.175535	352	1.39822595	Root MSE =	.75273

log_salary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
years	.071318	.012505	5.70	0.000	.0467236	.0959124
gamesyr	.0201745	.0013429	15.02	0.000	.0175334	.0228156
_cons	11.2238	.108312	103.62	0.000	11.01078	11.43683

Now, our H_0 and H_a becomes

8 How the Hypothesis Testing is done in Practice

1. Check the values of $t-statistic$ reported by the statistical software (i.e. STATA, SPSS, SAS)

   $\Rightarrow$ These $t-statistics$ are to test $H_0 : \beta_i = 0$

   $\Rightarrow$ If the d.f. > 30, then when $t > 1.96$, we can reject H_0

   $\Rightarrow$ **When $t > 1.96$**, we can say that β_i is **statistically significant** at 5% level. (value of $\beta_i \neq 0$)

   $\Rightarrow$ **When $t < 1.96$** we can say that β_i is **not statistically significant** at 5% level.

   $\Rightarrow$ If $t < 1.96$ we can drop x_i from the model

   $\Rightarrow$ After we drop x_i, we estimate the new regression function and obtain a new set of $\hat{\beta}$.

2. We can also perform other hypothesis testings of interest.

   e.g. $H_0 : \beta_i = \beta_j$

   or $H_0 : \beta_i = 5$ etc.

   or perform an F-test for testing multiple linear restrictions

3. Usually, in economics, the estimation results are reported using this form

Dependent Variable: log(*salary*)			
Independent Variables	**(1)**	**(2)**	**(3)**
log(*sales*)	.224 (.027)	.158 (.040)	.188 (.040)
log(*mktval*)	——	.112 (.050)	.100 (.049)
profmarg	——	−.0023 (.0022)	−.0022 (.0021)
ceoten	——	——	.0171 (.0055)
comten	——	——	−.0092 (.0033)
intercept	4.94 (0.20)	4.62 0.25)	4.57 (0.25)
Observations *R*-squared	177 .281	177 .304	177 .353

Multiple Regression Analysis : Further Issues

1 Data scaling on OLS statistics

When we change the unit of measurment of a variable, the value of estimators would chance accordingly. For example

$$\widehat{bweght} = \widehat{\beta}_0 + \widehat{\beta}_1 cigs + \widehat{\beta}_2 fa\min c,$$

where
$bwght =$ child birth weight, in grams.
$cigs =$ number of cigarettes smoked by the mother while pregnant, per day.
$fa\min c =$ annual family income, in thousands of dollars.

2 More on functional forms

- Logarithmic Functional Form

$$\log(y) = \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + u$$

- Models with Quadratics

Example : Effects of Pullution on Housing Prices

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 \log(dist) + \beta_3 rooms + \beta_4 room^2 + \beta_5 stratio + u$$

where

price = housing price

nox = level of pollution

dist = distance from downtown

rooms = number of rooms

stratio = average student per teacher ratio

The estimation result is given by

```
regress  lprice lnox dist rooms rooms_sq stratio
```

Source	SS	df	MS		
Model	51.4933152	5	10.298663	Number of obs =	506
Residual	33.0889098	500	.06617782	F(5, 500) =	155.62
				Prob > F =	0.0000
				R-squared =	0.6088
				Adj R-squared =	0.6049
Total	84.582225	505	.167489554	Root MSE =	.25725

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnox	-.9767545	.0995938	-9.81	0.000	-1.172429	-.7810806
dist	-.0321972	.0094013	-3.42	0.001	-.050668	-.0137264
rooms	-.5528032	.1612965	-3.43	0.001	-.8697056	-.2359007
rooms_sq	.0624697	.0124867	5.00	0.000	.0379368	.0870025
stratio	-.0486667	.0058131	-8.37	0.000	-.0600879	-.0372455
_cons	13.59154	.5650901	24.05	0.000	12.4813	14.70178

Consider the effect of "room"

What wold be the % change in price when the number of room increases from 5 to 6?

3 Models with Interaction Terms

Consider

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + \beta_3 sqrft \times bdrms + \beta_4 bthrms + u$$

where
$price$ = housing price
$sqrft$ = house size (square feet)
$bdrms$ = number of bedrooms
$bthrms$ = number of bathrooms

4 More on the Goodness-of-Fit and Selection of Regressors

- Adding more regressors ALWAYS improve fit

Using adjusted R-squared to choose between non-nested models (one model is not a subset of another).

Consider Model 1

$$\begin{aligned} \widehat{salary} &= \underset{(223.90)}{830.63} \underset{(0.0089)}{+0.0163 sales} \underset{(11.08)}{+19.63 roe} \\ n &= 209, \quad R^2 = 0.029, \ \bar{R}^2 = 0.020 \end{aligned}$$

Consider Model 2

$$\begin{aligned} \widehat{\log(salary)} &= \underset{(0.29)}{4.36} \underset{(0.033)}{+0.2751 \log(sales)} \underset{(0.004)}{+0.0179 roe} \\ n &= 209, \quad R^2 = 0.282, \ \bar{R}^2 = 0.275 \end{aligned}$$

Matrix Algebra (Quick Review)

(For more basic background, please read Wooldridge Appendix D)

1 Matrix Operation

- Addition - $A + B$ is possible when A and B are of the same dimension.

$$\begin{aligned} A &= \begin{bmatrix} 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 \end{bmatrix}_{2\times 4}, \qquad B = \begin{bmatrix} 1 & 0 & -1 & 3 \\ -2 & 0 & 1 & 5 \end{bmatrix}_{2\times 4} \\ A+B &= \begin{bmatrix} 3 & 3 & 3 & 8 \\ 4 & 7 & 9 & 14 \end{bmatrix}_{2\times 4} \end{aligned}$$

- Subtraction

$$A - B = \begin{bmatrix} & & & \\ & & & \end{bmatrix}_{2\times 4}$$

- Scalar Multiplication

$$\lambda A = \lambda \left[a_{ij}\right]_{i\times j}$$

Suppose $\lambda = 2$

$$\begin{aligned} A &= \begin{bmatrix} -3 & 5 \\ 8 & 7 \end{bmatrix} \\ \lambda A &= \begin{bmatrix} & \\ & \end{bmatrix} \end{aligned}$$

- Matrix Multiplication

$A_{(M\times N)}B_{(O\times P)}$ is possible if an only if $N = O$ ($\leftarrow$ Letter O). The multiplication result would be of dimension _____.

Let

$$\begin{aligned} A &= \begin{bmatrix} 3 & 4 & 7 \\ 5 & 6 & 1 \end{bmatrix}_{2\times 3}, \qquad B = \begin{bmatrix} 2 & 1 \\ 3 & 5 \\ 6 & 2 \end{bmatrix}_{3\times 2} \\ C &= A \times B = \begin{bmatrix} & \\ & \end{bmatrix}_{2\times 2} \\ C &= \begin{bmatrix} & \\ & \end{bmatrix}_{2\times 2} \\ C_{ij} &= \end{aligned}$$

- Matrix Differentiation

Rule1 - If $a' = \begin{bmatrix} a_1 & a_2 & a_3 & ... & ... & a_n \end{bmatrix}$ and $X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ ... \\ x_n \end{bmatrix}$

Then

- Rule2 - If $X'AX = \begin{bmatrix} x_1 & x_2 & x_3 & ... & ... & x_n \end{bmatrix}_{1\times n} \begin{bmatrix} \quad \\ \quad \\ \quad \\ \quad \end{bmatrix}_{n\times n} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ ... \\ x_n \end{bmatrix}_{n\times 1}$

Then

2 OLS Estimation using Matrix Notation

Let

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ .. \\ .. \\ .. \\ .. \\ y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & x_{11} & x_{21} & ... & x_{k1} \\ 1 & x_{12} & & & .. \\ 1 & x_{13} & & & .. \\ .. & .. & & & .. \\ & & & & \\ & & & & \\ & & & & \\ 1 & x_{1n} & & & x_{kn} \end{bmatrix}_{n\times(k+1)} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ ... \\ \hat{\beta}_k \end{bmatrix}_{(k+1)\times 1} + \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ .. \\ .. \\ .. \\ .. \\ \hat{u}_n \end{bmatrix}_{n\times 1}$$

Example: Let the estimated model be

$$y = \beta_0 + \beta_1 x_1 + u$$

Let
$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}_{2\times 1}$

$X'X =$

3 Variance-Covariance Matrix of $(\hat{\beta})$

$$\begin{aligned} \text{From Variance-Covariance Matrix of } (\hat{\beta}) &= E\{[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]'\} \\ &= E\{[\hat{\beta} - \beta][\hat{\beta} - \beta]'\} \\ \text{from } \hat{\beta} &= (X'X)^{-1}X'Y \end{aligned}$$

Multiple Regression Analysis with Qualitative Information:

1 Outline

- Describing qualitative information
- Using a single dummy independent variable
- Using dummy variables for multiple categories
- Interactions involving dummy variables
- A binary dependent variable (Y variable): The linear probability model

2 Describing Qualitative Information

- "Female" and "Married" are qualitative variable.
- We arbitarily assign a dummy variable to decribe them.

$$famale = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases}$$

$$married = \begin{cases} 1 & \text{if married} \\ 0 & \text{otherwise (of if single)} \end{cases}$$

TABLE 7.1

A Partial Listing of the Data in WAGE1.RAW

person	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

3 Models with a single dummy independent variable

Consider

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u.$$

where

$$famale = \begin{cases} 1 & \text{if female} \\ 0 & \text{otherwise (or if male)} \end{cases}$$

In this case, the δ_0 notation is used to highlight the interpretation of the parameters multiplying dummy variables. In other cases, we can use any notation that is the most convenient.

4 It is not possible to include all of the dummy alternatives in the same model

- If we include all alternatives of a dummy variable in the same model, we will face the "perfect collinearity" problem.

For example:

$$\begin{aligned} 1 &= female + male \\ female &= male + 1 \end{aligned}$$

or

$$\begin{aligned} 1 &= winter + spring + summer + fall \\ winter &= 1 - spring - summer - fall \end{aligned}$$

- At least one alternative has to be dropped. We treat the dropped alternative as the "BASE GROUP" or "BASELINE" or "BENCHMARK GROUP".

```
. regress lwage female male married educ exper
note: male omitted because of collinearity

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  4,   521) =   75.27
       Model |  54.3265253     4  13.5816313            Prob > F      =  0.0000
    Residual |  94.0032262   521  .180428457           R-squared     =  0.3663
-------------+------------------------------           Adj R-squared =  0.3614
       Total |  148.329751   525   .28253286           Root MSE      =  .42477

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.3251146   .0377061    -8.62   0.000    -.3991892     -.25104
        male |           0  (omitted)
     married |   .1380145   .0411197     3.36   0.001     .0572338    .2187953
        educ |   .0872644   .0071554    12.20   0.000     .0732075    .1013213
       exper |   .0076213   .0015314     4.98   0.000     .0046129    .0106297
       _cons |   .4690918   .1040575     4.51   0.000      .264668    .6735156
------------------------------------------------------------------------------
```

5 Using dummy variables for multiple categories

Case 1 We can use many dummy variables in the same model

Consider a model which includes 2 dummy variables– $female$ and $married$.

$$\begin{aligned}\log(wage) &= \beta_0 + \delta_0 female + \delta_1 married + \beta_1 educ + \beta_2 \exp er + \beta_3 \exp er^2 \\ &\quad + \beta_4 tenure + \beta_5 tenure^2 + u.\end{aligned}$$

```
regress lwage female married educ exper expersq tenure tenursq
```

Source	SS	df	MS	Number of obs =	526
				F(7, 518) =	58.76
Model	65.6482326	7	9.37831895	Prob > F =	0.0000
Residual	82.6815188	518	.159616832	R-squared =	0.4426
				Adj R-squared =	0.4351
Total	148.329751	525	.28253286	Root MSE =	.39952

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.2901838	.0361121	-8.04	0.000	-.3611279	-.2192396
married	.0529219	.0407561	1.30	0.195	-.0271456	.1329894
educ	.0791547	.0068003	11.64	0.000	.0657952	.0925143
exper	.0269535	.0053258	5.06	0.000	.0164907	.0374163
expersq	-.0005399	.0001122	-4.81	0.000	-.0007603	-.0003196
tenure	.0312962	.0068482	4.57	0.000	.0178426	.0447499
tenursq	-.0005744	.0002347	-2.45	0.015	-.0010355	-.0001134
_cons	.4177837	.0988662	4.23	0.000	.2235557	.6120116

Comments:

Consider a model which includes dummy variables for each gender/marital status combination– $marrmale$, $marrfem$ and $singfem$.

$$\begin{aligned}\log(wage) &= \beta_0 + \delta_0 marrmale + \delta_1 marrfem + \delta_3 singfem + \beta_1 educ + \beta_2 exper \\ &\quad + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u. \end{aligned} \tag{9.1}$$

```
regress lwage marrmale marrfem singfem educ exper expersq tenure tenursq
```

Source	SS	df	MS	Number of obs =	526
				F(8, 517) =	55.25
Model	68.3617623	8	8.54522029	Prob > F =	0.0000
Residual	79.9679891	517	.154676961	R-squared =	0.4609
				Adj R-squared =	0.4525
Total	148.329751	525	.28253286	Root MSE =	.39329

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
marrmale	.2126757	.0553572	3.84	0.000	.103923	.3214284
marrfem	-.1982676	.0578355	-3.43	0.001	-.311889	-.0846462
singfem	-.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
educ	.0789103	.0066945	11.79	0.000	.0657585	.092062
exper	.0268006	.0052428	5.11	0.000	.0165007	.0371005
expersq	-.0005352	.0001104	-4.85	0.000	-.0007522	-.0003183
tenure	.0290875	.006762	4.30	0.000	.0158031	.0423719
tenursq	-.0005331	.0002312	-2.31	0.022	-.0009874	-.0000789
_cons	.3213781	.100009	3.21	0.001	.1249041	.5178521

Comments:

Case 2 We can use dummy variables to represent multiple categories of a variable Consider the relationship between law school rankings and starting salaries

$$\begin{aligned}\log(salary) &= \beta_0 + \delta_0 top10 + \delta_1 r11\_25 + \delta_3 r26\_40 + \delta_4 r41\_60 + \beta_1 LSAT \\ &\quad + \beta_2 GPA + \beta_3 \log(libvol) + \beta_4 \log(cost) + u.\end{aligned}$$

where $top10$, $r11\_25$, $r26\_40$, $r41\_60$ would be equal to 1 when the variable $rank$ falls into the appropriate range.

** Rank below 60 would be the base case.

```
. regress lsalary top10 r11_25 r26_40 r41_60 LSAT GPA llibvol lcost
```

Source	SS	df	MS	Number of obs =	136
				F(8, 127) =	120.15
Model	9.16538532	8	1.14567316	Prob > F =	0.0000
Residual	1.2109665	127	.009535169	R-squared =	0.8833
				Adj R-squared =	0.8759
Total	10.3763518	135	.076861865	Root MSE =	.09765

lsalary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
top10	.5393428	.053542	10.07	0.000	.4333927	.6452928
r11_25	.4716199	.0390921	12.06	0.000	.3942637	.548976
r26_40	.2790977	.0346972	8.04	0.000	.2104383	.3477571
r41_60	.182382	.0283098	6.44	0.000	.126362	.238402
LSAT	.0060482	.0034919	1.73	0.086	-.0008616	.012958
GPA	.1305893	.0818678	1.60	0.113	-.0314122	.2925908
llibvol	.0725522	.0289213	2.51	0.013	.0153221	.1297824
lcost	.0249169	.0283224	0.88	0.381	-.031128	.0809619
_cons	8.363103	.4457314	18.76	0.000	7.481081	9.245125

Comments:

6 Interactions involving dummy variables

Case 1 Interactions among dummies

** We can use interactions among dummies to account for the effect of each combination of dummies as well:

A different way to estimate eq.(9.1) is

$$\begin{aligned}\log(wage) &= \beta_0 + \delta_0 female + \delta_1 married + \delta_3 female \cdot married + \beta_1 educ \\ &\quad + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u.\end{aligned}$$

where $female{\cdot}married = female \times married$.

```
. gen female_married = female*married

. regress lwage female married female_married educ exper expersq tenure tenursq
```

Source	SS	df	MS		
Model	68.3617623	8	8.54522029	Number of obs	= 526
Residual	79.9679891	517	.154676961	F(8, 517)	= 55.25
				Prob > F	= 0.0000
				R-squared	= 0.4609
				Adj R-squared	= 0.4525
Total	148.329751	525	.28253286	Root MSE	= .39329

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.1103502	.0557421	-1.98	0.048	-.219859	-.0008414
married	.2126757	.0553572	3.84	0.000	.103923	.3214284
female_married	-.3005931	.071767	-4.19	0.000	-.4415838	-.1596024
educ	.0789103	.0066945	11.79	0.000	.0657585	.092062
exper	.0268006	.0052428	5.11	0.000	.0165007	.0371005
expersq	-.0005352	.0001104	-4.85	0.000	-.0007522	-.0003183
tenure	.0290875	.006762	4.30	0.000	.0158031	.0423719
tenursq	-.0005331	.0002312	-2.31	0.022	-.0009874	-.0000789
_cons	.3213781	.100009	3.21	0.001	.1249041	.5178521

Comments:

Case 2 Interaction between a dummy and a continuous variable

$$\begin{aligned}\log(wage) &= \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female \cdot educ + \beta_2 exper \\ &\quad + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + u.\end{aligned}$$

```
. gen female_educ = female*educ

. regress lwage female married female_educ educ exper expersq tenure tenursq
```

Source	SS	df	MS		
Model	65.677852	8	8.2097315	Number of obs	= 526
Residual	82.6518994	517	.159868277	F(8, 517)	= 51.35
				Prob > F	= 0.0000
Total	148.329751	525	.28253286	R-squared	= 0.4428
				Adj R-squared	= 0.4342
				Root MSE	= .39984

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-.2197774	.1675154	-1.31	0.190	-.5488721	.1093172
married	.0529779	.0407884	1.30	0.195	-.0271535	.1331092
female_educ	-.0056186	.0130532	-0.43	0.667	-.0312625	.0200254
educ	.0813472	.0085008	9.57	0.000	.0646469	.0980476
exper	.0268542	.005335	5.03	0.000	.0163733	.0373351
expersq	-.0005375	.0001124	-4.78	0.000	-.0007583	-.0003167
tenure	.0314803	.0068669	4.58	0.000	.0179898	.0449709
tenursq	-.0005792	.0002351	-2.46	0.014	-.0010412	-.0001173
_cons	.389629	.1186101	3.28	0.001	.1566119	.6226461

Comments:

7 Testing for Differences in Regression Functions across Groups

- Is it reasonable to believe that the population regression function that explains the dependent variable is the same across subsamples of populations?
- For example, is it reasonable to believe that the function that explains "GPA of college athlete" is the same for male and female students?

Consider

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u,$$

where
$cumgpa = cummulative\ GPA$
$sat = SAT$ score
$hsperc =$ high school rank percentile
$tothrs =$ total hours of college courses

- If we want to test whether "male" students and "female" students have the same values of $\beta_0, \beta_1, \beta_2, \beta_3$ we can estimate the following model

$$\begin{aligned} cumgpa &= \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc + \delta_2 female \cdot hsperc \\ &\quad + \beta_3 tothrs + \delta_3 female \cdot tothrs + u, \end{aligned}$$

and the null hypothesis would be

$$\begin{aligned} H_0 &: \delta_0 = \delta_1 = \delta_2 = \delta_3 = 0 \\ H_a &: otherwise \text{ (at least one } \delta_j = 0) \end{aligned} \tag{9.2}$$

We can use the F-test to test this type of null hypothesis:

1. The restricted model (r)

```
. regress cumgpa sat hsperc tothrs
```

Source	SS	df	MS		
Model	168.533658	3	56.1778861	Number of obs	= 732
Residual	547.364897	728	.751874858	F(3, 728)	= 74.72
				Prob > F	= 0.0000
				R-squared	= 0.2354
				Adj R-squared	= 0.2323
Total	715.898555	731	.979341389	Root MSE	= .86711

cumgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat	.0009028	.0002079	4.34	0.000	.0004947	.0013109
hsperc	-.0063791	.0015678	-4.07	0.000	-.0094572	-.0033011
tothrs	.0119779	.0009314	12.86	0.000	.0101494	.0138064
_cons	.9291105	.2285515	4.07	0.000	.4804118	1.377809

2. The unrestricted model (ur)

```
. gen female_sat = female*sat

. gen female_hsperc = female*hsperc

. gen female_tothrs = female*tothrs

. regress cumgpa female sat female_sat hsperc female_hsperc tothrs female_tothrs
```

Source	SS	df	MS	Number of obs =	732
				F(7, 724) =	35.15
Model	181.589407	7	25.9413439	Prob > F =	0.0000
Residual	534.309148	724	.73799606	R-squared =	0.2537
				Adj R-squared =	0.2464
Total	715.898555	731	.979341389	Root MSE =	.85907

cumgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-1.113638	.528539	-2.11	0.035	-2.15129	-.0759859
sat	.0006113	.000235	2.60	0.009	.0001499	.0010727
female_sat	.0011167	.0005	2.23	0.026	.0001351	.0020984
hsperc	-.0059675	.0017765	-3.36	0.001	-.0094551	-.0024798
female_hsperc	.0000508	.0041025	0.01	0.990	-.0080035	.008105
tothrs	.0103004	.0010928	9.43	0.000	.0081549	.0124459
female_tothrs	.0055599	.0020696	2.69	0.007	.0014968	.009623
_cons	1.213984	.2648281	4.58	0.000	.6940617	1.733907

Comments:

7.1 We can use the "Chow statistics" to test this type of hypothesis as well

- When there are many variables in the model, adding an interaction for every explanatory variable would make the regression analysis messy.
- In which case, we can use the "Chow test" or "Chow statistic" to test the hypothesis expressed in (9.2).
- Chow statistic is a type of F-statistic.

$$F = \frac{SSR_p - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k+1)]}{k+1},$$

where n is the total number of observations.

$SSR_p = SSR$ from the pooled model (include observations from both subsamples)

$SSR_1 = SSR$ from subsample 1
$SSR_2 = SSR$ from subsample 2

```
regress cumgpa sat hsperc tothrs if female == 0
```

Source	SS	df	MS	Number of obs =	552
				F(3, 548) =	41.94
Model	89.6937042	3	29.8979014	Prob > F =	0.0000
Residual	390.619421	548	.712809162	R-squared =	0.1867
				Adj R-squared =	0.1823
Total	480.313125	551	.871711661	Root MSE =	.84428

cumgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat	.0006113	.000231	2.65	0.008	.0001576	.001065
hsperc	-.0059675	.0017459	-3.42	0.001	-.0093969	-.002538
tothrs	.0103004	.001074	9.59	0.000	.0081907	.0124101
_cons	1.213984	.2602697	4.66	0.000	.7027359	1.725233

```
regress cumgpa sat hsperc tothrs if female == 1
```

Source	SS	df	MS	Number of obs =	180
				F(3, 176) =	34.08
Model	83.4816253	3	27.8272084	Prob > F =	0.0000
Residual	143.689727	176	.816418902	R-squared =	0.3675
				Adj R-squared =	0.3567
Total	227.171352	179	1.2691137	Root MSE =	.90356

cumgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sat	.0017281	.0004642	3.72	0.000	.0008119	.0026442
hsperc	-.0059167	.0038895	-1.52	0.130	-.0135927	.0017594
tothrs	.0158603	.0018485	8.58	0.000	.0122122	.0195085
_cons	.1003465	.4810947	0.21	0.835	-.8491105	1.049803

Comments:

8 A Binary Dependent Variable (y variable): The Linear Probability Model

- So far, our Y variables are continuous.
- What if we are interested in explaining a qualitative Y variable (that is, Y is a dummy variable)?

Consider

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \\ E(y|\mathbf{x}) &= \beta_0 + \beta_1 x_1 + ... + \beta_k x_k, \end{aligned}$$

where $\mathbf{x}$ denotes all of the explanatory variables $(x_1, ..., x_k)$.

```
. regress inlf nwifeinc educ exper expersq age kidslt6 kidsge6
```

Source	SS	df	MS		
Model	48.8080578	7	6.97257969	Number of obs	= 753
Residual	135.919698	745	.182442547	F(7, 745)	= 38.22
				Prob > F	= 0.0000
Total	184.727756	752	.245648611	R-squared	= 0.2642
				Adj R-squared	= 0.2573
				Root MSE	= .42713

inlf	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nwifeinc	-.0034052	.0014485	-2.35	0.019	-.0062488	-.0005616
educ	.0379953	.007376	5.15	0.000	.023515	.0524756
exper	.0394924	.0056727	6.96	0.000	.0283561	.0506287
expersq	-.0005963	.0001848	-3.23	0.001	-.0009591	-.0002335
age	-.0160908	.0024847	-6.48	0.000	-.0209686	-.011213
kidslt6	-.2618105	.0335058	-7.81	0.000	-.3275875	-.1960335
kidsge6	.0130122	.013196	0.99	0.324	-.0128935	.0389179
_cons	.5855192	.154178	3.80	0.000	.2828442	.8881943

where

$inlf$ = 1 if the woman reports working for a wage outside the home at some point during the year, zero otherwise.

$nwifeinc$ = husband's earnings (in thousands of dollars)

$educ$ = years of education

$exper$ = past years of labor market experience

age = age

$kidslt6$ = number of children less than 6 years old

$kidsage6$ = number of kinds between 6 - 18 years old

Heteroscedasticity Problem

1 Nature and Consequences of heteroscedasticity for OLS

- Heteroskedasticity (broadly) -

- Heteroskedasticity (in econometrics) -

1.1 Nature of Heteroskedasticity

1.2 Consequences of Heteroskedasticity

1. Does not affect the biasedness of the OLS estimators

2. Does not affect the value of R^2 and $adj.R^2$

3. Make the estimated value of $Var(\widehat{\beta}_{OLS})$ wrong

4. Affect the correctness of our inference

1.3 How can the estimated value of $Var(\widehat{\beta}_{OLS})$ be wrong?

Suppose

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

Given that assumption 1 to 4 are true, but assumption 5 (homoskedasticity) is violated. Thus,

$$Var(u_i|x_i) =$$

And from the OLS estimation steps, we can write

$$\hat{\beta}_1 = \beta_1 +$$

1.4 Two types of remedies

1. Passive

2. Active

2 Testing for heteroskedasticity

- The main point -

Suppose

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

Assume that assumption 1 to 4 are true. Our hypotheses to test for heteroskeroskedasticity would be

We know that $Var(u|\mathbf{x}) = E(u^2|\mathbf{x}) - [E(u|\mathbf{x})]^2$. But _____ according to assumption 4. Thus, H_0 and H_a can be written as

2.1 Breusch-Pagan test (BP test)

To perform the Breusch-Pagan Test in STATA

STATA commands (in case $k = 4$):

regress y x_1 x_2 x_3 x_4

predict u_hat, residual

generate u_hat_sq = u_hat^2

regress u_hat_sq x_1 x_2 x_3 x_4

** Then, check the F-statistic on the top right-hand corner of the result table.

Example: Finding the determinants of GPA.

```
. regress  termgpa attend priGPA final frosh soph
```

Source	SS	df	MS		
Model	226.077541	5	45.2155081	Number of obs	= 680
Residual	142.319996	674	.211157264	F(5, 674)	= 214.13
				Prob > F	= 0.0000
Total	368.397537	679	.542558964	R-squared	= 0.6137
				Adj R-squared	= 0.6108
				Root MSE	= .45952

termgpa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attend	.046594	.0036082	12.91	0.000	.0395093	.0536787
priGPA	.5329307	.0403281	13.21	0.000	.4537468	.6121146
final	.0503197	.0040339	12.47	0.000	.0423992	.0582403
frosh	.0974307	.0560211	1.74	0.082	-.0125662	.2074276
soph	.0689273	.0467006	1.48	0.140	-.0227689	.1606236
_cons	-1.361077	.1316861	-10.34	0.000	-1.619642	-1.102513

```
. predict u_hat, residual

. generate u_hat_sq = u_hat^2

. regress u_hat_sq  attend priGPA final frosh soph

regress u_hat_sq attend priGPA final frosh soph
```

Source	SS	df	MS	Number of obs =	680
				F(5, 674) =	14.54
Model	8.22606613	5	1.64521323	Prob > F =	0.0000
Residual	76.2624962	674	.113149104	R-squared =	0.0974
				Adj R-squared =	0.0907
Total	84.4885623	679	.124430872	Root MSE =	.33638

u_hat_sq	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
attend	-.0088079	.0026413	-3.33	0.001	-.0139941	-.0036218
priGPA	-.1454432	.029521	-4.93	0.000	-.2034074	-.0874791
final	.0061879	.0029529	2.10	0.036	.0003899	.0119859
frosh	-.1077493	.0410085	-2.63	0.009	-.1882692	-.0272294
soph	-.0975658	.0341858	-2.85	0.004	-.1646892	-.0304423
_cons	.7368933	.0963968	7.64	0.000	.5476191	.9261674

Alternatively, you can use the following set of STATA commands:
regress y x_1 x_2 x_3 x_4
estat hettest x_1 x_2 x_3 x_4

```
. regress termgpa attend priGPA final frosh soph

      Source |       SS       df       MS              Number of obs =     680
-------------+------------------------------           F(  5,   674) =  214.13
       Model |  226.077541     5  45.2155081            Prob > F      =  0.0000
    Residual |  142.319996   674  .211157264           R-squared     =  0.6137
-------------+------------------------------           Adj R-squared =  0.6108
       Total |  368.397537   679  .542558964           Root MSE      =  .45952

------------------------------------------------------------------------------
     termgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      attend |    .046594   .0036082    12.91   0.000     .0395093    .0536787
      priGPA |   .5329307   .0403281    13.21   0.000     .4537468    .6121146
       final |   .0503197   .0040339    12.47   0.000     .0423992    .0582403
       frosh |   .0974307   .0560211     1.74   0.082    -.0125662    .2074276
        soph |   .0689273   .0467006     1.48   0.140    -.0227689    .1606236
       _cons |  -1.361077   .1316861   -10.34   0.000    -1.619642   -1.102513
------------------------------------------------------------------------------

. estat hettest attend priGPA final frosh soph

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: attend priGPA final frosh soph

         chi2(5)      =    93.90
         Prob > chi2  =   0.0000
```

If the null hypothesis is rejected (we have the heteroskedasticity problem), we can use the ", robust" option in STATA. This option gives us the correct standard error, or "heteroskedasticity-robust standard error". We can now use the t-statistics in this case.

```
. regress termgpa attend priGPA final frosh soph, robust

Linear regression                               Number of obs =     680
                                                F(  5,   674) =  234.28
                                                Prob > F      =  0.0000
                                                R-squared     =  0.6137
                                                Root MSE      =  .45952

------------------------------------------------------------------------------
             |               Robust
     termgpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      attend |    .046594   .0044101    10.57   0.000     .0379348    .0552532
      priGPA |   .5329307   .0426426    12.50   0.000     .4492023     .616659
       final |   .0503197   .0041066    12.25   0.000     .0422564     .058383
       frosh |   .0974307   .0633543     1.54   0.125    -.0269648    .2218262
        soph |   .0689273   .0520495     1.32   0.186    -.0332713    .1711259
       _cons |  -1.361077   .1448208    -9.40   0.000    -1.645431   -1.076723
------------------------------------------------------------------------------
```

2.2 The White Test

Similar to the Breush-Pagan test, but is stricter because it does not allow $\widehat{u}^2$ to be correlated with x^2 or interactions among different x_s.

Suppose

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

The White Test (special case) (save degree of freedom)

1. Get $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ by OLS.
2. Calculate $\hat{u}_i^2 = [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k)]^2$
3. Calculate $\hat{y}_i = (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k)$
4. Calculate $\hat{y}_i^2$
5. Estimate $\hat{u}_i^2 = \gamma_0 + \gamma_1 \hat{y}_i + \gamma_2 \hat{y}^2 + error$ (keep R^2 of this regression)
6. $LM = nR^2$
7. If $p - value > significance\ level$, cannot reject H_o.

3 Remedial measures

As mentioned before, there are 2 types of remedies – passive and active.

- The passive remedies just re-calculate the *std.err.* or $\hat{\beta}$ using the heteroskedasticity-robust standard error formula(s).
- The active remedies include the "weighted least squares (WLS) estimators", "generalized least squares (GLS) estimators", or "feasible GLS estimator".

3.1 Weighted Least Squares (WLS)

We assume that the heteroskedasticity may take the pattern

From

$$\begin{aligned} Var(u_i|\mathbf{x}) &= E(u_i^2|\mathbf{x}) - [E(u_i|\mathbf{x})]^2 \\ &= \end{aligned}$$

We get

To make the error term become homoskedastic, we weight every term by $\sqrt{h_i}$.

How do we find h_i *or* $\sqrt{h_i}$, the heteroskedasticity function?

1. If the heteroskedasticity is "known" to be caused by a multiplicative constant, we can adjust using that constant.

2. If the heteroskedasticity pattern is not known, we can estimate it. This procedure is called "Feasible Generalized Least Squares" (also called Feasible GLS or FGLS)

3.2 Feasible GLS

Since $Var(\hat{\beta}_j)$ would not be unbiased, we can make valid inferences about β_j, e.g. can use t-test, F-test, etc.

Feasible GLS in practice

1. Get $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ by OLS. (regress y x_1 x_2 .... x_k)
2. Calculate $\hat{u}_i = [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_k x_k)]$ (predict u_hat, residual)
3. Create $\log(\hat{u}_i^2)$ (generate log_u_sq = log(u_hatˆ2))
4. Estimate $\log(\hat{u}_i^2) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_k x_k + error$ (regress log_u_sq x_1 x_2 ... x_k)
5. Obtain the fitted value of $\widehat{\log(\hat{u}_i^2)}$, called $\hat{g}$.(predict g_hat, xb)
6. Create $\hat{h} = \exp(\hat{g})$ (generate h_hat = exp(g_hat))
7. Divide y and each x_{ij} by $\sqrt{\hat{h}}$
8. Estimate $\frac{y}{\sqrt{\hat{h}}} = \frac{\lambda_0}{\sqrt{\hat{h}}} + \lambda_1 \frac{x_1}{\sqrt{\hat{h}}} + \lambda_2 \frac{x_2}{\sqrt{\hat{h}}} + ... + \lambda_k \frac{x_k}{\sqrt{\hat{h}}} + \frac{error}{\sqrt{\hat{h}}}$

Steps 7 & 8 on STATA would be: regress y x_1 x_2 ... x_k [aweight = $\frac{1}{\sqrt{\hat{h}}}$]

3.3 What if the assumed h_i function is wrong?

More on Specification and Data Issues

- Heteroskedasticity problem - not serious

- Specification error problem -serious!

1 Functional Form Misspecification

Suppose the true model is

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 \exp er + \beta_3 \exp er^2 + u.$$

- But we do not include $\exp er^2$, and estimate

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 \exp er + u.$$

- Our $\beta_0, \beta_1, \beta_2$ could be biased.
- The severity of biasedness depends on

1.1 The RESET test

Rationale behind – add polynomial terms $(x^2, x^3, ...)$ to the model and test whether the coefficients are significant. If significant, then we need to include non-linear terms.

Suppose we have a linear baseline model to begin with

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \tag{11.1}$$

To save the # of regressors (save d.f.), we can instead estimate

1.2 Tests against Nonnested Alternatives

For example, to test

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \qquad \text{(Model \#1)}$$

against

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u. \qquad \text{(Model \#2)}$$

Which one should we use? .... There are many methods which can help us justify:

1. Use the RESET test

2. Mizon and Richard (1986)

3. Davidson-MacKinnon Test

Rationale - If Model#1 is true, then the fitted values ($\hat{y}$) from Model#2 should not be significant!

2 Using Proxy Variables for Unobserved Explanatory Variables

In many cases, we cannot observe some important variables

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 \exp er + \beta_3 ability + u.$$

- Ability is unobserved, but is very likely to explain wage or log(wage).

- Ability is very likely to be correlated with *educ* and $\exp er$. If ability is omitted, $\hat{\beta}_1$ and $\hat{\beta}_2$ will be biased.

Remedy (use proxy variables (this chapter) or use instrumental variable method (not in this chapter))

2.1 Using Proxy variables

2.2 Using Lagged Dependent Variables as Proxy Variables

3 OLS with Measurement Error

Suppose

$$GDP = \beta_0 + \beta_1 \exp ort + \beta_2 investment + \beta_3 \inf lation + \beta_4 educ + ... + u$$

- How do you know if a country's GDP is correctly measured?
- What about the measurement of other variables?

3.1 Measurement Error in the Dependent Variable

If there is a measurement error in y variable (dependent variable), then

3.2 Measurement Error in the Explanatory Variable

Suppose the true model is

$$y = \beta_0 + \beta_1 x_1^* + u.$$

4 Missing Data, Nonrandom Samples, Outlying Observations

4.1 Missing Data

Household id.	head_educ	total_income	members
1	12	12,000	2
2	14	25,000	3
3	.	60,000	2
4	4	.	4
5	8	120,000	4
6	16	32,900	2

- STATA would drop the observation with missing data. So, household 3 and 4 would not be used in the regression.
- **If the missing data are missing at random, then $\hat{\beta}_{OLS}$ would not be biased.**

4.2 Nonrandom Samples

4.3 Outliers and Influential Observations

5 Models with Random Slopes

- What if the partial effects of a variale depends on unobserved factors that "vary" by population unit?

$$y_i = a_i + b_i x_i$$

- This is called a random coefficient model, or a random slope model (u_i would then be absorbed into a_i).
- a_i, b_i are viewed as a random draw from the population along with the observed data.
- In any case, we cannot estimate the actual a_i, b_i (for each population unit). We can only find their average (average partial effect, APE).

$$\beta_0 = E(a_i) \quad and \quad \beta_1 = E(b_i).$$

Basic Regression Analysis with Time Series Data

1 The Nature of Time Series Data

TABLE 10.1

Partial Listing of Data on U.S. Inflation and Unemployment Rates, 1948–2003

Year	Inflation	Unemployment
1948	8.1	3.8
1949	−1.2	5.9
1950	1.3	5.3
1951	7.9	3.3
⋮	⋮	⋮
1998	1.6	4.5
1999	2.2	4.2
2000	3.4	4.0
2001	2.8	4.7
2002	1.6	5.8
2003	2.3	6.0

2 Examples of Time Series Regression Models

There are many time series regression models. Different models would be suitable for different types of relationship we want to estimate. Some examples of time series models are Static Model, AR (Autoregressive), ADL (Autoregressive Distributed Lag), FDL (Finite Distributed Lag), ARMA (Autoregressive Moving Average), ARCH (Autoregressive Conditional Heteroskedasticity) etc.

In this class we will talk about 2 examples 1) Static Models and 2) FDL.

2.1 Static Models

Studies a contemperaneous (occurring in the same period of time) relationship of variables.

For example:

2.2 Finite Distributed Lag Models

For example,

In general,

$$y_t = \alpha_0 + \delta_0 x_t + \delta_1 x_{t-1} + \delta_2 x_{t-2} + u_t$$

$\delta_0 = \frac{dy_t}{dx_t}$

$\delta_1 = \frac{dy_t}{dx_{t-1}}$

- Temporary change – assume that "x" changes from "c" to "$c+1$" in period t and remain at c in other periods.

- Permanent change – assume that "x" takes the value of "$c+1$" instead of "c" from period t on.

Compared with y_{t-1}, the value of y from period $t+2$ increases "permanently" by "$\delta_0+\delta_1+\delta_2$". This is called the "long-run" propensity or "long-run multiplier".

3 Properties of OLS under classical assumptions

Assumption TS1. Linear in Parameter – Y is linear in X.

Assumption TS2. No Perfect Collinearity

Assumption TS3. Zero Conditional Mean

***** Under Assumptions TS1 to TS3, $\hat{\beta}_{OLS}$ would be unbiased *****

Assumption TS4. Homoskedasticity

Assumption TS5. No Serial Correlation

***** Under Assumptions TS1 to TS5, $\hat{\beta}_{OLS}$ would be BLUE (best linear unbiased estimators)*****

3.1 The variance of $\hat{\beta}_{OLS}$ (under ass.TS1-TS5)

Serial Correlation and Heteroskedasticity in Time Series Regressions

1 Properties of OLS with Serially Correlated Errors

2 Unbiasedness and Consistency

3 Efficiency and Inference

With serial correlation, $\hat{\beta}_{OLS}$ would not be BLUE ($var(\hat{\beta}_{OLS})$ would not be minimized). Consider

$$u_t = \rho u_{t-1} + e_t \ ; \ t = 1, 2, ..., n \quad \text{and} \quad |\rho| < 1$$

where u_t is from a regression model

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

4 Testing for Serial Correlation

Given the model

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} + u_t$$

4.1 A "t-test" for AR(1) serial correlation with strictly exogeneous regressors

The most common type of serial correlation or autocorrelation is the AR(1) type:

To perform the test:

1. Estimate $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} + u_t$
2. Obtain $\hat{u}_t$,$\hat{u}_{t-1}$; $\forall\ t = 1, 2, ..., n$
3. Estimate $\hat{u}_t = \rho\ \hat{u}_{t-1} + error$

4. Perform the $t-test$ for

4.2 The Durbin-Watson Test (DW test)

This implies

$$\begin{aligned} \hat{\rho} &= 0 \quad \Rightarrow \quad DW = 2 \\ \hat{\rho} &> 0 \quad \Rightarrow \quad DW < 2 \\ \hat{\rho} &< 0 \quad \Rightarrow \quad DW > 2 \end{aligned}$$

$$\begin{aligned} H_o &: \text{ no positive autocorrelation, serial-correlation} \\ H_a &: \text{ no negative serial correlation} \end{aligned}$$

To perform the test:

1. Estimate $y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} + u_t$
2. Obtain $\hat{u}_t$,$\hat{u}_{t-1}$; $\forall\ t = 1, 2, ..., n$
3. Calculate DW from eq.(2)
4. Find the critical d_L and d_u values (say, at the 5% level of significance) for the given sample size and # of regressors.
5. Follow the decision rule in the picture.

Example:
Suppose the calculated value of $DW = 0.80$, $n = 45$, $k = 4$.
From this, we get $d_L =$ _____ and $d_u =$ _____

4.3 Testing for AR(1) serial correlation "without" strictly exogenous regressors

4.4 Testing for AR(q) serial correlation "without" strictly exogenous regressors

5 Correcting for serial correlation

5.1 Passive way

Use the type of standard error that is robust to the serial correlation, autocorrelation problem

5.2 Active way –

Lab 2 – Dummy, Heteroskedasticity, Specification Issues

1 Does "beauty" help increase wage?

1. Download the datafile "beauty.xlsx" from your EE425 Moodle page.
2. Open the STATA software program. Click on the "Data Editor" icon.
3. Copy the entire dataset from the excel file and paste it onto the STATA's Data Editor page.

   Choose "Treat first raw as variable names".
4. Save the new STATA dataset.

   Choose File -> Save As -> (then name the new dataset "beauty_stata")
5. Open a new do-file and save it.

   Choose "New Do-file Editor" icon

   On the Do–file's top panel, choose File -> Save As -> (then name the new do-file "second_stata_lab")

   - **To explore and understand the data**
6. type: sum
7. type: codebook
8. type: tab look
9. type: tab look female

   - **We want to test whether the fraction of above-average-looking women and men are the same**

$$\begin{aligned} abvavg &= \beta_1 famale \\ H_0 &: \beta_1 = 0.5 \\ H_0 &: \beta_1 \neq 0.5 \end{aligned}$$

10. type: gen abvavg = 0
11. type: replace abvavg = 1 if look > 3
12. type: regress abvavg female, nocon

13. Can we reject H_0?

- **We want to test whether "good look" has a positive impact on wage**

14. type: gen belavg = 0
15. type: replace belavg = 1 if look < 3
16. type: gen log_wage = log(wage)
17. regress log_wage belavg abvavg

- **Seems like we may have the omitted variable bias. Let's take into account other variables.**

18. type: regress log_wage abvavg belavg educ
19. type: regress log_wage abvavg belavg educ exper expersq
20. type: regress log_wage abvavg belavg educ exper expersq bigcity
21. type: regress log_wage abvavg belavg educ exper expersq bigcity black
22. type: regress log_wage abvavg belavg educ exper expersq bigcity black union
23. type: regress log_wage abvavg belavg educ exper expersq bigcity black union female

- **Check if we have the heteroskedasticity problem. (Let's use the White Test (special case))**

24. type: regress log_wage abvavg belavg educ exper expersq bigcity black union female
25. type: predict u_hat, resid
26. type: predict y_hat, xb
27. type: gen u_hat_sq = u_hat^2
28. type: gen y_hat_sq = y_hat^2
29. type: regress u_hat_sq y_hat y_hat_sq
30. Calculate $LM = nR^2$
31. Do we reject H_0 : homoskedasticity at 5% level of confidence?

32. Now, try using the ready-made test by STATA
33. type: regress log_wage abvavg belavg educ exper expersq bigcity black union female
34. type: estat hettest
35. Do we reject H_0 : homoskedasticity at 5% level of confidence?

    - **Should we believe that the value of β are the same for female and male?** (Chow Test)
    - Chow statistic is a type of F-statistic $F = \frac{SSR_p-(SSR_1+SSR_2)}{SSR_1+SSR_2} \cdot \frac{[n-2(k+1)]}{k+1}$

36. To get SSR_p: regress log_wage abvavg belavg educ exper expersq bigcity black union
37. To get SSR_1: regress log_wage abvavg belavg educ exper expersq bigcity black union if female == 0
38. To get SSR_2: regress log_wage abvavg belavg educ exper expersq bigcity black union if female == 1
39. What is the value of the F-statistic? Can we reject H_0 (can use the same model)?

2 Fixing Heteroskedasticity

1. Download the "GPA1.DTA" dataset from your EE425 Moodle page and open it in STATA.

   Choose File -> Open -> (then direct to the location of the file)
2. type: des
3. type: regress colGPA hsGPA ACT skipped PC

   - **Check if we have the heteroskedasticity problem. (Let's use the White Test (special case))**

4. type: predict u_hat, resid
5. type: predict y_hat, xb
6. type: gen u_hat_sq = u_hat^2

7. type: gen y_hat_sq = y_hat^2
8. type: regress u_hat_sq y_hat y_hat_sq
9. Calculate $LM = nR^2$
10. Do we reject H_0 : homoskedasticity at 5% level of confidence?
    - **Now we think the fitted value in #8. is a reasonable candidate for** $\hat{h}_i$ So,
11. type: predict h_hat, xb
12. To check if $\hat{h}_i$ are all positive type: sum h_hat
13. type: gen sqrt_h = h_hat^0.5
14. type: gen wcolGPA = colGPA/sqrt_h
15. type: gen whsGPA = hsGPA/sqrt_h
16. type: gen wACT = ACT/sqrt_h
17. type: gen wskipped = skipped/sqrt_h
18. type: gen wPC = PC/sqrt_h
19. type: gen w = 1/sqrt_h
20. type: regress colGPA hsGPA ACT skipped PC
21. type: regress wcolGPA w whsGPA wACT wskipped wPC, nocon
22. type: regress wcolGPA w whsGPA wACT wskipped wPC, nocon robust

3 Labor Force Participation of Female

1. Download the "MORA.DTA" dataset from your EE425 Moodle page and open it in STATA.

   Choose File -> Open -> (then direct to the location of the file)
2. type: des
3. type: regress inlf nwifeinc educ exper expersq age kidslt6 kidsge6
4. type: estat hettest
5. type: regress inlf nwifeinc educ exper expersq age kidslt6 kidsge6, robust
6. type: predict y_hat, xb
7. type: twoway scatter inlf educ || line y_hat educ
8. type: sort educ

9. type: twoway scatter inlf educ || line y_hat educ

    - **We need to "hold other things constant". Suppose nwifeinc = 30, exper = 10, age = 35, kidslt6 = 0, kidsage6 = 0.**

10. type: gen y_new = 0.585 + 30*(-0.0034) + educ*(0.0379) + 10*(0.0395) + 100*(-0.0006) + 35*(-0.0161)

11. type: twoway scatter inlf educ || line y_new educ

Instrumental Variables Estimation and Two Stage Least Squares

Wooldridge Ch.15 section 15.1-15.4

1 Motivation

Assumption MLR4 or TS3 does not hold! So, $\hat{\beta}_{OLS}$ would be biased.

Because of various possible reasons. For example

1. Omitted Variable

2. Endogeneity or Simultaneity

From problem 1) Omitted Variable bias, if we have a good proxy variable for "ability", we can estimate

Let "z" be an instrumental variable, it has to satisfy two conditions

What might be correlated with *educ* but not *abil*?

Example: Wongsurawat (2004) Pornography and Social Ill

$$crime\_rate = \theta_0 + \theta_1 porn\_magazine + \theta_2 population + \theta_3 dangerous + e$$

- If we cannot find a measurement for "$dangerous$", then the error term becomes "$u = \theta_3 dangerous + e$".
- $Cov(porn\_magazine, u) \neq 0$
- Use $z =$ _____
- This z satisfies both requirements for a valid IV because ..

2 Consistency of the estimators

As long as the two requirements for a valid IV are satisfied, $\hat{\beta}$ would be consistent.

** Note that we don't call $\hat{\beta}$ under the IV method $\hat{\beta}_{OLS}$ anymore.

- $\hat{\beta}_{OLS}$ usually refers to $\hat{\beta}$ obtained from an estimation that does not employ IV to correct the problem.
- $\hat{\beta}_{IV}$ usually refers to $\hat{\beta}$ obtained from an estimation that employs IV to correct the problem.

To prove for consistency of $\hat{\beta}_{IV}$ we start with a simple regression function

$$y = \beta_0 + \beta_1 x + u$$

Inference

3 Properties of IV with a Poor Instrumental Variable

For $\hat{\beta}$ to be consistent, we require $plim\hat{\beta} = \beta$.

4 IV Estimation for the Multiple Regression Model

given

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

Let
y_1 = dependent variable of interest
y_2 = explanatory variable that is endogenous ($corr(y_2, u_1) \neq 0$)
z_1 = explanatory variable that is exogenous ($corr(z_1, u_1) = 0$)
z_2 = instrumental variable (also exogenous because $corr(z_2, u_1) = 0$)
We know that

5 Two-Stage Least Squares (2SLS)

Suppose now we have 1 more IV

$$z_3 = \text{cost of input}$$

$corr(z_3, u_1) = 0$

$corr(z_3, y_2) \neq 0$

We know that if $corr(z_1, y_2) \neq 0$, $corr(z_2, y_2) \neq 0$, $corr(z_3, y_2) \neq 0$, we can write:

The 2SLS method

Stage 1:

Stage 2:

STATA Command: