APPLIED ECONOMETRICS I NOTES 4

NOVEMBER 11, 2024

UNIVERSITY OF NOTTINGHAM
HAOLING WANG

Applied Econometrics notes 4

L11-L15

Table of Contents

Preface 3

L11 Functional form and interaction terms 4

The meaning of linear regression 4

The natural logarithm transformation 4

Note (S10) 5

Interpretation of coefficients 6

Example1 7

Example2-dependent variable (DV) is a percentage 8

Quadratic terms 9

Problem that should be noticed: multicollinearity in the model with quadratic term 10

Example 11

Functional form and model selection 12

Interaction terms 14

Q: Why do we need an interaction term? 14

Example1 14

Example2 15

L12 Incorporating qualitative information in regression models 17

Binary/Dummy variables ... 17
Incorporating binary information in regression models ... 18
Example1 – one category ... 18
Example2- one category with multiple constant coefficients ... 20
Example3- multiple categories ... 21
Example4-Interactions involving dummy variables ... 22
Example5-Interactions involving dummy variables ... 23
Example6- sample pooled across the two times, Interactions involving dummy variables ... 25
Lecture 13 over-specification and under-specification ... 26
Example ... 28
Lecture 15 The asymptotic properties of estimators ... 29
Consistency ... 30
Asymptotic distribution ... 31
The normality assumption ... 32
Lecture 16-17 Heteroscedasticity, Part I-II ... 33
Test for Homoscedasticity ... 35
The Breusch-Pagan test ... 35
The white test ... 37
Heteroscedasticity-consistent robust standard errors ... 39

Preface

This set of notes originates from my Applied Econometrics I module at UNNC. Rather than adding new material, I have reorganised the logical flow of the key points presented in the lecture slides. The module is closely linked to ECON1049 Mathematical Economics and Econometrics (SPR)—now renamed ECON1057 Introductory Econometrics—so many topics will already be familiar. If you find something related to statistics or econometrics, please go back to the lecture notes in ECON1049 (/ECON 1057). This will benefit to your memory and smooth your further study.

That said, the course should not be taken lightly. Its true value lies not in memorizing the slide-deck formulas, but in learning how to apply the statistical and econometric theory covered in Year 2 to real economic-data analysis. From a practical standpoint, I strongly recommend gaining proficiency in STATA, as intensive coding will return in Applied Econometrics II and even your final year dissertation.

Because I studied this module on exchange at UNUK, I also completed a piece of coursework. That assignment—requiring the direct application of PPT concepts, extensive STATA work, and a basic research framework—proved immensely beneficial. If you know classmates heading to UNUK, do not hesitate to ask for a look at their coursework. At UNNC, Applied Econometrics I is assessed solely by a final exam; therefore, if you are curious about econometrics or research methods, consider obtaining the UNUK coursework and experimenting with it yourself (perhaps over the summer, or even with GPT's guidance). Treat it as a scaffold for practicing real applications, but ultimately try to write the code on your own.

My guiding principle is to look beyond earning marks (though I know marks are important) for isolated knowledge points and instead ask whether each step you take builds the capabilities you will need next year or even in the future. Much like the Permanent Income Hypothesis in macroeconomics, cultivating a habit of extension and forward thinking will smooth your progression to the next stage, or if you learn repeated game in game theory, you will find out that the choices in each stage will be different depending on whether the game is one-stage or multi-stage (I think that sometimes, in real life, acting in player 1 or 2 in game theory is important, because they always have rational strategy).

This is the synthesized note 4, the last one. If you do not have the previous notes, please visiting the website: https://github.com/wang95483/notebook. And since this is just a synthesized lecture notes, its function is limited. So if you have any other ideas of how to learn this course, feel free to drop me an email first hmyhw8@nottingham.edu.cn (or if this doesn't work, use wang95483@gmail.com , my personal email).

Good luck with your Applied Econometric I study!

L11 Functional form and interaction terms

The meaning of linear regression

<table>
<tr><td>Linear in parameters</td><td>non-linearity in variables</td><td>Cannot accommodate in OLS regression models is relationships that are non-linear in parameters.</td></tr>
<tr><td>$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$
y is a linear function of x_1 and x_2</td><td>any non-linear function of other variables
Example
<table><tr><td>$x_1 = \ln(v)$</td><td>$x_1 = \sqrt{w}$</td><td>$x_2 = g^2$</td><td>$y = e^h$</td></tr></table>
Marginal effects can be allowed to change magnitude and even sign.

Two or more types of non-linearity can be accommodated within the same model.</td><td>$y = \beta_0 + \beta_1 x_1 + \beta_2^2 x_2 + \beta_1 \beta_2 x_3 + u$</td></tr>
</table>

The natural logarithm transformation

Definition

The natural logarithm transformation, denoted ln(y) or lny or log(y), is often used to introduce some non-linearity into regression models.

For small changes, the change in the natural log of a variable is approximately equal to the **proportional change** in that variable.

The formula below is a convenient way to calculate ε_y.

$$\varepsilon_y = \frac{\mathrm{d}\ln y}{\mathrm{d}\ln x}.$$

This is due to the chain rule and inverse function rule.

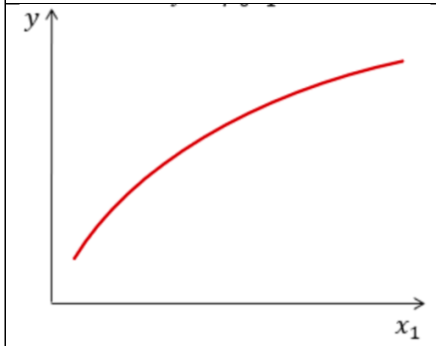
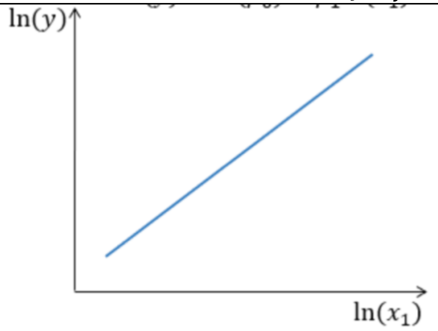
$$\frac{\mathrm{d}\ln y}{\mathrm{d}\ln x} \overset{\text{chain rule}}{=} \frac{\mathrm{d}\ln y}{\mathrm{d}y}\frac{\mathrm{d}y}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}\ln x} \overset{\text{inverse function rule}}{=} \frac{1}{y}\frac{\mathrm{d}y}{\mathrm{d}x}\frac{1}{\frac{\mathrm{d}\ln x}{\mathrm{d}x}} \overset{\text{ln rule}}{=} \frac{1}{y}\frac{\mathrm{d}y}{\mathrm{d}x}\frac{1}{\frac{1}{x}}.$$

$$\lim_{\Delta x \to 0}\left(\frac{\Delta y}{y}\right) \Big/ \left(\frac{\Delta x}{x}\right) = \lim_{\Delta x \to 0}\left(\frac{\Delta y}{\Delta x}\right) \Big/ \left(\frac{y}{x}\right) = \frac{\mathrm{d}y}{\mathrm{d}x} \Big/ \left(\frac{y}{x}\right) \equiv \varepsilon_y.$$

So it is particularly useful when we want to **measure relative (or percentage) changes.**

<u>Example</u>

$$y = \beta_0 {x_1}^{\beta_1}$$

<table>
<tr><td>the marginal effect (slope) of $x1$ varies with $x1$.
We cannot estimate this directly using OLS.</td><td>ln(y) = ln(β_0) + $\beta_1$ln(x_1)
linear and can be estimated using OLS.</td></tr>
<tr><td></td><td>$\beta_1 = \frac{\Delta \ln(y)}{\Delta \ln(x)} \approx \frac{\Delta y/y}{\Delta x/x}$ the elasticity of y with respect to x</td></tr>
<tr><td>
</td><td>
</td></tr>
</table>

Note (S10)

<table>
<tr><td rowspan="5">Applying logs allows marginal effects to change magnitude but not sign.</td><td>Not all variables can be logged</td><td>It is not always useful to log a variable</td><td colspan="2">percentage change vs percentage point change</td></tr>
<tr><td rowspan="4">Variables with observations that take zero or negative values.</td><td rowspan="4">Variables that are already measured in proportion or percent.</td><td>percentage change</td><td>percentage point change</td></tr>
<tr><td>a change relative to the initial value</td><td>a change in a percentage</td></tr>
<tr><td colspan="2"><u>Example</u>
If 24% of 18 year olds went to university in 2014 and 30% went in 2015.</td></tr>
<tr><td>(30-24)/24=25%</td><td>6 percentage points</td></tr>
</table>

Interpretation of coefficients

Model type	Equation	β_1	Interpretation
Linear (level-level)	$y = \beta_0 + \beta_1 x$	$\beta_1 = \frac{\Delta y}{\Delta x}$	**Linear,** β_1 is the change in y resulting from a unit change in x
Logarithmic (log-log)	$\ln(y) = \beta_0 + \beta_1 \ln(x)$	$\beta_1 = \frac{\Delta \ln(y)}{\Delta \ln(x)}$	**Elasticity,** β_1 is the % change in y that results from a 1% change in x
Semi-logarithmic (log-level)	$\ln(y) = \beta_0 + \beta_1 x$	$\beta_1 = \frac{\Delta \ln(y)}{\Delta x}$	$\beta_1 \times 100$ is the % change in y resulting from a unit change in x
Semi-logarithmic (level-log)	$y = \beta_0 + \beta_1 \ln(x)$	$\beta_1 = \frac{\Delta y}{\Delta \ln(x)}$	$\beta_1 \div 100$ is the change in y resulting from a 1% change in x

Example1

Some researchers were interested in the effect of air pollution on house prices, so, they formulated the following econometric model:

$$\ln price = \beta_0 + \beta_1 \ln(nox) + \beta_2 rooms + u$$

$price$	average house price in a neighbourhood
nox	amount of nitrogen oxide in the air in the neighbourhood
$rooms$	average number of rooms in houses within the neighbourhood

$\beta_1 = \frac{\partial ln(price)}{\partial ln(nox)} \approx \frac{\frac{\Delta price}{price}}{\frac{\Delta nox}{nox}}$	β_1 is the elasticity of $price$ w.r.t. air pollution	Ceteris paribus, a 1% increase in nitrogen oxide changes house prices by β_1% (holding house size fixed).
$\beta_2 = \frac{\partial \ln(\text{price})}{\partial \text{rooms}}$	$\beta 2(\times 100)$ is the semi-elasticity of $price$ w.r.t. house size	$\beta 2\times 100$ is the percentage change in $price$ that results from a house having one extra room ($\Delta rooms = 1$) (holding pollution fixed).

$$ln(\widehat{price}) = 9.23 - 0.718 \ln nox + 0.306\, rooms$$

Coefficient	Interpretation
$\beta_0 = 9.23$	In log regressions the constant term typically has no useful interpretation. The constant is the predicted value of $\ln(price)$ when $\ln(nox)$ and $rooms$ are zero
$\beta_1 =- 0.718$	When nox increases by 1%, price falls by ($\widehat{\beta_1}$) 0.718% (holding house size fixed). This is the house price elasticity wrt pollution.
$\beta_2 = 0.306$	When house size increases by one room, price increases by ($\widehat{\beta_2} \times 100$) 30.6%, (holding pollution fixed). This is the house price semi-elasticity wrt house size.

Example2-dependent variable (DV) is a percentage

$maths = \beta_0 + \beta_1 flunch + \beta_2 ptratio + \beta_3 expppup + \beta_4 avgtsal + \beta_5 enrol + \beta_6 grant + u$

$maths$	the **percentage** of pupils achieving a satisfactory mark in GCSE maths		
$flunch$	the **percentage** of pupils who are eligible to free school lunches	β_1	Ceteris paribus, **a percentage point** increase in the proportion of pupils eligible for free school lunches causes a β_1 **percentage point** increase in the proportion of pupils who achieve a satisfactory mark in GCSE maths.
$ptratio$	the number of pupils per teacher	$\beta_2 = \frac{\partial maths}{\partial \text{ptratio}}$	Ceteris paribus, one more pupil per teacher causes a β_2 **percentage point** increase in the proportion of pupils who achieve a satisfactory mark in GCSE maths.
$expppup$	expenditure per pupil per year, excluding teacher salaries, in thousands of pounds		
$avgtsal$	the average teacher's annual salary in thousands of pounds	$\beta_4 = \frac{\partial maths}{\partial \text{avgtsal}}$	Ceteris paribus, a £1,000 increase in the average teacher's salary causes a β_4 **percentage point** increase in the proportion of pupils who achieve a satisfactory mark in GCSE maths.
$enrol$	the total number of pupils in the school		
$grant$	grant income per year received from charities and private sponsors		
u	the error term		

Quadratic terms

Introducing quadratic terms allows:	
Marginal effects to change magnitude and sign.	Models to accommodate non-linearity in relationships involving variables with observations that **take zero or negative values.**

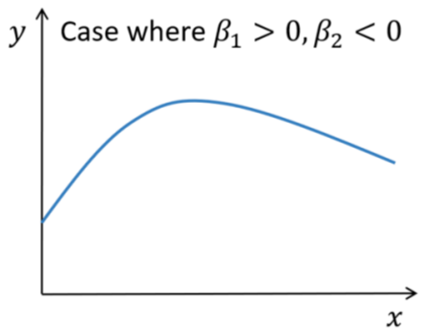
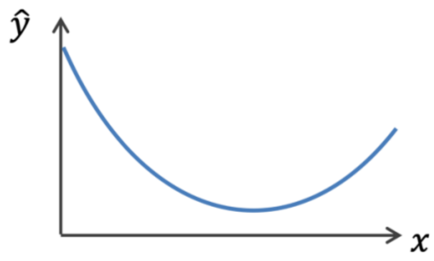
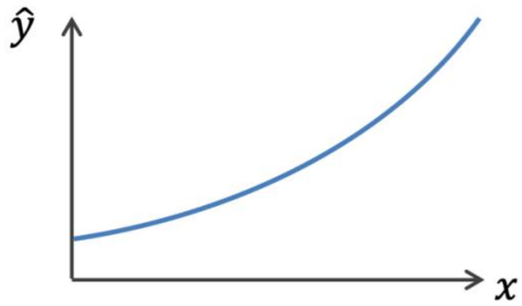
Model

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x + \widehat{\beta_2} x^2$$

Marginal effect (slope) of x on y now depends on the value of x:

$$\frac{\Delta \hat{y}}{\Delta x} = \widehat{\beta_1} + 2\widehat{\beta_2} x$$

$\widehat{\beta_1} > 0, \widehat{\beta_2} < 0$	$\widehat{\beta_1} < 0, \widehat{\beta_2} > 0$	$\widehat{\beta_1} > 0, \widehat{\beta_2} > 0$
		

Problem that should be noticed: multicollinearity in the model with quadratic term

For instance:

$Exper$ = number of years as a professional player in the NBA

$Expersq$ = the square of $exper$

$exper$ and $expersq$ are **not independent**:

$expersq$ is the square of $exper$, so they are mathematically related.

This relationship likely **introduces a high correlation** between these variables.

Multicollinearity can make the coefficients of $exper$ and $expersq$ unstable, leading to:

Large standard errors ($se(\hat{\beta})$) for their coefficients.

Difficulty in determining the individual contribution of each variable.

Example

In many empirical applications $\beta 1 > 0$ and $\beta 2 < 0$, implying an inverted u-shaped relationship.

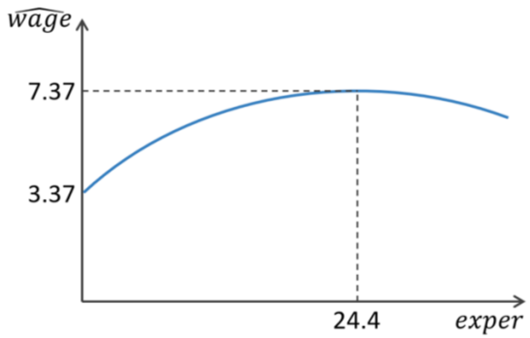
The relationship between wages and experience:

$$\widehat{wage} = 3.73 + 0.298 exper - 0.0061 exper^2$$
$$(0.35)\ (0.041) \qquad (0.0009) \qquad \text{(s.e.)}$$

The relationship is inverted u-shaped because the coefficient on $exper$ is positive and the coefficient on $exper^2$ is negative. The negative coefficient on the quadratic term captures the diminishing effect of experience on wages.

The marginal effect of each additional year of experience depends on experience: $\frac{\Delta \widehat{wage}}{\Delta exper} = 0.298 - (2 \times 0.0061 \times exper)$

$\widehat{wage}$ is at its maximum when:



$$exper^* = \left|\frac{\hat{\beta}_1}{2\hat{\beta}_2}\right| = \left|\frac{0.298}{2 \times 0.0061}\right| = 24.4 \text{ years}$$

$exper = 0$	The effect of an additional year of experience on wage is \$0.298.
$exper = 1$	The effect of an additional year of experience on wage is 0.298 − 2 × 0.0061 × 1 = \$0.286.
$exper = 10$	The effect of an additional year of experience on wage is 0.298 − 2 × 0.0061 × 10 = \$0.176.

Functional form and model selection

The linear model is nested within the quadratic model.

A t test of H_0: $\beta_2 = 0$, the model is linear, can be used to select the functional form.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

If $\beta_2 = 0$ then $y = \beta_0 + \beta_1 x + u$

<table>
<tr><td colspan="2">All other things equal, we would prefer the model that captures the non-linearity best</td></tr>
<tr><td>$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$</td><td>$y = \alpha_0 + \alpha_1 \ln(x) + e$</td></tr>
<tr><td colspan="2">Non-nested and do not contain any of the same explanatory variables.
Use $\overline{R^2}$ to select the preferred model since they have the same dependent variable.</td></tr>
</table>

Coefficient of correlation R^2

Definition
A measure of model performance or 'fit'
The ratio of the explained variation to total variation.
The fraction of the sample variation in y that is explained by $\boldsymbol{x}$.
A higher R^2 indicates a better 'fit' of the model to the data.

Properties

R^2 encompasses the **co-linearity of the regressors**; we cannot attribute this part of the variation in y to either x_1 alone or x_2 alone, but we can attribute it to them both and, hence, to the model.

R^2 can be used to compare rival models as long as they have the **same dependent variable.**

Formula

Adjustment $\overline{R^2}$

$$\overline{\boldsymbol{R^2}} = \mathbf{1} - \left[(\mathbf{1} - \boldsymbol{R^2}) \times \frac{\boldsymbol{n} - \mathbf{1}}{\boldsymbol{n} - \boldsymbol{k} - \mathbf{1}}\right]$$

Properties

(1) For any given model**,** $\overline{\boldsymbol{R^2}} < \boldsymbol{R^2}$**.**

(2) As $\boldsymbol{k}$ **increases**, the **size of the adjustment increases**.

(3) $\overline{\boldsymbol{R^2}}$ is often used to **compare** the fit of rival models with **different** $\boldsymbol{k}$**.**

Interaction terms

Definition

The effect of an explanatory variable, x_1, on the dependent variable, y, depends on the value of another explanatory variable x_2.

Q: Why do we need an interaction term?

A: **Marginal effect** of a factor is **variable conditional** on **another factor.**

Example1

Theory: in demand theory, consumer responsiveness to price (price elasticity) depends on income level.

(1) At low incomes, consumers are price conscious and ceteris paribus, exhibit price elastic demand (elasticity of demand is high).
(2) At high incomes, consumers are indifferent to prices, exhibit price inelastic demand (demand elasticity is low).

Model

$$\ln(q) = \beta_0 + \beta_1 \ln(p) + \beta_2 \ln(inc) + \beta_3 \ln(p) \times \ln(inc) + \mu$$

q	demand
p	price
inc	income
u	error or disturbance term

$\ln(p) \times \ln(inc)$ is an interaction term

$$\text{Price elasticity} = \frac{\Delta\log(q)}{\Delta\log(p)} = \beta_1 + \beta_3 \log(inc)$$

Example2

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

$price$	House price
$sqrft$	Floor area of house
$bdrms$	Number of bedrooms
$bthrms$	Number of bathrooms
u	Error or disturbance term

<table>
<tr><th>Terms</th><th>Relationship between Terms and Coefficient</th><th colspan="3">Interpretation</th></tr>
<tr><td rowspan="3">$bdrms$</td><td rowspan="3">$\frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 \mathrm{sqrft}$</td><td rowspan="3">the marginal effect of another bedroom is a function of house size</td><td rowspan="2">$\beta_3 = 0$</td><td>$\beta 2$ is now the constant marginal effect of number of bedroom in a house with floor area = 0</td></tr>
<tr><td>can be set up as a null hypothesis and a t test undertaken</td></tr>
<tr><td colspan="2">Effect of bedrooms on price must be evaluated at meaningful values of $sqrft$ such as the mean.</td></tr>
<tr><td rowspan="2">interaction term
$sqrft \times bdrms$</td><td rowspan="2">β_3</td><td rowspan="2">The marginal effect of the number of bedrooms ($bdrms$) on house price ($price$) depends on the size of the house ($sqrft$)
⇔ for interaction term $sqrft \times bdrms$</td><td colspan="2">$\beta_3 \neq 0$</td></tr>
<tr><td colspan="2">If $\beta_3 > 0$ an extra bedroom adds more to the value of a large house compared to a small house.</td></tr>
<tr><td>$bthrms$</td><td>$\beta 4 = \frac{\Delta price}{\Delta bthrms}$</td><td>the marginal effect of another bathroom</td><td colspan="2"></td></tr>
</table>

We ca rewrite the model as:

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

NOTE Many different types of **non-linear relationship** can be accommodated within the Classical Linear Model.

Assumption MLR1	linear in parameters
Assumption MLR2	random sampling
Assumption MLR3	no perfect collinearity (形容的是 variable 之间的，不是 variable 和 y 之间的)
Assumption MLR4	zero conditional mean
Assumption MLR5	homoscedasticity
Assumption MLR6	Normality of u

L12 Incorporating qualitative information in regression models

Binary/Dummy variables

Definition

When the qualitative information we are interested in takes an **either-or form**, i.e., each of our observations falls in **one of two categories**, then the qualitative information can be captured using a binary variable, i.e., a variable that equals either 0 or 1.

Binary variables are usually referred to as dummy variables, sometimes they are referred to as indicator variables.

Example

a power station is either coal fired or not coal fired

a university is either in the Russell Group or not

a person either owns a smart phone or does not own a smart phone

a country either has a coast or does not have a coast

Incorporating binary information in regression models

Example1 – one category

Question The gender wage gap and, specifically, in **whether male and female wages differ** even when education is held constant.

Model

$$wage = \beta_0 + \boldsymbol{\delta_1 female} + \beta_1 educ + u$$

$wage$	hourly wage in $
$female$	= 1 if individual female
	= 0 if individual male male is our base group/benchmark group/basis for comparison
$educ$	education in years
u	error or disturbance term

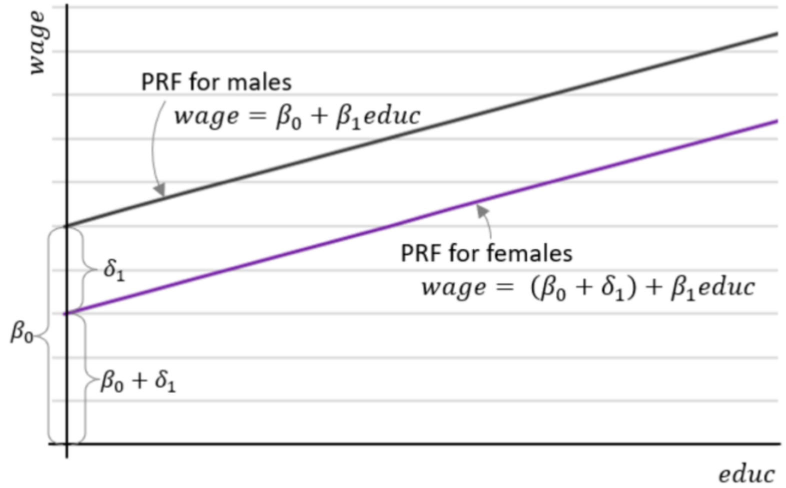
NOTE Two dummies $female$ and $male$ would be **redundant** and are **perfectly colinear**, so a model including both violates MLR3: No perfect collinearity. Including both is referred to as falling into the **dummy variable trap.**

要想好怎么设置 dummy variable 就把所有的 categories 全部算进去。

male	$female$ = 0	$wage = \beta_0 + \beta_1 educ + u$
female	$female$ = 1	$wage = (\boldsymbol{\beta_0 + \delta_1}) + \beta_1 educ + u$

$\boldsymbol{\delta_1 = E\,(wage|female = 1, educ) - E\,(wage|female = 0, educ)}$

If δ_1<0



wage
PRF for males
$wage = \beta_0 + \beta_1 educ$
δ_1
β_0
$\beta_0 + \delta_1$
PRF for females
$wage = (\beta_0 + \delta_1) + \beta_1 educ$
educ

Example2- one category with multiple constant coefficients

Question Do male and female wages differ even when education, experience, and tenure with current employer are held constant?

Model1

$$wage = \beta_0 + \boxed{\delta_1 \ female} + \beta_1 \ educ + \beta_2 \ exper + \beta_3 \ tenure + u$$

$$\underset{(.72)}{wage = -1.57} - \underset{(.26)}{1.81} female + \underset{(.049)}{.572} educ + \underset{(.012)}{.025} exper + \underset{(.021)}{.141} tenure \quad s.e.$$

$$n = 526, \qquad R^2 = .364$$

Do male and female wages differ?

$$wage = \beta_0 + \delta_1 \ female + e$$

$$wage = \underset{(.72)}{7.10} - \underset{(.26)}{2.51} female \quad s.e.$$

$$n = 526, \qquad R^2 = .116$$

Model2 Apply the **natural log transformation** to our dependent variable, our interpretation of δ_1, the coefficient on $female$, has to change accordingly.

$$\ln(wage) = \beta_0 + \delta_1 \ female + \beta_1 \ educ + \beta_2 \ exper + \beta_3 \ tenure + \varepsilon$$

$$\ln(wage) = .50 - .301 female + .087 educ + .005 exper + .017 tenure$$

δ_1 = -0.301 Ceteris paribus, women earn approximately 30% less than men per hour.

Example3- multiple categories

Question The effect of gender on earnings, we are interested in the effect of marriage on earnings and whether that effect differs between women and men.

Model

$$ln(wage) = \beta_0 + \boldsymbol{\delta_1 marmale + \delta_2 singfem + \delta_3 marfem} + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon$$

Category

	Male	Female
Married	$marmale$ = 1	$marfem$ = 1
Single	$marmale = singfem = marfem$ =0	$singfem$ = 1

$marmale$ = 1	the individual is male and married	$ln(wage) = \beta_0 + \boldsymbol{\delta_1 marmale} + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon$
$singfem$ = 1	the individual is female and unmarried	$ln(wage) = \beta_0 + \boldsymbol{\delta_2 singfem} + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon$
$marfem$ = 1	the individual is female and married	$ln(wage) = \beta_0 + \boldsymbol{\delta_3 marfem} + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon$
$marmale = singfem = marfem$ =0	$single\ male$	$ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon$

$$\ln(wage) = .388 + \mathbf{.292} marmale - \mathbf{.097} singfem - \mathbf{.120} marfem + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon$$

$marmale$ = 1	the individual is male and married	$\ln(wage) = .388 + \mathbf{.292} marmale + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon$	Ceteris paribus, married men earn approximately **29% more than single men.**
$singfem$ = 1	the individual is female and unmarried	$\ln(wage) = .388 - \mathbf{.097} singfem + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon$	Ceteris paribus, single women earn approximately **10% less than single men.**
$marfem$ = 1	the individual is female and married	$\ln(wage) = .388 - \mathbf{.120} marfem + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon$	Ceteris paribus, married women earn approximately **12% less than single men.**
$singfem$ = 1 $marfem$ = 1	the individual is female and unmarried the individual is female and married	$\ln(wage) = .388 - \mathbf{.097} singfem - \mathbf{.120} marfem + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \epsilon$	Ceteris paribus, **married women** earn approximately **2% less than single women.** -.120-(-.097)=-.023

Example4-Interactions involving dummy variables

Model

$$\ln(wage) = \beta_0 + \delta_1 female + \delta_2 married + \delta_3 female * married + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \varepsilon$$

Category

	Male	Female
Married	$married$=1	$female$=1, $female * married$ = 1, $married$=1
Single	$female = married = female * married$ = 0	$female$=1

$married$=1	the individual is male and married	$\ln(wage) = \beta_0 + \delta_2 married + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \varepsilon$
$female$=1, $female * married$ = 1, $married$=1	the individual is female and married	$\ln(wage) = \beta_0 + \delta_1 female + \delta_2 married + \delta_3 female * married + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \varepsilon$
$female$=1	the individual is female and single	$\ln(wage) = \beta_0 + \delta_1 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \varepsilon$
$marmale = singfem = marfem$ =0	Single male	$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \varepsilon$

$$\ln(wage) = .388 - .097 female + .292 married - .316 female * married + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \varepsilon$$

$married$=1	the individual is male and married	$\ln(wage) = .388 + .292 married + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \varepsilon$	Ceteris paribus, married men earn approximately **29%** more than single men.
$female$=1	the individual is female and single	$\ln(wage) = .388 - .097 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \varepsilon$	Ceteris paribus, single women earn approximately **10%** less than single men.
$female$=1, $female * married$ = 1, $married$=1	the individual is female and married	$wage) = .388 + .292 married - .316 female * married + \beta 1 educ + \beta 2 exper + \beta 3 tenure + \varepsilon$	Ceteris paribus, married women earn approximately 2% less than single women. -.316+.292 $female * married$ 中排除 married 的影响
$female$=1	the individual is female and single		

NOTE Using interactions involving dummy variables, we can also investigate differences in the **slopes** of relationships between categories.

Example5-Interactions involving dummy variables

Model

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \boldsymbol{\delta_0 female + \delta_1 female * educ + \delta_2 female * exper + \delta_3 female * tenure} + \varepsilon$$

Male	$female = 0$	$wage = \beta 0 + \beta 1 educ + \beta 2 exper + \beta 3 tenure + \varepsilon$
Female	$female = 1$	$wage = \boldsymbol{(\beta_0 + \delta_0) + (\beta_1 + \delta_1) educ + (\beta_2 + \delta_2) exper + (\beta_3 + \delta_3) tenure} + \varepsilon$

Question1 Whether the marginal effects of **education, experience and tenure** are the same for married men and married women

$$wage = -2.81 + .69 educ + .03 exper + .17 tenure + \boldsymbol{3.38 female - .36 female * educ - .03 female * exper - .17 female * tenure}$$

Male	$female = 0$	$wage = -2.81 + .69 educ + .03 exper + .17 tenure$	for married men, one more year of **education** increases the hourly wage by 69 cents
			for married men, one more year of **tenure** increases the hourly wage by 17 cents
Female	$female = 1$	$wage = (\mathbf{3.38} - 2.81) + (.69 \mathbf{-.36})\ educ + (.03 \mathbf{-.03})\ exper + (.17 \mathbf{-.17})\ tenure$	for married women, one more year of **education** increases the hourly wage by 33 cents+
			for married women, one more year of **tenure** does not increase the hourly wage

Question2 Is the regression function for wages the same for men and for women or is it different for men and for women?

Null Hypothesis

$$H_0: \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$$

***F* test**

Restricted model

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \varepsilon$$

Unrestricted model

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \boldsymbol{\delta_0 female + \delta_1 female * educ + \delta_2 female * exper + \delta_3 female * tenure} + \varepsilon$$

→ Chow test

Definition

A statistical test used to determine whether the coefficients in two different linear regressions on different data sets are equal.

[Comparing Groups] To test if different groups (like males vs. females, different regions, or different economic periods) have the same regression relationship between dependent and independent variables.

Example6- sample pooled across the two times, Interactions involving dummy variables

Question a country has **introduced new legislation** aimed at reducing the gender wage gap, investigating whether the legislation has been effective. 时间点不同，进行前后效果反差的对比

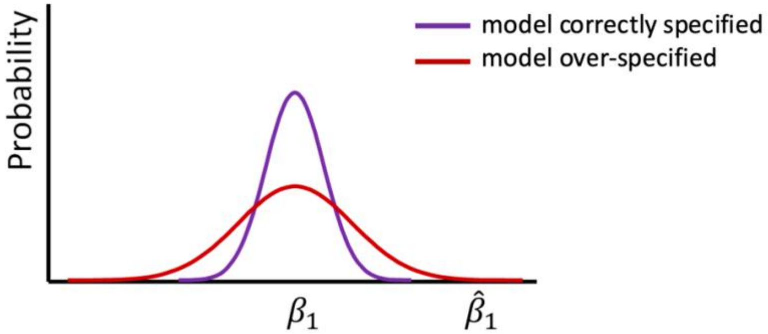
Model

$$wage = \beta_0 + \boldsymbol{\delta_0 Y_2 + \delta_1 female + \delta_2 Y_2 * female} + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

Y_2=0	Sample 1 observations	collected a month before the legislation took effect	$wage = \beta_0 + \boldsymbol{\delta_1 female} + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$	
Y_2=1	Sample 2 observations	collected 12 months later	$wage = \boldsymbol{(\beta_0 + \delta_0) + (\delta_1 + \delta_2) female} + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$	ceteris paribus, $\delta_1 + \delta_2$ = the gender wage gap after the change in legislation
				δ_2 = the change in the, ceteris paribus, gender wage gap owing to the change in legislation

Lecture 13 over-specification and under-specification

Mis-specified → OLS estimators may not be unbased and efficient.

<table>
<tr><td></td><td colspan="2">Over-specification</td><td colspan="2">Under-specification</td></tr>
<tr><td>Definition</td><td colspan="2">Irrelevant regressors included in the model</td><td colspan="2">Important regressors omitted from the model</td></tr>
<tr><td>True model</td><td colspan="2">$y = \beta_0 + \beta_1 x_1 + u$</td><td colspan="2">$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$</td></tr>
<tr><td rowspan="2">Our Model</td><td colspan="2">$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$</td><td colspan="2">$y = \beta_0 + \beta_1 x_1 + u$</td></tr>
<tr><td colspan="2">$y = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2$</td><td colspan="2">$y = \widehat{\beta_0} + \widehat{\beta_1} x_1$</td></tr>
<tr><td>OLS estimators</td><td colspan="2">Unbiased
$$E[\widehat{\beta_1}] = \beta_1$$
$$E[\beta_2] = \beta_2 = 0$$
But inefficient
→ OLS is more likely to produce estimates that are imprecise.

</td><td colspan="2">(I) Biased if $x1$ and $x2$ are correlated.
$$E[\widehat{\beta_1}] = \beta_1 + \beta_2 \frac{cov(x_1, x_2)}{var(x_1)}$$
Omitted variable bias = $\beta_2 \frac{cov(x_1,x_2)}{var(x_1)}$

(II) Unbiased if $x1$ and $x2$ are uncorrelated.
$E[\widehat{\beta_1}] = \beta_1$ since $cov(x_1, x_2) = 0$

Sign of the bias $\beta_2 \frac{cov(x_1,x_2)}{var(x_1)}$
(1) Sign of $\boldsymbol{\beta_2}$, the effect of x_2 on y in the population model (correct specification).
(2) $\boldsymbol{cov(x_1, x_2)}$ Sign of correlation between x_1 and x_2.
<table><tr><td></td><td>x_1 and x_2 positively correlated</td><td>x_1 and x_2 negatively correlated</td></tr><tr><td>$\beta_2 > 0$</td><td>Positive bias</td><td>Negative bias</td></tr><tr><td>$\beta_2 < 0$</td><td>Negative bias</td><td>Positive bias</td></tr></table></td></tr>
<tr><td></td><td>The correct specification</td><td>Over-specification</td><td>The correct specification</td><td>Under-specification</td></tr>
</table>

<table>
<tr>
<td>Standard error
$se(\widehat{\beta_1})$</td>
<td>$$se(\widehat{\beta_1}) = \sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(x_{i1} - \overline{x_1})^2}}$$
If there is no correlation between x_1 and x_2 ($R_1^2 = 0$), ‘our’ standard errors will be the same as had we correctly specified the model.</td>
<td>$$se(\widehat{\beta_1}) = \sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(x_{i1} - \overline{x_1})^2(1 - R_1^2)}}$$
R_1^2: the R^2 of the regression of x_1 on x_2.
If **x_1 and x_2 are correlated**, ‘our’ standard errors will be **larger.**
→ Over-specification inflates standard errors.
↔ over-specification renders the OLS estimators **inefficient**</td>
<td>$$se(\widehat{\beta_1}) = \sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(x_{i1} - \overline{x_1})^2(1 - R_1^2)}}$$</td>
<td>$$se(\widehat{\beta_1}) = \sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(x_{i1} - \overline{x_1})^2}}$$
Our $\boldsymbol{\widehat{\sigma^2}}$ will be a **biased estimate** of σ^2 because of the under-specification.
$$\widehat{\sigma^2} = Var(\hat{u}_i) = \frac{\sum_{i=1}^{n}\widehat{u_i}^2}{n - (k+1)}$$
No matter x_1 and x_2 are correlated or not, when under-specification, $\widehat{\sigma^2}$ is always biased.</td>
</tr>
<tr>
<td>t ratios</td>
<td colspan="2">Smaller (because $se(\widehat{\beta_1})$ is larger)
t-ratio = $\frac{\widehat{\beta_j}}{se(\widehat{\beta_j})}$
→ power is reduced
→ more likely to make Type II error
Power of a test = 1 − $Prob(Type\ II)$
↔ more likely to fail to reject false nulls</td>
<td colspan="2"></td>
</tr>
<tr>
<td>Explanation</td>
<td colspan="2">Over-specification has reduced the amount of information that can be used to estimate β_1 because x_1 and x_2 are correlated.</td>
<td colspan="2">Part of the influence of x_2 on y is being attributed to x_1.</td>
</tr>
<tr>
<td>Extent of the problems depends on the **correlations between variables**</td>
<td colspan="2">relevant and irrelevant</td>
<td colspan="2">included and omitted</td>
</tr>
<tr>
<td>Conclusion</td>
<td colspan="2">Estimation and hypothesis testing are still valid, but Type II errors more likely.</td>
<td colspan="2">Test statistics will be incorrect and our inferences will be wrong. Omission of important variables is generally more problematic than inclusion of irrelevant ones.</td>
</tr>
</table>

Example

Model

$$\widehat{bwght} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \boldsymbol{faminc} + \boldsymbol{\beta_2} \boldsymbol{order} + \boldsymbol{\beta_3} \boldsymbol{cigs} + \boldsymbol{\beta_4} \boldsymbol{fatheduc}$$

After running the regression:

$$\begin{array}{llllllll} \widehat{bwght} = & 112.1 & + 0.048 faminc & + 1.854 order & - 0.580 cigs & + 0.298 fatheduc & \\ & (35.90) & (1.34) & (2.82) & (-5.29) & (1.27) & (t\ ratio) \\ & [0.000] & [0.180] & [0.005] & [0.000] & [0.205] & [\text{p-value}] \end{array}$$

The coefficient on father's education ($fatheduc$) is statistically insignificant.

→ Dropping $fatheduc$ and re-estimating yields:

$$\begin{array}{llllll} \widehat{bwght} = & 114.2 & + 0.098 faminc & + 1.61 order & - 0.477 cigs & \\ & (77.73) & (3.35) & (2.68) & (-5.21) & (t\ ratio) \\ & [0.000] & [0.001] & [0.008] & [0.000] & [\text{p-value}] \end{array}$$

Dropping $fatheduc$ induces positive bias in the coefficient on family income ($faminc$)

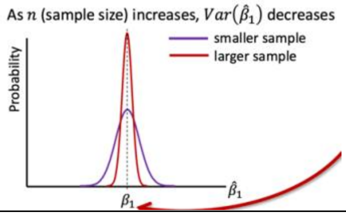
$cov(\text{fatheduc}, \text{faminc}) > 0$: $fatheduc$ and $faminc$ are positively correlated

$\boldsymbol{\beta_4 > 0}$: coefficient on father education positive

The correlation in family income and father's education is also giving rise to **multicollinearity** (low t ratios).

t-ratio = $\dfrac{\widehat{\beta_J}}{se(\widehat{\beta_J})} = \dfrac{\widehat{\beta_J}}{\sqrt{\dfrac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(x_{i1}-\overline{x_1})^2(1-R_1^2)}}} = \dfrac{\widehat{\beta_J}\sqrt{\sum_{i=1}^{n}(x_{i1}-\overline{x_1})^2(1-R_1^2)}}{\widehat{\sigma}}$

Lecture 15 The asymptotic properties of estimators

Asymptotic Properties			
Important when	**Properties**	**Explanation**	**Example**
the **finite** sample properties of estimators **cannot be determined**	unbiasedness		OLS estimator
	Efficiency		
the **assumptions** of the CLM **do not hold**	Consistency	A consistent estimator of a regression coefficient, $\widehat{\beta_k}$, has a sampling distribution **that converges on the true value** as **sample grows**. Captures the idea of <u>increasing precision</u> with <u>more information.</u> SLR5: homoscedasticity is not required to hold this. But SLR2, SLR3, and SLR4 must be valid.  	
	Asymptotically efficient		

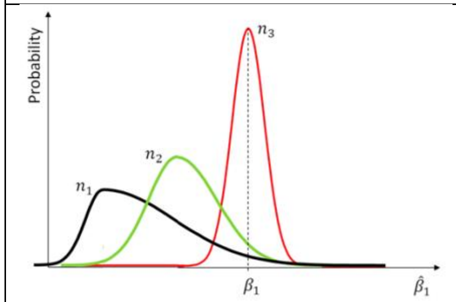
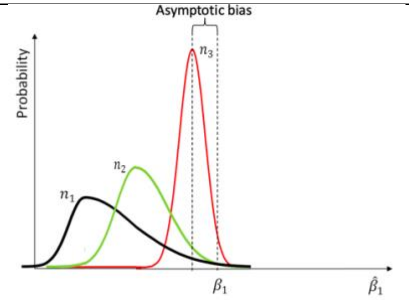
If all of the CLM assumptions are satisfied, then we can get an unbiased, efficient, and consistent estimator.

Consistency

<table>
<tr><td colspan="2">MLR4: zero conditional mean $E\,(u|x_1, x_2, x_3, \ldots, x_k)$ = 0</td></tr>
<tr><td>Unbiased
u is uncorrelated with:</td><td>Consistency
u is uncorrelated with:</td></tr>
<tr><td colspan="2">each and every explanatory variable</td></tr>
<tr><td>with any possible function of the explanatory variables</td><td></td></tr>
</table>

→ An estimator can be **biased** but **consistent:**

When the sample is small, the estimator is biased, but as the sample grows, the bias declines, and asymptotically, the bias disappears.

<table>
<tr><td>A biased but consistent estimator</td><td>A biased and inconsistent estimator</td></tr>
<tr><td colspan="2">Sample size n: $n1 < n2 \ll n3$, $n3 \to \infty$</td></tr>
<tr><td>
</td><td>
</td></tr>
<tr><td>When the sample size is small, n_1, the estimator is biased (and imprecise and skewed).</td><td rowspan="3">As the sample grows towards infinity, the estimator, $\widehat{\beta_1}$, fails to converge on the true parameter value, β_1.</td></tr>
<tr><td>When the sample size increases to n_2, the estimator is less biased (more precise and less skewed).</td></tr>
<tr><td>When the sample size is very large, $n_3 \to \infty$, the estimator is unbiased (much more precise and not skewed).</td></tr>
</table>

Asymptotic distribution

If we cannot have unbiased and efficient estimators, we seek estimators that are consistent and asymptotically efficient. i.e.

When an estimator with desirable finite sample properties cannot be found, we base our choice of estimator on **large sample properties.**

Consistency and asymptotic efficiency are the large sample counterparts to unbiasedness and minimum variance.

Definition

The asymptotic distribution of an estimator is the sampling distribution of that estimator when **the sample size is infinitely large.**

Theorem

If the asymptotic distribution of an estimator becomes **concentrated on a particular value k**, k is referred to as **the probability limit of $\widehat{\beta}_j$:**

$$plim\ \widehat{\beta}_j = k$$

If $plim\ \widehat{\beta}_j = \beta_j$, then $\widehat{\beta}_j$ is consistent.

If the **variance** of the asymptotic distribution of $\widehat{\beta}_j$ is **smaller** than any other consistent estimator, then $\widehat{\beta}_j$ is said to be **asymptotically efficient**.

***NOTE** **If MLR5, homoscedasticity, is not valid**, our **inferences will not be valid regardless of whether MLR6 is valid or not** and regardless of whether the sample is small or large.

The normality assumption

Theorem **Central Limit Theorem**

Irrespective of the distribution of the parent population, the distribution of sample averages approaches the **normal** as sample size grows.

When the data are non-normal the sampling distributions of OLS estimators are **asymptotically normal**, i.e., normal when the sample is large.

→ OLS estimators are (effectively) **sample averages.**

If the population is **symmetric**, sample **statistics are normal for $n \geq 30$.**

By n = 200 normality of sample statistics is assured.

Model errors or disturbance terms are described as **homoscedastic** if the error variance is the **same** for all the observations in the sample.
Model errors or disturbance terms are described as **heteroscedastic** if the error variance is **not the same** for all observations.

Definition **Heteroscedastic**

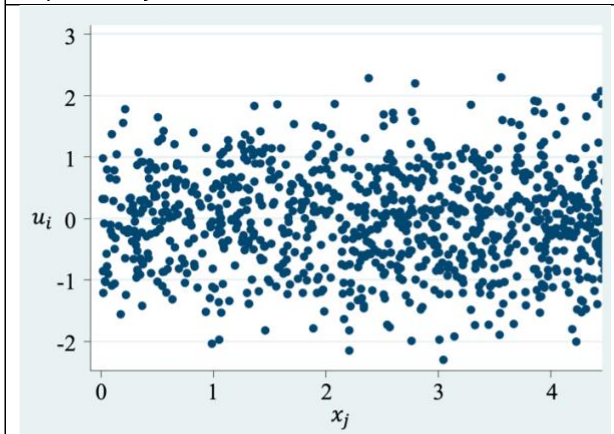
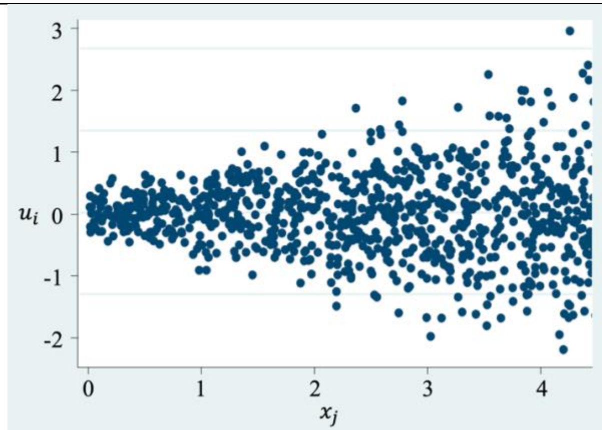
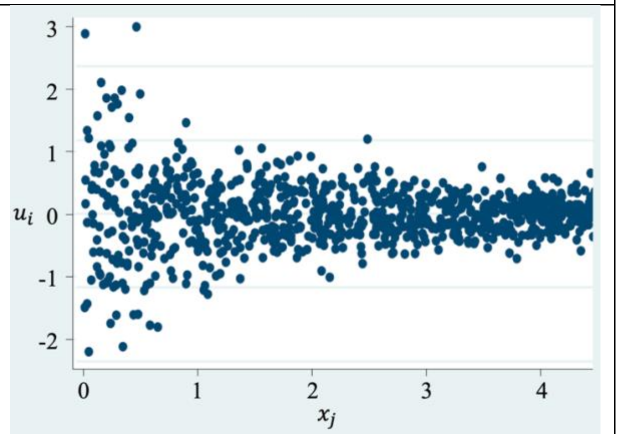
Gauss-Markov assumption MLR5: Homoscedasticity is not hold.

For the general multiple linear regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i$$

If $Var(u_i) = \sigma_i^2$ (note: i subscript on σ^2), **the variance is not the same** for all i, the errors are **heteroscedastic**, the variance may vary systematically with one or more of the explanatory variables, $x_1, x_2, x_3, \ldots x_k$.

If there is heteroscedasticity but we calculate s.e.s assuming homoscedasticity and use these when conducting hypothesis tests, our **inferences will be wrong.** When heteroscedasiticy is present the s.e.s have to be calculated in a way that accounts for that heteroscedasticity.

Homoscedastic errors		
The spread of the errors is **unrelated** to the explanatory variables, $x_1, x_2, x_3, \ldots x_k$	The spread of the errors **increasing** with x_j	The spread of the errors **decreasing** with x_j
		

Result

When there is Homoscedastic errors, as long as **the other Gauss-Markov assumptions are valid**,

(1) **OLS** still yields **unbiased** estimates of the regression parameters, β_0, β_1, β_2, β_3... β_k, but they are **inefficient.** i.e., they are not BEST.

    In the presence of heteroscedasticity, GLS is BEST.

(2) The standard errors of the OLS estimators $\boldsymbol{se(\widehat{\beta_j})} = \sqrt{\frac{\widehat{\sigma^2}}{\sum_{i=1}^{n}(x_{ji}-\overline{x_j})^2(1-R_j^2)}}$ is incorrect.

(3) Error in t-statistic, inferences could be wrong.

(4) Inferences drawn from F-tests could be wrong.

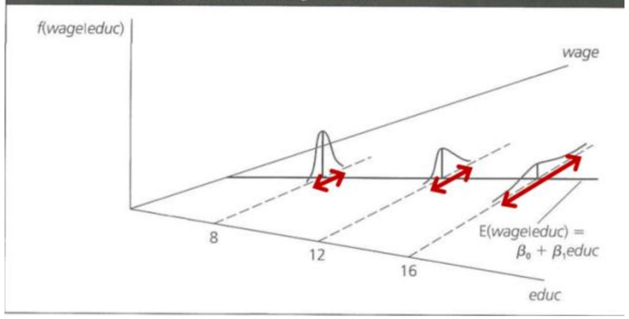
Example

In model:

$$wage = \beta_0 + \beta_1 education + u$$

people with more education have access to a **wider variety of jobs→the variability of wages is likely to increase with education**

FIGURE 2.9 Var(*wage*|*educ*) increasing with *educ*.



Test for Homoscedasticity

The Breusch-Pagan test

Hypothesis	Null hypothesis	Alternative hypothesis
	H_0: the errors are homoscedastic $Var\,(u_i \mid \boldsymbol{x}_i) = \sigma^2$	H_1: there is heteroscedasticity of a **specific** form
Step 1	Estimate the model using OLS, generate the residuals, $\widehat{u_i}$, compute the squared residuals, $\widehat{u_i}^2$.	
Step 2	Decide on the variable or list of variables, $\boldsymbol{Z}$, which you suspect are responsible for the heteroscedasticity, let p = the number of variables in $\boldsymbol{Z}$	
Step 3	Choose a significance level, say 5%, find the critical value in the χ^2_p distribution.	
Step 4	Using OLS (again), estimate **a regression model** in which the dependent variable is $\widehat{u_i}^2$ and the list of explanatory variables is $\boldsymbol{Z}$, include a constant in the model. $$Z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}$$ $$\widehat{u_i^2} = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_n z_n$$	
Step 5: Test statistic	the **Breusch-Pagan LM statistic** $LM = nR^2 \sim \chi^2_p$ distribution (R^2 is the coefficient of determination for the regression in Step 4; n = sample size)	
Step 6	Reject the null (homoscedasticity) if LM > critical value	

Example **Test for Homoscedasticity**

Research question: Are CEOs' salaries determined by profits?

Regression model:

$$\ln(CEOpay)_i = \beta_0 + \beta_1 \ln(assets)_i + \beta_2 profit_i + u_i$$

where

i indexes the observation and an observation is a company

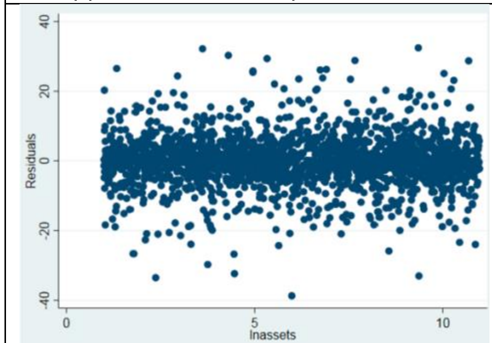
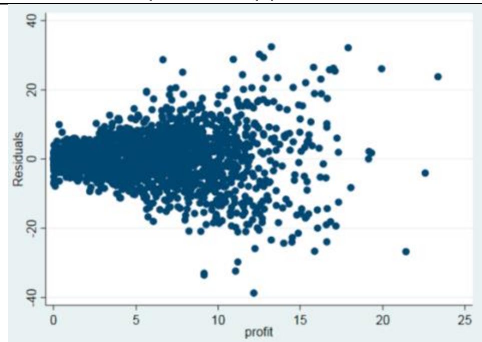
$CEOpay$ = the annual salary of the CEO

$assets$ = total value of company's assets

$profit = \left(\frac{pretax\ profits}{sales}\right) * 100$

u_i = error or disturbance term

Null hypothesis: H_0: $\beta_2 = 0$; Alternative hypothesis: H_1: $\beta_2 \neq 0$

<table>
<tr><td colspan="2">Detecting heteroscedasticity using graph</td></tr>
<tr><td>No apparent relationship between residual dispersion and ln($assets$).</td><td>Residual dispersion appears to increase as $profit$ increases.</td></tr>
<tr><td>
</td><td>
</td></tr>
</table>

The white test

```
imtest, white
```

Test for heteroscedasiticy (put another way – test for the validity of the homoscedasticity assumption) and, if there is heteroscedasticity, adjust the s.e.s to account for it.

The `imtest, white` command is asking STATA to conduct the test. More specifically, it is asking for a White's test for heteroscedasticity of a **general form.**

Null hypothesis: homoscedasticity,

Alternative hypothesis: heteroscedasticity of a general form.

The test statistic has a **Chi-squared distribution** with **degrees of freedom** that depend on the **number of variables in the model** and **the nature of those variables.**

```
regress crime popd taxpc pctue west police parr pconv parrw pconvw, vce(hc3)
```

The, `vce(hc3)` added to the end of the `regress` command is asking STATA to **adjust the standard errors to account for heteroscedasticity**.

Hypothesis	Null hypothesis	Alternative hypothesis
	The errors are homoscedastic, $Var\,(u_i\|\boldsymbol{x}_i) = \sigma^2$	There is heteroscedasticity of a **general** form
Step 1	Estimate the model using OLS, generate the residuals, $\widehat{u_i}$, compute the squared residuals, $\widehat{u_i}^2$	
Step 2	Generate the squares and the cross-products of all the regressors (EVs) in the model. So, if the model contains two regressors, x_1 and x_2, we generate x_1^2, x_2^2, and x_1x_2. Set $Z = [x_1, x_2, x_1x_2, x_1^2, x_2^2]. So, p = 5$	
Step 3	Choose a significance level, say 5%, find the critical value in the $\chi_p{}^2$ distribution, i.e., the chi-squared distribution with p degrees of freedom.	
Step 4	Using OLS (again), estimate a regression model in which the dependent variable is $\widehat{u_i}^2$ and the list of explanatory variables is Z, include a constant in the model. $Z = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}$ $\widehat{u_i}^2 = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_n z_n$	

Step 5 Test statistic	the **White statistic**, which has a χ_p^2 distribution, i.e., a chi-squared distribution with p degrees of freedom. Calculate the White statistic: $LM = nR^2$ where R^2 is the coefficient of determination for the regression in Step 4 and n = sample size
Step 6	Reject the null (homoscedasticity) if White statistic > critical value

Heteroscedasticity-consistent robust standard errors

Definition

Standard errors that have been adjusted to account for heteroscedasticity, that can be used to draw valid inferences in the presence of heteroscedasticity.

When the errors or disturbance terms are **heteroscedastic**, **OLS estimators are no longer best**, i.e., they are inefficient.

Best estimators: Weighted Least Squares Estimation

Formula

$$\widehat{Var\left(\widehat{\beta}_1\right)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \widehat{u_i^2}}{SST_x^2}$$

Derivation

$$\begin{aligned}
\hat{\beta}_1 &= \beta_1 + \frac{Cov(x,u)}{Var(x)} \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \overset{=0}{\bar{u}})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}$$

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

<table>
<tr><td colspan="2">SLR model
$y_i = \beta_0 + \beta_1 x_i + u_i$</td><td colspan="2">MLR model
$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_k x_k + u_i$</td></tr>
<tr><td>If the errors are homoscedastic (SLR5 valid), then for the SLR model $y_i = \beta_0 + \beta_1 x_i + u_i$ with SLR 1-4 valid, then $\sigma_i^2 = \sigma^2$</td><td>In the presence of heteroscedasticity (SLR5 invalid), for the SLR model $y_i = \beta_0 + \beta_1 x_i + u_i$ with SLR 1-4 valid</td><td>Under MLR 1-4, in the presence of heteroscedasticity, it can be show that the following is a valid estimator</td><td>Heteroscedasticity-consistent robust standard error for $\widehat{\beta_j}$ is</td></tr>
<tr><td rowspan="2">$Var(\widehat{\beta_1}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}$</td><td rowspan="2">$Var(\widehat{\beta_1}) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2}{SST_x^2}$
So it can be estimated as:
$\widehat{Var(\widehat{\beta_1})} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \widehat{u_i}^2}{SST_x^2}$</td><td>$\widehat{Var(\widehat{\beta_j})} = \frac{\sum_{i=1}^n \widehat{r_{ij}^2}\,\widehat{u_i^2}}{SSR_j^2}$</td><td>$se(\widehat{\beta_j}) = \sqrt{\frac{\sum_{i=1}^n \widehat{r_{ij}^2}\,\widehat{u_i^2}}{SSR_j^2}}$</td></tr>
<tr><td colspan="2">• $\widehat{r_{ij}^2}$ is the residual for the i^{th} observation when x_j is regressed on all the other explanatory variables
• SSR_j is the sum of the squared residuals from that regression</td></tr>
</table>