

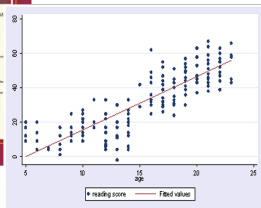
# Introduction to Quantitative Methods for Development

Renate Hartwig, PhD

June 16, 2016

# What this course is all about

- Introduce you to the basic principles of regression analysis.
- Apply theoretical concepts to 'real life' examples using Stata.
- Give you tools to understand, appraise and conduct studies on development issues.



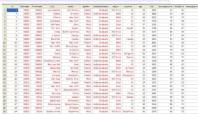
```
. logit union age hours ttl_exp topage logtenure  
Iteration 0: log likelihood = -3826.4263  
Iteration 1: log likelihood = -977.12718  
Iteration 2: log likelihood = -976.1985  
Iteration 3: log likelihood = -976.1985  
Iteration 4: log likelihood = -976.1985
```

Logistic regression

Log likelihood = -976.1985

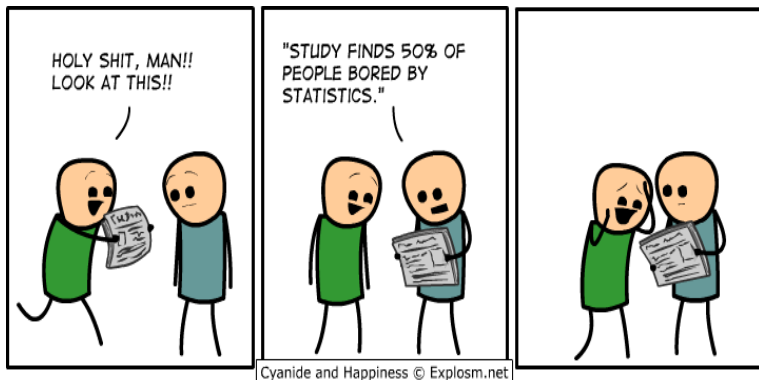
	coef.	std. err.	z	p> z	[95% Conf. Interval]
age	.022564	.0585345	0.68	0.497	-.0227566 .0696033
hours	-.0889733	.0061722	-3.45	0.001	-.0951238 -.0828229
ttl_exp	-.055001	.0563838	-1.40	0.001	-.0967208 -.0132813
topage	.0637068	.1250053	0.91	0.000	-.0169267 1.1466399
logtenure	.2798357	.0668345	4.18	0.000	.1466226 .4130489
_cons	-3.343205	.7896228	-4.23	0.000	-4.893229 -1.793181

```
Number of obs = 1801  
LR chi2(5) = 386.33  
Prob > chi2 = 0.0000  
Pseudo R2 = 0.9434
```



## A few organisational matters

- Course consists of a set of lectures and tutorials.
  - Lectures: Theory  
Tuesdays, 12-14h
  - Tutorials: Practise  
Wednesdays, 8-10h, WiWi 030
- Course material, i.e. slides and exercises will be uploaded in StudIP on Fridays prior to the respective sessions.
- THE Book: Introductory Econometrics: A Modern Approach  
by J.M. Wooldridge
- Grade: Exam (2h) at the end of the semester



# Introduction

# What is econometrics and why study it?

- Use of statistical methods to estimate economic relationships, test economic theories and evaluate policies and programmes.

Example: What is the effect of fertilizer on maize yields?

- Rare in economics to have experimental data (even though it is increasing).
- Use non-experimental data to make inferences.
- Theory may be ambiguous as to the effect of a programme or policy change  
⇒ can use econometrics to evaluate the programme.

## What is econometrics and why study it?

- Use of statistical methods to estimate economic relationships, test economic theories and evaluate policies and programmes.

Example: What is the effect of fertilizer on maize yields?

- Rare in economics to have experimental data (even though it is increasing).
- Use non-experimental data to make inferences.
- Theory may be ambiguous as to the effect of a programme or policy change  
⇒ can use econometrics to evaluate the programme.

## What is econometrics and why study it?

- Use of statistical methods to estimate economic relationships, test economic theories and evaluate policies and programmes.

Example: What is the effect of fertilizer on maize yields?

- Rare in economics to have experimental data (even though it is increasing).
- Use non-experimental data to make inferences.
- Theory may be ambiguous as to the effect of a programme or policy change  
⇒ can use econometrics to evaluate the programme.



## What is econometrics and why study it?

- Use of statistical methods to estimate economic relationships, test economic theories and evaluate policies and programmes.

Example: What is the effect of fertilizer on maize yields?

- Rare in economics to have experimental data (even though it is increasing).
- Use non-experimental data to make inferences.
- Theory may be ambiguous as to the effect of a programme or policy change  
⇒ can use econometrics to evaluate the programme.

# Types of data

## 1. Cross-sectional data

- Cross-sectional data is a random sample.
- Each observation is a new individual, firm, etc. with information at a point in time.
- If the data is not a random sample, we may have a sample-selection problem.
- Pooled cross-sectional data combines two or more cross-sectional data sets from different points in time.

## Types of data (cont.)

### 2. Time series data

- Time series data consists of a separate observation for each time period, e.g. share prices.
- Not a random sample and therefore different problems to consider.
- Trends and seasonality are important.

## Types of data (cont.)

### 3. Panel or longitudinal data

- Have repeated measurements, i.e. follow the same random individual observations over time.
- More difficult and costly to obtain.
- Valuable because they allow to study lags in behaviour and the results of individual decision making.

## The question of causality

- Simply establishing a relationship between variables is rarely sufficient.
- Need the effect to be considered causal.
- If we've truly controlled for all other effects, then the estimated *ceteris paribus* effect can be considered to be causal.
- Can be difficult to establish causality.

## Example: Returns to education

A model of human capital investment implies getting more education should lead to higher earnings. In the simplest case:

$$Earning = \beta_0 + \beta_1 Education + u \quad (1)$$

- The estimates of  $\beta_1$  is the return to education, but can it be considered causal?
- Note, that the error term,  $u$ , includes all other factors affecting earnings: Can we 'control' for them this way? What if some of them are related to education?

## Example: Returns to education

A model of human capital investment implies getting more education should lead to higher earnings. In the simplest case:

$$Earning = \beta_0 + \beta_1 Education + u \quad (1)$$

- The estimates of  $\beta_1$  is the return to education, but can it be considered causal?
- Note, that the error term,  $u$ , includes all other factors affecting earnings: Can we 'control' for them this way? What if some of them are related to education?

## Example: Returns to education

A model of human capital investment implies getting more education should lead to higher earnings. In the simplest case:

$$Earning = \beta_0 + \beta_1 Education + u \quad (1)$$

- The estimates of  $\beta_1$  is the return to education, but can it be considered causal?
- Note, that the error term,  $u$ , includes all other factors affecting earnings: Can we 'control' for them this way? What if some of them are related to education?



# The Simple Regression Model

## The simple regression model

Study the relationship between 2 variables.

$$y = \beta_0 + \beta_1 x + u \quad (2)$$

$y$

- dependent variable
- left hand variable
- regressand
- outcome variable

$x$

- independent variable
- control variable
- regressor
- explanatory variable
- covariate

## Functional relationship

If  $u$  is held fixed, i.e.  $\Delta u = 0$ , then  $x$  has a linear effect on  $y$ .

$$\Delta y = \beta_1 \Delta x \quad (3)$$

### **Assumption:**

The average value of  $u$ , the error term, in the population is 0.

$$E(u) = 0. \quad (4)$$

This is not a restrictive assumption, since we always use  $\beta_0$  to normalize  $E(u)$  to 0.

## Zero conditional mean assumption

We need to make a crucial assumption about how  $u$  and  $x$  are related.

We want it to be the case that knowing something about  $x$  does not give us any information about  $u$ , so that they are completely (here linear is sufficient) unrelated. That is, that

$$E(u|x) = E(u) \quad (5)$$

with

$$E(u) = 0 \Rightarrow E(u|x) = 0. \quad (6)$$

Example:  $u$  is land quality.  $E(u|x) = E(u) = 0$  means that land quality is the same no matter how much fertilizer we use.

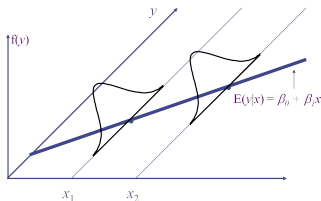
## Zero conditional mean assumption (cont.)

Thus, the conditional mean assumption means:

$$E(y|x) = \beta_0 + \beta_1 x \quad (7)$$

Tells us how much the average value of  $y$  changes when one unit of  $x$  changes.

See:  $E(y|x)$  as a linear function of  $x$ , where for any  $x$  the distribution of  $y$  is centred about  $E(y|x)$ .



## Ordinary Least Squares (OLS)

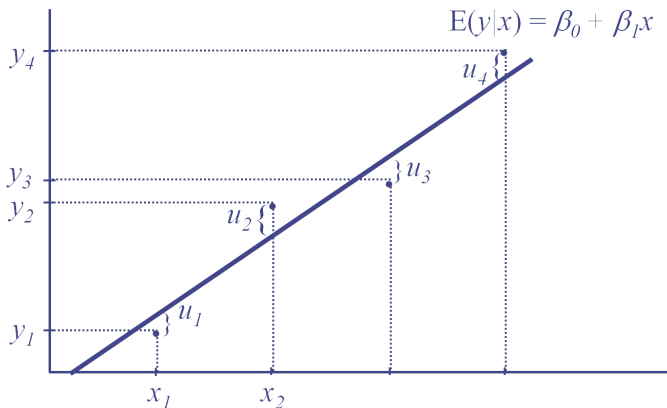
The idea of a regression is to estimate population parameters from a sample.

- Let  $(x_i, y_i) : i = 1, \dots, n$  denote a random sample of size  $n$  from the population of interest;
- for each observation in this sample it will be the case that

$$y_i = \beta_0 + \beta_1 x_i + u_i. \quad (8)$$

## Ordinary Least Squares (OLS) (cont.)

So, we can derive a regression line, sample data points and the associated error terms.



## Deriving OLS estimates

To derive the OLS estimates we need to realize that our main assumption of  $E(u|x) = E(u) = 0$  also implies that

$$\text{Cov}(x, u) = E(xu) - E(x)E(u) = 0. \quad (9)$$

We can write our two restrictions just in terms of  $x$ ,  $y$ ,  $\beta_0$  and  $\beta_1$ , since  $u = y - \beta_0 - \beta_1 x$

$$E(y - \beta_0 - \beta_1 x) = 0 \quad (10)$$

$$E[x(y - \beta_0 - \beta_1 x)] = 0 \quad (11)$$

These are called moment restrictions.



## Deriving OLS estimates (cont.)

The method of moments approach to estimation implies imposing the population moment restrictions on the sample moments.

We want to choose values of the parameters that will ensure that the sample versions of our moment restrictions are true.

What does this mean? Recall that for  $E(X)$ , the mean of a population distribution, a sample estimator of  $E(X)$  is simply the arithmetic mean of the sample.

## Deriving OLS estimates (cont.)

Recall law of large numbers, i.e.  $\frac{1}{n} \sum_{i=1}^n x_i = \bar{x} \rightarrow E[X]$ .

So the sample equivalent of the moment restrictions are:

$$n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (12)$$

$$n^{-1} \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (13)$$

Given the definition of a sample mean, and properties of summation, we can rewrite the first condition as follows

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \quad (14)$$

or,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (15)$$

## Deriving OLS estimates (cont.)

Plugging (15) into (13) (note: we drop  $n^{-1}$ ), we get

$$\sum_{i=1}^n x_i [y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i] = 0$$

and can rearrange it to derive

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

Provided that

$$\sum_{i=1}^n (x_i - \bar{x})^2 > 0 \quad (16)$$

the slope estimate is

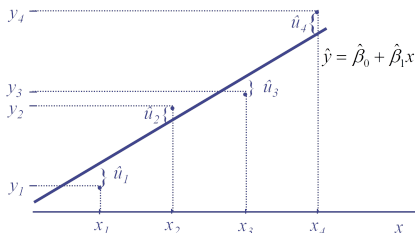
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (17)$$

## Summary: OLS estimates

- The intercept,  $\hat{\beta}_0$ , is  $(\bar{y} - \hat{\beta}_1\bar{x})$ .
- The slope estimate,  $\hat{\beta}_1$ , is the sample covariance between  $x$  and  $y$  divided by the sample variance of  $x$ .
- If  $x$  and  $y$  are positively correlated, the slope will be positive.
- If  $x$  and  $y$  are negatively correlated, the slope will be negative.
- To derive estimates, need  $x$  to vary in our sample.

## A few more things about OLS

- Intuitively, OLS is fitting a line through the sample points such that the sum of squared residuals is as small as possible, hence the term 'least squares'.
- The residual  $\hat{u}$ , is an estimate of the error term  $u$ , and is the difference between the fitted line and the sample point.



## Alternative approach to derivation

We want to choose our parameters such that we minimize the sum of squared residuals, i.e. the unexplained variance:

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad (18)$$

If we solve the minimization problem for the two parameters you obtain the following first order conditions, which are the same as before, multiplied by  $n$ .

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (19)$$

$$\sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (20)$$

## Algebraic properties of OLS

- The sum of the OLS residuals is zero, and thus the sample average of the OLS residuals is zero as well

$$\sum_{i=1}^n \hat{u}_i = 0 \quad \Rightarrow \quad \frac{\sum_{i=1}^n \hat{u}_i}{n} = 0 \quad (21)$$

- The sample covariance between the regressors and the OLS residuals is zero

$$\sum_{i=1}^n x_i \hat{u}_i = 0 \quad (22)$$

- The OLS regression line always goes through the mean of the sample

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \quad (23)$$

## A few more terms

Think of each observation as being made up of an explained part and an unexplained part,  $y_i = \hat{y}_i + \hat{u}_i$ . We then define

- $\sum_{i=1}^n (y_i - \bar{y})^2$  is the total sum of squares; (SST)
- $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  is the explained sum of squares (SSE);
- $\sum_{i=1}^n \hat{u}_i^2$  is the residual sum of squares (SSR);

which gives us  $SST = SSE + SSR$ .

Proof:

$$\begin{aligned}\sum (y_i - \bar{y})^2 &= \sum [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum [\hat{u}_i + (\hat{y}_i - \bar{y})]^2 \\ &= \sum \hat{u}_i^2 + 2 \sum \hat{u}_i (\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2 \\ &= SSR + 2 \sum \hat{u}_i (\hat{y}_i - \bar{y}) + SSE\end{aligned}\tag{24}$$



## A few more terms

Think of each observation as being made up of an explained part and an unexplained part,  $y_i = \hat{y}_i + \hat{u}_i$ . We then define

- $\sum_{i=1}^n (y_i - \bar{y})^2$  is the total sum of squares; (SST)
- $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  is the explained sum of squares (SSE);
- $\sum_{i=1}^n \hat{u}_i^2$  is the residual sum of squares (SSR);

which gives us  $SST = SSE + SSR$ .

Proof:

$$\begin{aligned}\sum (y_i - \bar{y})^2 &= \sum [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum [\hat{u}_i + (\hat{y}_i - \bar{y})]^2 \\ &= \sum \hat{u}_i^2 + 2 \sum \hat{u}_i (\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2 \\ &= SSR + 2 \sum \hat{u}_i (\hat{y}_i - \bar{y}) + SSE\end{aligned}\tag{24}$$

## Goodness-of-fit

How do we think about how well our sample regression line fits our sample data?

We compute the fraction of the total sum of squares (SST) that is explained by the model.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \quad (25)$$

**Interpretation:**

$R^2 \times 100$  = percentage of sample variation explained by  $x$ .

**Question:**

What should the value of  $R^2$  be, ideally?

## Unbiasedness of OLS

### Assumptions:

- The population model is linear in parameters, i.e.  
 $Y = \beta_0 + \beta_1 x + u$ .
- We have a random sample of size  $n$ ,  $(x_i, y_i) : i = 1, 2, \dots, n$ , from the population model.
- $E(u|x) = 0$  and thus  $E(u_i|x_i) = 0$ .
- There is variation in the  $x_i$ .

In order to think about unbiasedness, we need to rewrite our estimator in terms of the population parameter. Start by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{SSD_x} \quad (26)$$

$SSD_x$  (Sum of squared deviation) =  $\sum_{i=1}^n (x_i - \bar{x})^2$

## Unbiasedness of OLS

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum(x_i - \bar{x})y_i}{SSD_x} = \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{SSD_x} \\ &= \frac{\beta_0 \sum(x_i - \bar{x}) + \beta_1 \sum(x_i - \bar{x})x_i + \sum(x_i - \bar{x})u_i}{SSD_x} \\ &= \frac{\beta_1 \sum(x_i - \bar{x})x_i + \sum(x_i - \bar{x})u_i}{SSD_x} \\ &= \beta_1 + \frac{\sum(x_i - \bar{x})u_i}{SSD_x}\end{aligned}\tag{27}$$

Note:

$$\sum(x_i - \bar{x})x_i = \sum(x_i - \bar{x})^2 = SSD_x.$$

If  $x$  and  $u$  are uncorrelated  $\Rightarrow \sum(x_i - \bar{x})u_i = 0$

$$\hat{\beta}_1 = \beta_1 + \frac{1}{SSD_x} \sum(x_i - \bar{x})u_i \Rightarrow E[\hat{\beta}_1] = \beta_1\tag{28}$$

## Summary: Unbiasedness

- The OLS estimates of  $\beta_1$  and  $\beta_0$  are unbiased.
- Proof of unbiasedness depends on our 4 assumptions  
→ if any assumption fails, then the OLS is not necessarily unbiased.
- Remember unbiasedness is a description of the estimator - in a given sample we may be close or far from the true parameter.
- In econometrics, much emphasis is given to unbiasedness (being considered at the heart of **identification**).

## Variance of the OLS estimators

We know now that the sampling distribution of our estimate is centred around the true parameter.

We want to think about how spread out this distribution is.

It is easier to think about this variance under the additional assumption, i.e.  $Var(u|x) = \sigma^2 \rightarrow$  **homoscedasticity**.

$$Var(u|x) = E(u^2|x) - [E(u|x)]^2$$

$$E(u|x) = 0, \text{ so } \sigma^2 = E(u^2|x) = E(u^2) = Var(u)$$

$\sigma^2$  is the unconditional variance of  $u$ , also called the error variance, and  $\sigma$  the standard deviation of the error.

## Variance of the OLS estimators (cont.)

Thus, we can say

$$E(y|x) = \beta_0 + \beta_1 x \text{ and } \text{Var}(y|x) = \sigma^2.$$

### Question:

Complete the sentence:

The larger  $\sigma$ , the ... is the distribution of ... affecting  $y$ .

## Variance of the OLS estimators (cont.)

Thus, we can say

$$E(y|x) = \beta_0 + \beta_1 x \text{ and } \text{Var}(y|x) = \sigma^2.$$

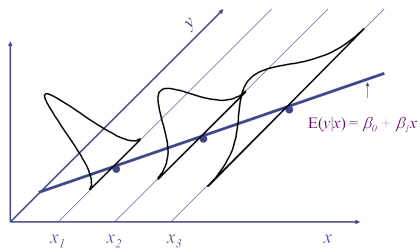
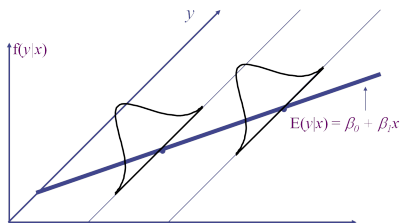
### **Question:**

Complete the sentence:

The larger  $\sigma$ , the ... is the distribution of ... affecting  $y$ .



# Homoscedasticity vs. Heteroscedasticity



## Variance of OLS estimator

$$d_i = (x_i - \bar{x})$$

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \text{Var}\left(\beta_1 + \frac{1}{SSD_x} \sum d_i u_i\right) = \left(\frac{1}{SSD_x}\right)^2 \text{Var}\left(\sum d_i u_i\right) \\ &= \left(\frac{1}{SSD_x}\right)^2 \sum d_i^2 \sigma^2 = \sigma^2 \left(\frac{1}{SSD_x}\right)^2 \sum d_i^2 = \frac{\sigma^2}{SSD_x} \end{aligned} \quad (29)$$

### Summary: Variance of OLS estimator

- The larger the error variance,  $\sigma^2$ , the larger the variance of the slope estimate
- The larger the variability of the  $x_i$ , the smaller  $\text{Var}[\hat{\beta}_1]$
- A larger sample size should decrease the variance of the slope estimate

## Variance of OLS estimator

$$d_i = (x_i - \bar{x})$$

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \text{Var}\left(\beta_1 + \frac{1}{SSD_x} \sum d_i u_i\right) = \left(\frac{1}{SSD_x}\right)^2 \text{Var}\left(\sum d_i u_i\right) \\ &= \left(\frac{1}{SSD_x}\right)^2 \sum d_i^2 \sigma^2 = \sigma^2 \left(\frac{1}{SSD_x}\right)^2 \sum d_i^2 = \frac{\sigma^2}{SSD_x} \end{aligned} \quad (29)$$

### Summary: Variance of OLS estimator

- The larger the error variance,  $\sigma^2$ , the larger the variance of the slope estimate
- The larger the variability of the  $x_i$ , the smaller  $\text{Var}[\hat{\beta}_1]$
- A larger sample size should decrease the variance of the slope estimate

## Estimating the Error Variance

We don't observe the errors,  $u_i$ , so we don't know what the error variance,  $\sigma^2$ , is.

What we observe are the residuals,  $\hat{u}_i$ .

We can use the residuals to form an estimate of the error variance.

$$\begin{aligned}\hat{u}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = (\hat{\beta}_0 + \hat{\beta}_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i\end{aligned}\tag{30}$$

Then, an unbiased estimator of  $\sigma^2$  is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum \hat{u}_i^2 = \frac{SSR}{n-2}\tag{31}$$

## Review questions

Let *kids* denote the number of children ever born to a women, and let *educ* denote years of education of the women. A simple model relating fertility and education is

$$kids = \beta_0 + \beta_1 educ + u$$

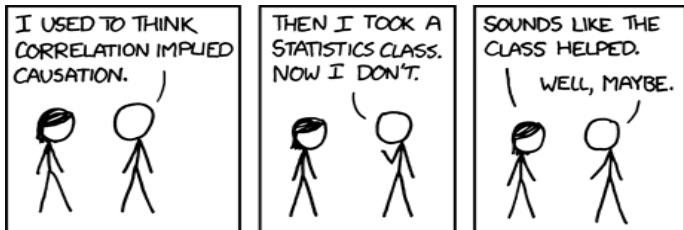
- 1 What kind of factors are contained in  $u$ ? Are these likely to be correlated with the level of education?
- 2 Will the simple regression analysis uncover the c.p. effect of education on fertility? Explain.
- 3 Write down the regression function you would estimate.
- 4 The regression analysis returns the following parameters:  $\hat{\beta}_0 = 5.38$ ;  $\hat{\beta}_1 = 0.47$ ; and  $R^2 = 0.034$ . How do you interpret these parameters?
- 5 Is  $\hat{\beta}_1$  unbiased? Explain why/why not?

# Introduction to Quantitative Methods for Development

Renate Hartwig, PhD

University of Passau

October 20, 2015



# Multiple Regression Analysis



## Multiple regression analysis

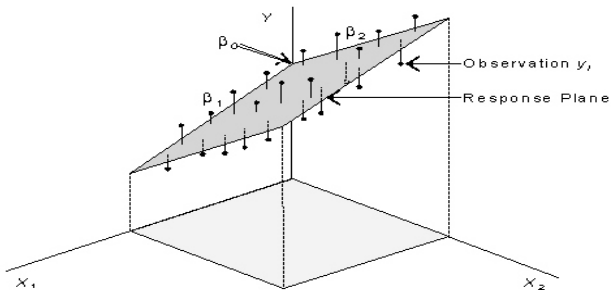
- Extension of the simple regression model.
- General form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_k x_k + u \quad (1)$$

### Parallels with simple regression

- $\beta_0$  is still the intercept.
- $\beta_1$  to  $\beta_k$  are all called slope parameters.
- $u$  is still the error term.
- Still need the zero conditional mean assumption, which now is  $E(u|x_1, x_2, \dots, x_k) = 0$ .
- Still minimizing the sum of squared residuals, now with  $k+1$  first order conditions.

## A graphic representation



## Interpreting multiple regression

Since we have a linear, additive model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \dots + \hat{\beta}_k x_k \quad (2)$$

we have

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \hat{\beta}_3 \Delta x_3 + \dots + \hat{\beta}_k \Delta x_k \quad (3)$$

So, holding  $x_2, \dots, x_k$  fixed implies that

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x_1 \quad (4)$$

i.e. each  $\hat{\beta}$  has a *ceteris paribus* interpretation.

## Simple vs. multiple regression estimates

Simple regression:  $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$

Multiple regression:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Generally,  $\tilde{\beta}_1 \neq \hat{\beta}_1$  unless:

- $\hat{\beta}_2 = 0$ , i.e. there is no partial effect of  $x_2$  on  $\hat{y}$ ; or
- $x_1$  and  $x_2$  are uncorrelated in the sample.

## Goodness-of-fit

Again, we can think of each observation as being made up of an explained part and an unexplained part, i.e.  $y_i = \hat{y}_i + \hat{u}_i$

Remember:

- $\sum_{i=1}^n (y_i - \bar{y})^2$  is the total sum of squares; (SST)
- $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  is the explained sum of squares (SSE);
- $\sum_{i=1}^n \hat{u}_i^2$  is the residual sum of squares (SSR);

and we already saw that,  $SST = SSE + SSR$ .

### Question:

What was the measure again that we use to say something about how well the regression line fits the data?

## Goodness-of-fit

Again, we can think of each observation as being made up of an explained part and an unexplained part, i.e.  $y_i = \hat{y}_i + \hat{u}_i$

Remember:

- $\sum_{i=1}^n (y_i - \bar{y})^2$  is the total sum of squares; (SST)
- $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  is the explained sum of squares (SSE);
- $\sum_{i=1}^n \hat{u}_i^2$  is the residual sum of squares (SSR);

and we already saw that,  $SST = SSE + SSR$ .

### Question:

What was the measure again that we use to say something about how well the regression line fits the data?

## R-squared

We can compute the fraction of the total sum of squares (SST) that is explained by the model, i.e.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \quad (5)$$

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

Alternatively, can also think of  $R^2$  as being equal to the squared correlation coefficient between the actual  $y_i$  and  $\hat{y}_i$ :

$$R^2 = \frac{[\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})]^2}{[\sum(y_i - \bar{y})^2][\sum(\hat{y}_i - \bar{\hat{y}})^2]} \quad (6)$$

## R-squared (cont.)

A few things about the  $R^2$  to bear in mind:

- $R^2$  can never decrease when another independent variable is added to a regression, and usually will increase.
- Because  $R^2$  will usually increase with the number of independent variables, it is not a good way to compare models.



## Assumptions for unbiasedness

- 1 The population model is linear in parameters, i.e.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \dots + \hat{\beta}_k x_k + u$$

- 2 We have a random sample of size  $n$ ,  
 $(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) : i = 1, 2, \dots, n$  from the population model, so that the sample model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + u_i$$

- 3  $E(u|x_1, x_2, \dots, x_k) = 0 \rightarrow$  zero conditional mean: implies that all explanatory variables are **exogenous**.
- 4 No perfect collinearity, i.e. none of the xes are constant and there are no exact linear relationships among them.

## Too many or too few variables?

What happens if we include variables in our specification that don't belong there?

⇒ There is no effect on the parameter estimates, i.e. the  $\hat{\beta}$ s remain unbiased.

What if we exclude a variable from our specification that does belong?

⇒ The OLS will be **biased** if the omitted variables are (cor)related with the included covariates.

## Too many or too few variables?

What happens if we include variables in our specification that don't belong there?

⇒ There is no effect on the parameter estimates, i.e. the  $\hat{\beta}$ s remain unbiased.

What if we exclude a variable from our specification that does belong?

⇒ The OLS will be **biased** if the omitted variables are (cor)related with the included covariates.

## Too many or too few variables?

What happens if we include variables in our specification that don't belong there?

⇒ There is no effect on the parameter estimates, i.e. the  $\hat{\beta}$ s remain unbiased.

What if we exclude a variable from our specification that does belong?

⇒ **The OLS will be biased if the omitted variables are (cor)related with the included covariates.**

## Example

You are interested in measuring the relationship between fertilizer and land quality.

What would your specification look like?

## Omitted variable bias

Suppose the true model is  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ .

But, we estimate  $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$

$$\Rightarrow \tilde{\beta}_1 = \frac{\sum (x_{i1} - \bar{x}_1) y_i}{\sum (x_{i1} - \bar{x}_1)^2} \quad (7)$$

Recall, that the true model is  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ , so that

$$\begin{aligned} \tilde{\beta}_1 &= \frac{\sum (x_{i1} - \bar{x}_1) (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i)}{\sum (x_{i1} - \bar{x}_1)^2} \\ &= \frac{\beta_1 \sum (x_{i1} - \bar{x}_1) x_{i1} + \beta_2 \sum (x_{i1} - \bar{x}_1) x_{i2} + \sum (x_{i1} - \bar{x}_1) u_i}{\sum (x_{i1} - \bar{x}_1)^2} \quad (8) \\ &= \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum (x_{i1} - \bar{x}_1)^2} + \frac{\sum (x_{i1} - \bar{x}_1) u_i}{\sum (x_{i1} - \bar{x}_1)^2} \end{aligned}$$

## Omitted variable bias

Suppose the true model is  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ .

But, we estimate  $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$

$$\Rightarrow \tilde{\beta}_1 = \frac{\sum (x_{i1} - \bar{x}_1) y_i}{\sum (x_{i1} - \bar{x}_1)^2} \quad (7)$$

Recall, that the true model is  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ , so that

$$\begin{aligned} \tilde{\beta}_1 &= \frac{\sum (x_{i1} - \bar{x}_1) (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i)}{\sum (x_{i1} - \bar{x}_1)^2} \\ &= \frac{\beta_1 \sum (x_{i1} - \bar{x}_1) x_{i1} + \beta_2 \sum (x_{i1} - \bar{x}_1) x_{i2} + \sum (x_{i1} - \bar{x}_1) u_i}{\sum (x_{i1} - \bar{x}_1)^2} \quad (8) \\ &= \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum (x_{i1} - \bar{x}_1)^2} + \frac{\sum (x_{i1} - \bar{x}_1) u_i}{\sum (x_{i1} - \bar{x}_1)^2} \end{aligned}$$

## Omitted variable bias

Suppose the true model is  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ .

But, we estimate  $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$

$$\Rightarrow \tilde{\beta}_1 = \frac{\sum(x_{i1} - \bar{x}_1)y_i}{\sum(x_{i1} - \bar{x}_1)^2} \quad (7)$$

Recall, that the true model is  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ , so that

$$\begin{aligned} \tilde{\beta}_1 &= \frac{\sum(x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i)}{\sum(x_{i1} - \bar{x}_1)^2} \\ &= \frac{\beta_1 \sum(x_{i1} - \bar{x}_1)x_{i1} + \beta_2 \sum(x_{i1} - \bar{x}_1)x_{i2} + \sum(x_{i1} - \bar{x}_1)u_i}{\sum(x_{i1} - \bar{x}_1)^2} \quad (8) \\ &= \beta_1 + \beta_2 \frac{\sum(x_{i1} - \bar{x}_1)x_{i2}}{\sum(x_{i1} - \bar{x}_1)^2} + \frac{\sum(x_{i1} - \bar{x}_1)u_i}{\sum(x_{i1} - \bar{x}_1)^2} \end{aligned}$$



## Omitted variable bias

Suppose the true model is  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ .

But, we estimate  $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$

$$\Rightarrow \tilde{\beta}_1 = \frac{\sum (x_{i1} - \bar{x}_1) y_i}{\sum (x_{i1} - \bar{x}_1)^2} \quad (7)$$

Recall, that the true model is  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ , so that

$$\begin{aligned} \tilde{\beta}_1 &= \frac{\sum (x_{i1} - \bar{x}_1) (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i)}{\sum (x_{i1} - \bar{x}_1)^2} \\ &= \frac{\beta_1 \sum (x_{i1} - \bar{x}_1) x_{i1} + \beta_2 \sum (x_{i1} - \bar{x}_1) x_{i2} + \sum (x_{i1} - \bar{x}_1) u_i}{\sum (x_{i1} - \bar{x}_1)^2} \quad (8) \\ &= \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum (x_{i1} - \bar{x}_1)^2} + \frac{\sum (x_{i1} - \bar{x}_1) u_i}{\sum (x_{i1} - \bar{x}_1)^2} \end{aligned}$$

## Omitted variable bias

Suppose the true model is  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ .

But, we estimate  $y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + u$

$$\Rightarrow \tilde{\beta}_1 = \frac{\sum (x_{i1} - \bar{x}_1) y_i}{\sum (x_{i1} - \bar{x}_1)^2} \quad (7)$$

Recall, that the true model is  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ , so that

$$\begin{aligned} \tilde{\beta}_1 &= \frac{\sum (x_{i1} - \bar{x}_1) (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i)}{\sum (x_{i1} - \bar{x}_1)^2} \\ &= \frac{\beta_1 \sum (x_{i1} - \bar{x}_1) x_{i1} + \beta_2 \sum (x_{i1} - \bar{x}_1) x_{i2} + \sum (x_{i1} - \bar{x}_1) u_i}{\sum (x_{i1} - \bar{x}_1)^2} \\ &= \beta_1 + \beta_2 \frac{\sum (x_{i1} - \bar{x}_1) x_{i2}}{\sum (x_{i1} - \bar{x}_1)^2} + \frac{\sum (x_{i1} - \bar{x}_1) u_i}{\sum (x_{i1} - \bar{x}_1)^2} \end{aligned} \quad (8)$$

## Omitted variable bias (cont.)

Taking the expectation yields

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \frac{\sum(x_{i1} - \bar{x}_1)x_{i2}}{\sum(x_{i1} - \bar{x}_1)^2} \quad (9)$$

Consider the regression of  $x_2$  on  $x_1$

$$\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1 \quad (10)$$

then

$$\tilde{\delta}_1 = \frac{\sum(x_{i1} - \bar{x}_1)x_{i2}}{\sum(x_{i1} - \bar{x}_1)^2} \quad (11)$$

So,

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \tilde{\delta}_1 \quad (12)$$

## The direction of the bias

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

## Summary: Omitted variable bias

- There are 2 cases where the bias is zero.
  - $\beta_2 = 0$ , that is  $x_2$  doesn't really belong in the model.
  - $x_1$  and  $x_2$  are uncorrelated in the sample.
- If the correlation between  $(x_2, x_1)$  and  $(x_2, y)$  is in the same direction, the bias will be positive.
- If the correlation between  $(x_2, x_1)$  and  $(x_2, y)$  is in opposite directions, the bias will be negative.
- Unfortunately, for  $k > 2$  these relations become much more complex, such that general statements are not possible any more.

### The more general case

- Technically, we can only put a sign on the bias for the more general case if all of the included  $x$ es are uncorrelated.
- Typically, we work through the bias assuming the  $x$ es are uncorrelated, as a useful guide even if this assumption is not strictly true.

## Variance of the OLS estimates

Now, that we know that the sampling distribution of our estimates is centred around the true parameter, we want to think about how spread out this distribution is.

It is easier to think about this variance under the additional assumption of homoskedasticity, i.e.

$$\text{Var}(u|x_1, x_2, \dots, x_k) = \sigma^2$$

Let  $x$  stand for  $(x_1, x_2, \dots, x_k)$

Assuming that  $\text{Var}(u|x) = \sigma^2$  also implies that  $\text{Var}(y|x) = \sigma^2$ .

⇒ The homoskedasticity assumption and the 4 assumptions for unbiasedness are known as the **Gauss-Markov assumptions**.

## Variance of the OLS estimates (cont.)

Given the Gauss-Makrov assumptions

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1 - R_j^2)} \quad (13)$$

where  $SST_j = \sum_{i=1}^n (x_{ij} - \hat{x}_j)^2$  and  $R_j^2$  is the  $R^2$  from regressing  $x_j$  on all other xes.

### Components of the OLS variances

- The error variance: a larger  $\sigma^2$  implies a larger variance for the OLS estimators.
- The total sample variation: a larger  $SST_j$  implies a smaller variance for the estimators.
- Linear relationships among the independent variables: a larger  $R_j^2$  implies are larger variance for the estimators.

## Misspecified models

Consider again the misspecified model

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 \quad (14)$$

so that  $\text{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{SST_1}$ .

Thus,  $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$  unless  $x_1$  and  $x_2$  are uncorrelated, then they are the same.

- While the variance of the estimator is smaller for the misspecified model, unless  $\beta_2 = 0$  the misspecified model is biased.
- As the sample grows, the variance of each estimator shrinks to zero, making the variance difference less important.



## Estimating the error variance

We don't know what the error variance,  $\sigma^2$ , is, because we don't observe the errors,  $u_i$ .

What we observe are the residuals,  $\hat{u}_i$ .

We can use the residuals to form an estimate of the error variance

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n - k - 1} = \frac{SSR}{df} \quad (15)$$

Thus,

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SST_j(1 - R_j^2)}} \quad (16)$$

where  $df = n - (k + 1) = n - k - 1$  (i.e. degrees of freedom) is the number of observations minus the number of estimated parameters.

# The Gauss-Markov theorem

Given the 5 Gauss-Markov assumptions, it can be shown that OLS is BLUE:

- **B**est (the one with the smallest variance)
- **L**inear
- **U**nbiased
- **E**stimator(s)

⇒ If the assumptions hold, use OLS!

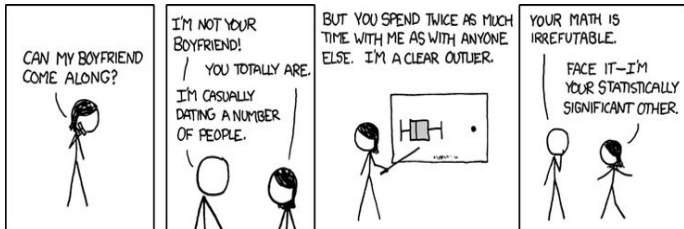
## Review questions

- 1 List the Gauss-Markov assumptions.
- 2 Explain in your own words what unbiasedness means.
- 3 Which of the following can cause OLS estimators to be biased?
  - (a) Heteroskedasticity
  - (b) Omitting an important variable
  - (c) A sample correlation coefficient of 0.95 between two independent variables included in the model
- 4 Suppose that average worker productivity at manufacturing firms (*avgprod*) depends on two factors, average hours of training (*avgtrain*) and average worker ability (*avgabil*) i.e.

$$avgprod = \beta_0 + \beta_1 avgtrain + \beta_2 avgabil + u$$

## Review questions (cont.)

Assume that this equation satisfies the Gauss-Markov assumptions. If grants have been given to firms whose workers have less than average ability, so that *avgtrain* and *avgabil* are negatively correlated, what is the likely bias in  $\tilde{\beta}_1$  obtained from the simple regression of *avgprod* and *avgtrain*?



# Multiple Regression Analysis: Inference

## Assumptions of the Classical Linear Model (CLM)

So far, we know that given the Gauss-Markov assumptions, OLS is BLUE.

To do hypothesis testing we need to make one more assumption:

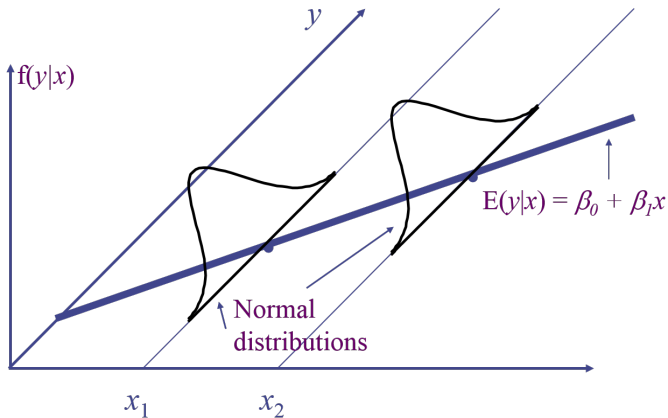
- $u$  is independent of  $x_1, x_2, \dots, x_k$  and  $u$  is normally distributed with zero mean and variance  $\sigma^2$  :  $u \sim Normal(0, \sigma^2)$

⇒ Under CLM, OLS is not only BLUE but is also the minimum variance unbiased estimator.

### Summary of the population assumptions of CLM

- $y | x \sim Normal(\beta_0 + \beta_1 x_1 + \dots + \beta_k x_k, \sigma^2)$
- For now we assume normality (clear that this is not the case sometimes). Large samples will let us drop normality.

# The homoskedastic normal distribution with a single explanatory variable





## Normal sampling distributions

Under the CLM assumptions, conditional on the sample values of the independent variables

$$\hat{\beta}_j \sim \text{Normal}[\beta_j, \text{var}(\hat{\beta}_j)] \quad (1)$$

so that

$$\frac{(\hat{\beta}_j - \beta_j)}{\text{sd}(\hat{\beta}_j)} \sim \text{Normal}(0, 1) \quad (2)$$

$\hat{\beta}_j$  is distributed normally because it is a linear combination of the errors.

## The t-test

Under the CML assumptions

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1} \quad (3)$$

Note, this is a  $t$ -distribution, because we have to estimate  $\sigma^2$ .

- Knowing the sampling distribution for the standardized estimator allows us to carry out hypothesis tests.
- Example: The null hypothesis  $H_0 : \beta_j = 0$ .
- Accepting the null means that  $x_j$  has no effect on  $y$ , controlling for other  $x$ es.

## The t-test (cont.)

To perform the t-test we need to determine the  $t$ -statistic for  $\hat{\beta}_j$

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \quad (4)$$

We will use the  $t$ -statistic together with the rejection rule to determine whether to accept the null hypothesis,  $H_0$ , or not.

## One- or two-sided alternatives

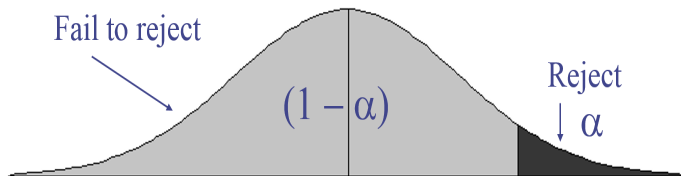
- Besides the null,  $H_0$ , we need an alternative hypothesis,  $H_1$  and a significance level.
- $H_1$  may be one- or two-sided
- One-sided:  $H_1 : \beta_j > 0$  or  $H_1 : \beta_j < 0$
- If we want to have only a 5% probability of rejecting  $H_0$  when it is really true, then we say our significance level is 5%.
- Given a significance level,  $\alpha$ , we look up the  $(1 - \alpha)^{th}$  percentile in a  $t$ -distribution with  $n - k - 1$  degrees of freedom and call this,  $c$ , the critical value.
- We reject the null hypothesis if the  $t$ -statistic is greater than the critical value.
- Is the  $t$ -statistic smaller than the critical value, then we fail to reject the null.

## One- or two-sided alternatives (cont.)

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + u_i$$

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j > 0$$



## One- or two-sided alternatives (cont.)

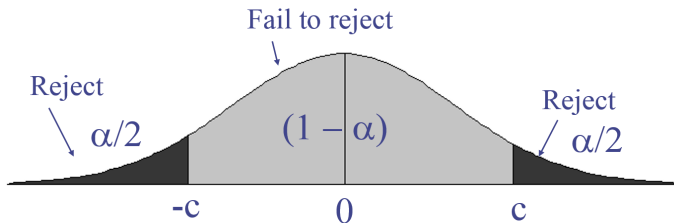
- Because the  $t$ -distribution is symmetric, testing  $H_1 : \beta_j < 0$  is straightforward. The critical value is just the negative of before.
- We reject the null if the  $t$ -statistic is smaller  $-c$ .
- If the  $t$ -statistic is larger  $-c$ , we fail to reject the null.
- For a two-sided test, we set the critical value based on  $\frac{\alpha}{2}$  and reject the  $H_1 : \beta_j \neq 0$  if the absolute value of the  $t$ -statistic  $> c$ .

## One- or two-sided alternatives (cont.)

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + u_i$$

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$



## Summary: $H_0 : \beta_j = 0$

- Unless stated otherwise, the alternative is assumed to be two-sided.
- If we reject the null, we typically say ‘ $x_j$  is statistically significant at the  $\alpha\%$  level’.
- If we fail to reject the null, we typically say ‘ $x_j$  is statistically insignificant at the  $\alpha\%$  level’



## Example

Table: Explaining wages in India

					R-squared	0.34
<b>wage</b>	<b>Coef.</b>	<b>Std. Err.</b>	<b>t</b>	$P > t$	<b>95% Conf. Intervall</b>	
age	4.97	0.61				
education	43.43	28.42				
gender	68.54	38.20				
_cons	-124.79	27.58				

## Testing other hypotheses

A more general form of the  $t$ -statistic recognizes that we may want to test something like  $H_0 : \beta_j = a_j$

In this case, the appropriate  $t$ -statistic is

$$t = \frac{(\hat{\beta}_j - a_j)}{se(\hat{\beta}_j)} \quad (5)$$

with  $a_j = 0$  for the standard test.

## Confidence intervals

Another way to use classical testing is to construct a confidence interval using the same critical value as for a two-sided test.

A  $(1 - \alpha)\%$  confidence interval is defined as

$$[\hat{\beta}_j - c * se(\hat{\beta}_j); \hat{\beta}_j + c * se(\hat{\beta}_j)] \quad (6)$$

where  $c$  is the  $1 - \frac{\alpha}{2}$  percentile in a  $t_{n-k-1}$  distribution.

We say that there is a chance of  $(1 - \alpha)\%$  that this interval contains the true value  $\beta_j$ .

## Example

Table: Explaining wages in India

					R-squared	0.34
<b>wage</b>	<b>Coef.</b>	<b>Std. Err.</b>	<b>t</b>	$P >  t $	<b>95% Conf. Intervall</b>	
age	4.97	0.61				
education	43.43	28.42				
gender	68.54	38.20				
_cons	-124.79	27.58				

## Computing $p$ -values for the $t$ -tests

- An alternative to the classical approach is to ask, ‘what is the smallest significance level at which the null can be rejected?’
- So, compute the  $t$ -statistic, and then look up what percentile it is in the appropriate  $t$ -distribution  $\Rightarrow$  this is the  $p$ -value.
- $p$ -value is the probability that we would observe the  $t$ -statistic we did, if the null were true.

# Economic vs. statistical significance

**What is what?**

**Why does it matter?**

## Testing a linear combination

Suppose instead of testing whether  $\beta_1$  is equal to a constant, you want to test if it is equal to another parameter, that is

$$H_0 : \beta_1 = \beta_2$$

We use the same procedure for the  $t$ -statistic

$$t = \frac{(\hat{\beta}_1 - \hat{\beta}_2)}{se(\hat{\beta}_1 - \hat{\beta}_2)} \quad (7)$$

Since

$$\begin{aligned} se(\hat{\beta}_1 - \hat{\beta}_2) &= \sqrt{Var(\hat{\beta}_1 - \hat{\beta}_2)} \\ &= \sqrt{Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2)} \quad (8) \\ &= \sqrt{[se(\hat{\beta}_1)]^2 + [se(\hat{\beta}_2)]^2 - 2s_{12}} \end{aligned}$$

## Testing a linear combination (cont.)

- Note:  $s_{12}$  is an estimate of  $Cov(\hat{\beta}_1, \hat{\beta}_2)$ .
- To use the formula, we need  $s_{12}$  which is not standard in many statistical packages.

- In Stata: after

```
reg y x1 x2 ... xk,
```

```
type
```

```
test = x1 = x2
```

to get the p-value for the test.



## Multiple linear restrictions

Everything we have done so far has involved testing a single linear restriction, e.g.  $\beta_1 = 0$  or  $\beta_1 = \beta_2$ .

However, we may want to jointly test multiple hypotheses about our parameters.

A typical example is testing exclusion restrictions, i.e. we want to know if a group of parameters are all equal to zero.

## Testing exclusion restrictions

- The null hypothesis might be something like

$$H_0 : \beta_{k-q-1} = 0, \dots, \beta_k = 0.$$

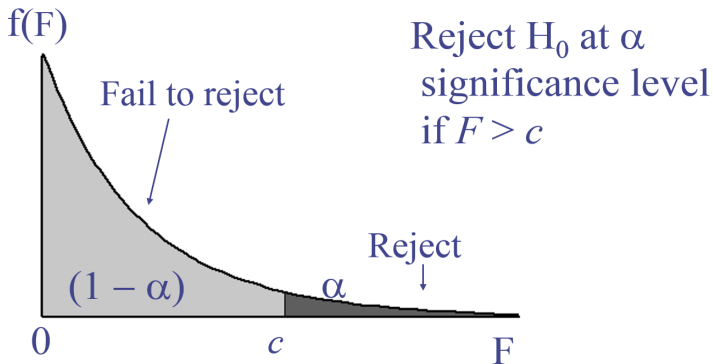
- The alternative:  $H_1 : H_0$  is not true.
- Cannot just check each  $t$ -statistic separately, because we want to know if the  $q$ -parameters are jointly significant at a given level. It is possible for none to be individually significant at that level.
- To do the test, we need to estimate the restricted model without  $x_{k-q-1}, \dots, x_k$  included, and the unrestricted model with all  $x$  included.
- Intuitively, we want to know if the change in SSR is big enough to warrant the inclusion of  $x_{k-q-1}, \dots, x_k$

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n - k - 1)} \quad (9)$$

## The F-statistic

- The  $F$ -statistic is always positive, since the SSR from the restricted model cannot be less than the SSR from the unrestricted.
- Essentially the  $F$ -statistic is measuring the relative increase in SSR when moving from the unrestricted to the restricted model.
- $q$  = number of restrictions, or  $df_r - df_{ur}$
- $n - k - 1 = df_{ur}$
- To decide if the increase in SSR when we move from the restricted model, is big enough to reject the exclusions, we need to know about the sampling distribution of the  $F$ -statistic.
- Not surprisingly,  $F \sim F_{q, n-k-1}$ , where  $q$  is referred to as the numerator degrees of freedom and  $n - k - 1$  as the denominator degrees of freedom.

## The F-statistic (cont.)



## Overall significance

- A special case of exclusion restrictions is to test  
 $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$
- Since the  $R^2$  from a model with only an intercept will be zero, the  $F$ -statistic is simply

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \quad (10)$$

## Summary: $F$ -statistic

- Just as with the  $t$ -statistic,  $p$ -values can be calculated by looking up the percentile in the appropriate  $F$ -distribution.
- Often in the economics literature the  $F$ -test is also known as the Wald-test.

## Review questions

- 1 Which of the following can cause the usual OLS  $t$ -statistic to be invalid (that is, not to have  $t$ -distributions under  $H_0$ )?
- (a) Heteroskedasticity.
  - (b) Omitting an important variable.
  - (c) A sample correlation coefficient of 0.95 between two independent variables included in the model.

## Review questions (cont.)

- 2 Consider an equation to explain child height (in cm) in terms of household income, mother education, the number of illness episodes since birth and if the child has been breast fed.
- (a) Write down the regression function to be estimated.
  - (b) In terms of model parameters, state the null hypothesis that, after controlling for income, illness and breastfeeding, mother education has no effect on height. State the alternative that better education increases child height.
  - (c) Test the null that education has no effect on height, against the alternative that education has a positive effect. The estimated parameter is 2.80, the standard error 0.35. Carry out the test at the 10% significance level.
  - (d) Would you include education in the final model? Explain.



<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
$\geq 0.1$	

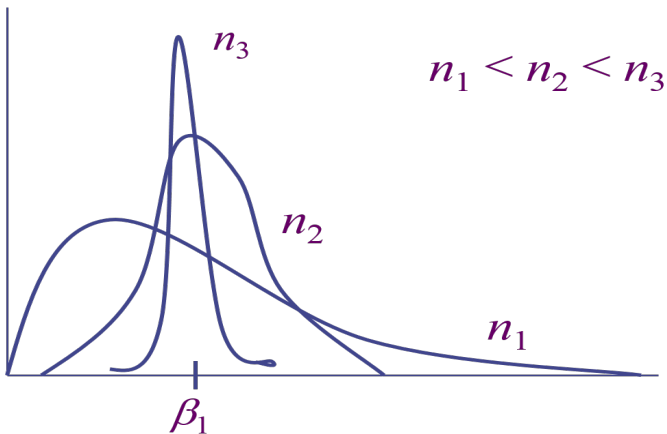
# Multiple Regression Analysis: Asymptotic Properties

## Consistency

- Under the Gauss-Markov assumptions, OLS is BLUE, but in other cases it won't always be possible to find unbiased estimators.
- In those cases, we may settle for estimators that are **consistent**.
- Consistent means that as  $n \rightarrow \infty$ , the distribution of the estimator collapses to the parameter value.
- Mathematically, a statistic  $\hat{\theta}_n$  is said to be a consistent estimator of  $\theta$  if:

$$plim(\hat{\theta}_n) = \theta$$

## Sampling distributions as $n$ increases



## A weaker assumption

- For unbiasedness, we assumed a zero conditional mean, i.e.  
 $E(u|x_1, x_2, \dots, x_k) = 0$
- For consistency, the weaker assumption of zero mean and zero correlation is sufficient, i.e.  
 $E(u) = 0$  and  $Cov(x_j, u) = 0$  for  $j = 1, 2, \dots, k$
- Without this assumption, OLS can be biased and inconsistent.

## Deriving inconsistency

- Just as we could derive the omitted variable bias earlier, now we want to think about the inconsistency. We also call it the **asymptotic bias**.

$$\text{True model: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v$$

$$\text{You may think: } y = \beta_0 + \beta_1 x_1 + u$$

So,

$$u = \beta_2 x_2 + v$$

$$plim \tilde{\beta}_1 = \beta_1 + \beta_2 \delta$$

where

$$\delta = \frac{Cov(x_1, x_2)}{Var(x_1)}$$

## Asymptotic bias

- Often thinking about the direction of the (asymptotic) bias is just like thinking about the direction of bias for omitted variables.
- The main difference is that the asymptotic bias uses the population variance and covariance, whereas the omitted variable bias uses the sample counterparts.
- Remember, inconsistency is a large sample problem - it doesn't go away as we add data.

## Asymptotic bias: An example

We are interested in housing prices.  $x_1$  is distance to the new football stadium,  $x_2$  is the quality of the house. Assume football stadiums depress housing prices, so  $\beta_1$  should be ... By definition  $\beta_2$  is positive. If football stadiums are on average built farther away from home, then distance and quality are positively correlated. So, what does this imply for  $\beta_1$ ?

Recall:

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias



## Large sample inference

- Recall that under the CLM assumptions, the sampling distributions are normal, so we could derive  $t$  and  $F$  distributions for testing.
- This exact normality was due to assuming the population error distribution was normal.
- This assumption of normal errors implied that the distribution of  $y$  given the  $x$ es was normal as well.
- Easy to come up with examples for which this exact normality will fail.

**Question:** Which ones can you think of?

- Any clearly skewed variable, like wages, arrests, savings, etc. can't be normal, since a normal distribution is symmetric.
- However, the normality assumption is not needed to conclude OLS is BLUE, only for inference.

## Large sample inference

- Recall that under the CLM assumptions, the sampling distributions are normal, so we could derive  $t$  and  $F$  distributions for testing.
- This exact normality was due to assuming the population error distribution was normal.
- This assumption of normal errors implied that the distribution of  $y$  given the  $x$ es was normal as well.
- Easy to come up with examples for which this exact normality will fail.

**Question:** Which ones can you think of?

- Any clearly skewed variable, like wages, arrests, savings, etc. can't be normal, since a normal distribution is symmetric.
- However, the normality assumption is not needed to conclude OLS is BLUE, only for inference.

## Law of Large Numbers and Central Limit Theorem (CLT)

Recall the Law of Large Numbers:

Sample moments are consistent estimates for the population moments, e.g. if the sample size is large enough the sample mean will converge to the true mean of the population

- Based on the CLT we can show that the OLS estimators are asymptotically normal.
- The CLT states that the standardized average of a random sample drawn from 'any' population with a mean  $\mu$  and variance  $\sigma^2$  is asymptotically normal, or

$$Z = \frac{\bar{Y} - \mu_y}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1) \quad (1)$$

## Asymptotic Normality

Under the Gauss-Markov assumptions we get

- 1  $\sqrt{n}(\hat{\beta}_j - \beta_j) \sim \text{Normal}(0, \frac{\sigma^2}{a_j^2})$ , where  $a_j^2 = \text{plim}(n^{-1} \sum \hat{r}_{ij}^2)$  and  $\hat{r}_{ij}$  is residual  $i$  from regression  $x_j$  on the other  $x_k$  including the intercept.
- 2  $\hat{\sigma}^2$  is a consistent estimator of  $\sigma^2$ .
- 3  $\frac{(\hat{\beta}_j - \beta_j)}{\text{se}(\hat{\beta}_j)} \sim N(0, 1)$

Because the t-distribution approaches the normal distribution for large df, we can also say  $\frac{(\hat{\beta}_j - \beta_j)}{\text{se}(\hat{\beta}_j)} \sim t_{n-k-1}$

Note: while we no longer need to assume normality with a large sample, we still assume homoscedasticity.

## Asymptotic Standard Errors

- If  $u$  is not normally distributed, we sometimes will refer to the standard error as an asymptotic standard error, since

$$\text{Var}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{SST_j(1 - R_j^2)} \quad (2)$$

$$\text{se}(\hat{\beta}_j) \approx \frac{c_j}{\sqrt{n}} \quad (3)$$

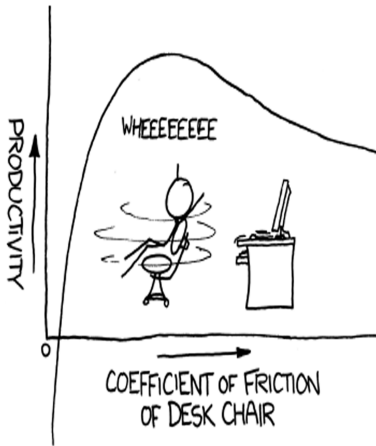
- So, we can expect standard errors to shrink at a rate proportional to the inverse of  $\sqrt{n}$  (so called 'rate of convergence')

## Asymptotic efficiency

- Estimators besides OLS will be consistent.
- However, under the Gauss-Markov assumptions, the OLS estimators will have the smallest asymptotic variances.
- We say that OLS is asymptotically efficient.
- Important to remember our assumptions though e.g. if not homoskedastic, not necessarily BLUE.
- Violations of other assumptions can be even worse, and will be discussed later.

## Review questions

- 1 What is the difference between unbiasedness and consistency?  
Explain
- 2 The sampling distribution of the mean becomes approximately normally distributed only when which of the following conditions is met?
  - (a) A single random sample is drawn from the population
  - (b) The population is normally distributed
  - (c) The sample size is large
  - (d) The standard deviation of the population is large





## Multiple Regression Analysis: Further Issues

## Redefining variables

- Changing the scale of the  $y$  variable will lead to a corresponding change in the scale of the coefficients and standard errors.  
⇒ The significance and interpretation will not change.
- Changing the scale of one  $x$  variable will lead to a change in the scale of that coefficient and standard error.  
⇒ The significance and interpretation will not change.

## Redefining variables

- Changing the scale of the  $y$  variable will lead to a corresponding change in the scale of the coefficients and standard errors.  
⇒ The significance and interpretation will not change.
- Changing the scale of one  $x$  variable will lead to a change in the scale of that coefficient and standard error.  
⇒ The significance and interpretation will not change.

## Redefining variables

- Changing the scale of the  $y$  variable will lead to a corresponding change in the scale of the coefficients and standard errors.  
⇒ The significance and interpretation will not change.
- Changing the scale of one  $x$  variable will lead to a change in the scale of that coefficient and standard error.  
⇒ The significance and interpretation will not change.

## Redefining variables: Example

Think of your Malawi dataset. You want to investigate the health status of the children in your dataset. There are many studies that use height (in meters) as a proxy for health assuming that healthier children are taller. You are interested if differences in the health status of the children can be explained by wealth proxied by monthly income (in USD). In your regression you also control for the age of the child (in yrs). The regression result that you get is as follows:

$$\hat{height} = 0.497 + 0.00001income + 0.157age$$

(0.034)    (0.000000)    (0.078)

### Questions:

- How do you interpret the results?
- Instead of measuring height in meters you want to measure it in cm. How does that change your regression results?

## Standardized coefficients

- Occasional you will see reference to a 'standardized coefficient' which has a specific meaning.
- The idea is to replace each  $y$  and  $x$  variable with a standardized version.
- To standardize we subtract the mean and divide by the standard deviation.  
(Note: The result is called a z-score. Examples that you will come across are height-for-age z-scores (HAZ), weight-for-height z-scores (WHZ) etc.)
- Coefficients reflect a change in standard deviation of  $y$  for a one standard deviation change in  $x$ .

# Standardized coefficients: Example



The screenshot shows the Guardian website's navigation bar with the logo and various menu items like 'sign in', 'subscribe', 'search', 'jobs', 'more', and 'International'. Below the navigation bar, there are category links such as 'UK', 'world', 'sport', 'football', 'opinion', 'culture', 'business', 'lifestyle', 'fashion', 'environment', 'tech', and 'travel'. A breadcrumb trail indicates the current page is 'home > environment > pollution > climate change > wildlife > energy'. The main headline reads 'Pollution London air pollution dangerously high, campaigners warn'.

Given the high level of air pollution in London we might wonder if it influences housing prices. The population model we are estimating is:

$$price = \beta_0 + \beta_1 nox + \beta_2 crime + \beta_3 rooms + u$$

While we can estimate the model in level terms we could also estimate it converting each variable into a z-score. That gives us:

$$zprice = -0.34znox - 0.143zcrime + 0.514zrooms$$



## Functional forms

- So far, we only looked at a strictly linear relationship between  $y$  and  $x$ .
- But, OLS can also be used for relationships that are not strictly linear in  $x$  and  $y$  by using non-linear functions of  $x$  and  $y$ .
- Remember, the model will still be linear in parameters (**Remember the Gauss-Markov assumptions**).
- Examples:
  - Natural log of  $x$ ,  $y$  or both
  - Quadratic form of  $x$
  - Use interactions of  $x$ 's

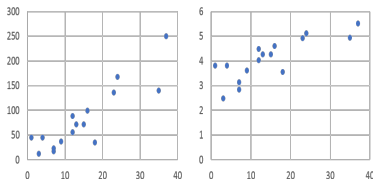


## Log models: Types and interpretation

Model	Dep.	Ind.	Example	Interpretation of $\beta_1$
Level-level	$y$	$x$	$y = \beta_0 + \beta_1 x + u$	Change in $y$ as $x$ changes by 1 unit
Level-log	$y$	$\ln(x)$	$y = \beta_0 + \beta_1 \ln(x) + u$	Change in $y$ for a 100% change in $x$
Log-level	$\ln(y)$	$x$	$\ln(y) = \beta_0 + \beta_1 x + u$	% change in $y$ given a one unit change in $x$
Log-log	$\ln(y)$	$\ln(x)$	$\ln(y) = \beta_0 + \beta_1 \ln(x) + u$	Elasticity of $y$ with respect to $x$

## Why use log models?

- Log models are invariant to the scale of the variables since they measure percentage changes.
- They give a direct estimate of elasticity.
- For models with  $y > 0$ , the conditional distribution is often heteroskedastic or skewed, while for  $\ln(y)$  this is much less so.
- The distribution of  $\ln(y)$  is more narrow, limiting the effect of outliers.



## A few rules of thumb using log vs. level

Types of variables that are often used in log form:

- Very large variables such as population figures
- Monetary values which must be positive (e.g. Dollar amounts)

Types of variables that are often used in level form:

- Variables that are measured in year
- Variables that are a proportion or percent

## A few rules of thumb using log vs. level

Types of variables that are often used in log form:

- Very large variables such as population figures
- Monetary values which must be positive (e.g. Dollar amounts)

Types of variables that are often used in level form:

- Variables that are measured in year
- Variables that are a proportion or percent

## A few rules of thumb using log vs. level

Types of variables that are often used in log form:

- Very large variables such as population figures
- Monetary values which must be positive (e.g. Dollar amounts)

Types of variables that are often used in level form:

- Variables that are measured in year
- Variables that are a proportion or percent

## A few rules of thumb using log vs. level

Types of variables that are often used in log form:

- Very large variables such as population figures
- Monetary values which must be positive (e.g. Dollar amounts)

Types of variables that are often used in level form:

- Variables that are measured in year
- Variables that are a proportion or percent

## Quadratic models

- Quadratic models are of the following functional form:  
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$
- We cannot interpret  $\beta_1$  alone as measuring the change in  $y$  with respect to  $x$ , we need to take into account  $\beta_2$  as well.
- Interpretation:  $\frac{\Delta \hat{y}}{\Delta x} = \hat{\beta}_1 + 2\hat{\beta}_2 x$

Example:

$$\begin{aligned} \hat{wage} = & 3.73 + 0.298 \text{exper} - 0.0061 \text{exper}^2 \\ & (0.35) \quad (0.041) \quad (0.0009) \end{aligned}$$

What is the effect of experience on wage?

## Quadratic models (cont.)

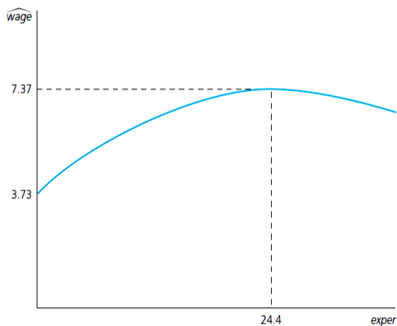
Suppose that the coefficient on  $x$  is positive and the coefficient on  $x^2$  is negative.



## Quadratic models (cont.)

Suppose that the coefficient on  $x$  is positive and the coefficient on  $x^2$  is negative.

- This means  $y$  is increasing in  $x$  first but will eventually turn and be decreasing in  $x$ .
- Turning point:  $x^* = \left| \frac{\hat{\beta}_1}{2\hat{\beta}_2} \right|$



## Quadratic models (cont.)

Suppose that the coefficient on  $x$  is negative and the coefficient on  $x^2$  is positive.

- This means  $y$  is decreasing in  $x$  first but will eventually turn and be increasing in  $x$ .
- Turning point: same as before.

**Question:** What would that look like graphically?

## Quadratic models (cont.)

Suppose that the coefficient on  $x$  is negative and the coefficient on  $x^2$  is positive.

- This means  $y$  is decreasing in  $x$  first but will eventually turn and be increasing in  $x$ .
- Turning point: same as before.

**Question:** What would that look like graphically?

## Quadratic models (cont.)

Suppose that the coefficient on  $x$  is negative and the coefficient on  $x^2$  is positive.

- This means  $y$  is decreasing in  $x$  first but will eventually turn and be increasing in  $x$ .
- Turning point: same as before.

**Question:** What would that look like graphically?

## Interaction terms

- Functional form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

- We cannot interpret  $\beta_1$  alone as measuring the change in  $y$  with respect to  $x_1$  we need to take into account  $\beta_3$  as well.
- Interpretation:  $\frac{\Delta \hat{y}}{\Delta x_1} = \beta_1 + \beta_3 x_2$
- To summarize the effect of  $x_1$  on  $y$  we typically evaluate it at  $\bar{x}_2$ .

## Adjusted R-squared

- Recall that the  $R^2$  will always increase as more variables are added to the model.
- The adjusted  $R^2$  takes into account the number of variables in the model, and may decrease.

$$adj - R^2 = 1 - \frac{[SSR/(n - k - 1)]}{[SST/(n - 1)]} \quad (1)$$

- The adjusted  $R^2$  is just  $(1 - R^2)(n - 1)/(n - k - 1)$ .
- Stata will give you both  $R^2$  and adjusted  $R^2$ .
- You can compare the fit of 2 models (with the same  $y$ ) using the adjusted  $R^2$ .
- **But**, you cannot use it to compare models with different  $y$ s (e.g.  $y$  vs.  $\ln(y)$ ).

## Adjusted R-squared

- Recall that the  $R^2$  will always increase as more variables are added to the model.
- The adjusted  $R^2$  takes into account the number of variables in the model, and may decrease.

$$adj - R^2 = 1 - \frac{[SSR/(n - k - 1)]}{[SST/(n - 1)]} \quad (1)$$

- The adjusted  $R^2$  is just  $(1 - R^2)(n - 1)/(n - k - 1)$ .
- Stata will give you both  $R^2$  and adjusted  $R^2$ .
- You can compare the fit of 2 models (with the same  $y$ ) using the adjusted  $R^2$ .
- **But**, you cannot use it to compare models with different  $y$ s (e.g.  $y$  vs.  $\ln(y)$ ).

## A few remarks on the Goodness-of-fit

- Despite all, it is important not to fixate too much on the adjusted  $R^2$  and lose sight of theory and common sense.
- If economic theory clearly predicts a variable belongs, generally leave it in.
- In contrast, you do not want to include a variable that prohibits a sensible interpretation of the variable of interest.

**Example?**



## Residual analysis

Information can be obtained from looking at residuals (i.e. predicted vs. observed values).

Example: Regress the price of cars on car characteristics. If the residuals are big it that good or bad?



## Review questions

- 1 The following model allows the returns to education to depend upon the total amount of both parents' education called  $par$ . The estimated equation is:

$$\log(\hat{w}age) = 5.65 + 0.47edu + 0.00078educ * par + 0.19exper$$

(0.13)    (0.01)    (0.00021)    (0.004)

$n = 722, R^2 = 0.169$

How do you interpret the results?

## Review questions

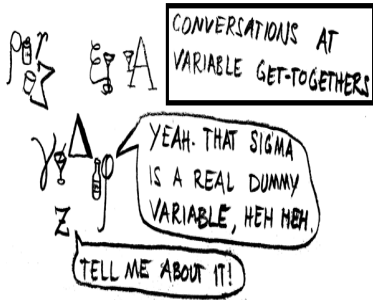
- 1 The following model allows the returns to education to depend upon the total amount of both parents' education called  $par$ . The estimated equation is:

$$\log(\hat{w}age) = 5.65 + 0.47edu + 0.00078educ * par + 0.19exper$$

(0.13)    (0.01)    (0.00021)    (0.004)

$n = 722, R^2 = 0.169$

How do you interpret the results?



# Multiple Regression Analysis: Dummy Variables

## Dummy variable

- A dummy variable is a variable that takes on values 1 or 0.
- Examples:
  - male (=1 if male, 0 otherwise)
  - south (=1 if south, 0 otherwise)
- Dummy variables are also called binary variables, for obvious reasons.

## Remember...

<i>person</i>	<i>wage</i>	<i>educ</i>	<i>exper</i>	<i>female</i>	<i>married</i>
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

## A dummy independent variable

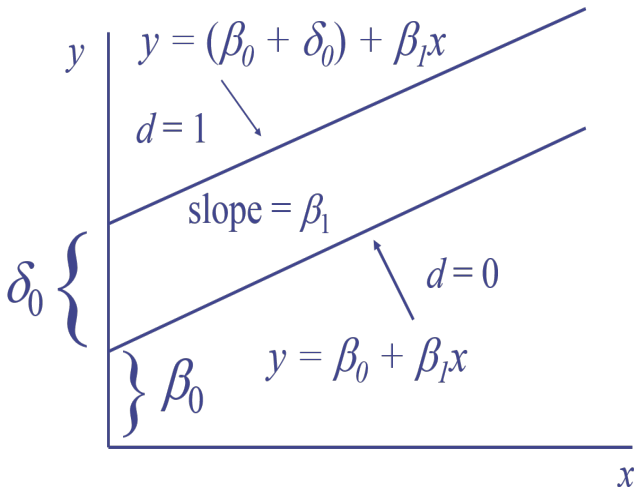
Consider a simple model with one continuous variable ( $x$ ) and one dummy ( $d$ ).

$$y = \beta_0 + \delta_0 d + \beta_1 x + u$$

- This can be interpreted as an intercept shift.
- If  $d = 0$ , then  $y = \beta_0 + \beta_1 x + u$
- If  $d = 1$ , then  $y = (\beta_0 + \delta_0) + \beta_1 x + u$
- $d = 0$  is the reference group.



Graphically:  $\delta_0 > 0$



## Dummies for multiple categories

- We can use dummy variables to control for something with multiple categories.
- Suppose, in your data people are categorized into one of three groups:
  - Those with no education,
  - those which have primary education, and
  - those that have secondary or higher education.
- To compare those with primary and secondary education to those without education, we have to include two dummy variables
  - *prime* = 1 if primary education and 0 otherwise
  - *second* = 1 if secondary education and 0 otherwise

## Dummies for multiple categories (cont.)

- Any categorical variable can be turned into a set of dummy variables.
- Because the reference group is represented by the intercept, if there are  $n$  categories there should be  $n - 1$  dummies.
- If there are a lot of categories, it may make sense to group them together, e.g. rankings 0-10, 11-25, etc.

## Interactions among dummies

- Interacting dummy variables is like subdividing the group.
- In our case, we have gender and education i.e. we have dummies for male, as well as, for the education level (primary, secondary or higher).
- For analysis we add  $male * prime$  and  $male * second$ , for a total of 5 dummy variables.  
⇒ We have 6 categories.
- Our reference category are **uneducated women**.
- $prime$  is for women with primary education.
- $second$  is for women with secondary education.
- the interaction terms  $male * prime$  and  $male * second$  reflect men with primary and secondary education respectively.

## Interactions among dummies (cont.)

Formally, the model is

$$y = \beta_0 + \delta_1 \text{male} + \delta_2 \text{prime} + \delta_3 \text{second} + \delta_4 \text{male} * \text{prime} + \delta_5 \text{male} * \text{second} + \beta_1 x + u \quad (1)$$

If  $\text{male} = 0$  and  $\text{prime} = 0$  and  $\text{second} = 0$ , then

$$y = \beta_0 + \beta_1 x + u \quad (2)$$

If  $\text{male} = 0$  and  $\text{prime} = 1$  and  $\text{second} = 0$ , then

$$y = \beta_0 + \delta_2 \text{prime} + \beta_1 x + u \quad (3)$$

If  $\text{male} = 1$  and  $\text{prime} = 0$  and  $\text{second} = 1$ , then

$$y = \beta_0 + \delta_1 \text{male} + \delta_3 \text{second} + \delta_5 \text{male} * \text{second} + \beta_1 x + u \quad (4)$$

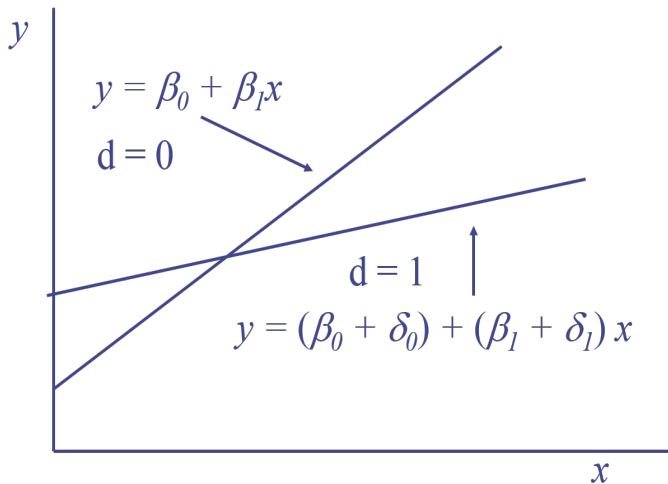
## Other interactions with dummies

- We can also consider interacting a dummy variable,  $d$ , with a continuous variable,  $x$ , i.e.,

$$y = \beta_0 + \delta_0 d + \beta_1 x + \delta_1 d * x + u \quad (5)$$

- If  $d = 0$ , then  $y = \beta_0 + \beta_1 x + u$
- If  $d = 1$ , then  $y = (\beta_0 + \delta_0) + (\beta_1 + \delta_1)x + u$
- This is interpreted as a change in slope.

Graphically:  $\delta_0 > 0$  and  $\delta_1 < 0$



## Interpretation: Example

We have estimate the following model:

$$\hat{w}age = 7.10 - 2.51 \textit{female}$$

(0.21)   (0.30)

Alternatively, we also estimated this:

$$\hat{w}age = -1.37 - 1.2 \textit{female} + 0.572 \textit{educ} + 0.025 \textit{exper}$$

(0.72)   (0.80)   (0.049)   (0.012)

- What does the model estimate/aim to achieve?
- How do you interpret the coefficients?
- How would the interpretation change if the dependent variable was measured in logs?



## Interpretation: Example

We have estimate the following model:

$$\hat{w}age = 7.10 - 2.51 \textit{female}$$

(0.21)   (0.30)

Alternatively, we also estimated this:

$$\hat{w}age = -1.37 - 1.2 \textit{female} + 0.572 \textit{educ} + 0.025 \textit{exper}$$

(0.72)   (0.80)   (0.049)   (0.012)

- What does the model estimate/aim to achieve?
- How do you interpret the coefficients?
- How would the interpretation change if the dependent variable was measured in logs?

## Interpretation: Example

We have estimate the following model:

$$\hat{w}age = 7.10 - 2.51 \textit{female}$$

(0.21)   (0.30)

Alternatively, we also estimated this:

$$\hat{w}age = -1.37 - 1.2 \textit{female} + 0.572 \textit{educ} + 0.025 \textit{exper}$$

(0.72)   (0.80)   (0.049)   (0.012)

- What does the model estimate/aim to achieve?
- How do you interpret the coefficients?
- How would the interpretation change if the dependent variable was measured in logs?

## Interpretation: Example

We have estimate the following model:

$$\hat{w}age = 7.10 - 2.51 \textit{female}$$

(0.21)   (0.30)

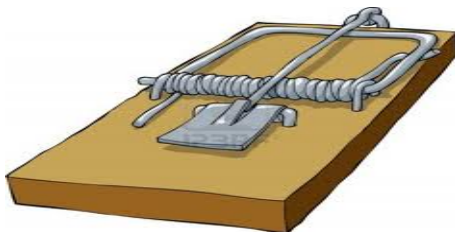
Alternatively, we also estimated this:

$$\hat{w}age = -1.37 - 1.2 \textit{female} + 0.572 \textit{educ} + 0.025 \textit{exper}$$

(0.72)   (0.80)   (0.049)   (0.012)

- What does the model estimate/aim to achieve?
- How do you interpret the coefficients?
- How would the interpretation change if the dependent variable was measured in logs?

Watch out



Don't step into the **dummy variable trap!**

## Testing for differences across groups

- Testing whether a regression function is different for one group versus another can be thought of as simply testing for the joint significance of the dummy and its interactions with all other  $x$  variables.
- So, you can estimate the model with all the interactions and without and form an  $F$ -statistic, but this could be unwieldy.

## The Chow test

- Turns out that you can compute the ‘proper’  $F$ -statistic without running the unrestricted model with interactions with all  $k$  continuous variables.
- Run the restricted model for group 1 and get  $SSR_1$ , then for group 2 to get  $SSR_2$ .
- Run the restricted model for all to get  $SSR$ , then

$$F = \frac{SSR - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} * \frac{n - 2(k + 1)}{k + 1} \quad (6)$$

- The Chow test is really just a simple  $F$ -test for exclusion restrictions, but we have realized that  $SSR_{ur} = SSR_1 + SSR_2$

## The Chow test (cont.)

- Note, we have  $k + 1$  restrictions (each of the slope coefficients and the intercept)
- Note, the unrestricted model would estimate 2 different intercepts and 2 different slope coefficients, so the df is  $n - 2k - 2$

## Linear probability model

- When we run an OLS when the dependent variable ( $y$ ) is a binary variable, we call this a linear probability model.
- Here, we interpret the  $\beta$ -coefficients as changes in the probability or likelihood as  $x$  changes.
- The problem here is that our predicted values of  $y$  can be outside  $[0, 1]$ .
- We will see logit and probit models later which restrict the range to  $[0, 1]$ .
- For now, it is a good place to start when our dependent variable is binary.



## Programme evaluation and some caveats

- A typical use of a dummy variable is when we are looking for a programme effect.
- For example, we may have individuals that received bed nets, a sanitation programme, a training programme, etc.
- We need to remember that many times individuals choose whether to participate in a programme or not, which may lead to a self-selection problem.

This is something you will look into in more detail in the advanced course. 😊

## Programme evaluation and some caveats

- A typical use of a dummy variable is when we are looking for a programme effect.
- For example, we may have individuals that received bed nets, a sanitation programme, a training programme, etc.
- We need to remember that many times individuals choose whether to participate in a programme or not, which may lead to a self-selection problem.

**This is something you will look into in more detail in the advanced course. 😊**

## ...But spoiler alert, sneak peak...

- If we can control for everything that is correlated with both participation and the outcome of interest then it is not a problem.
- Often, though, there are unobservables that are correlated with participation.
- In this case, the estimate of the programme effect is biased, and we do not want to set policies based on it... (or well,...)



## Review questions

- 1 Suppose you collected data from youth working in the informal sector in Laos (Nigeria). Your survey asks information on wages, education, experience, and gender. In addition, you ask for information about marijuana usage. The original question is: 'On how many separate occasions last month did you smoke marijuana?'
  - (a) Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, 'smoking marijuana five more times per month is estimated to change wage by x%.'
  - (b) Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug usage for men and women?

## Review questions (cont.)

- (c) Suppose you think it is better to measure marijuana usage by putting people into one of four categories: non-user, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana usage on wage.
- (d) Using the model in part (c), explain in detail how to test the null hypothesis that marijuana usage has no effect on wage. Be very specific and include a careful listing of the degrees of freedom.
- (e) What are some potential problems with drawing causal inference using the survey data that you collected?

I hope you do not feel like this...



Remember, it's half-time! You are almost there...

## Review questions

- 1 The following model allows the returns to education to depend upon the total amount of both parents' education called  $par$ . The estimated equation is:

$$\log(\hat{w}age) = 5.65 + 0.47edu + 0.00078educ * par + 0.19exper$$

(0.13)    (0.01)    (0.00021)    (0.004)

$n = 722, R^2 = 0.169$

How do you interpret the results?

## What you should know by now

- 1 What is OLS and what are the underlying assumptions?
- 2 What happens when the assumptions break down and what are potential remedies?
- 3 How to calculate and interpret estimates.
- 4 How to do hypothesis testing.
- 5 How to account for different functional forms of the relationship between your outcomes and covariates and what this implies.
- 6 When and how to use dummy variables.



## What you should know by now

- 1 What is OLS and what are the underlying assumptions?
- 2 What happens when the assumptions break down and what are potential remedies?
- 3 How to calculate and interpret estimates.
- 4 How to do hypothesis testing.
- 5 How to account for different functional forms of the relationship between your outcomes and covariates and what this implies.
- 6 When and how to use dummy variables.

## What you should know by now

- 1 What is OLS and what are the underlying assumptions?
- 2 What happens when the assumptions break down and what are potential remedies?
- 3 How to calculate and interpret estimates.
- 4 How to do hypothesis testing.
- 5 How to account for different functional forms of the relationship between your outcomes and covariates and what this implies.
- 6 When and how to use dummy variables.

## What you should know by now

- 1 What is OLS and what are the underlying assumptions?
- 2 What happens when the assumptions break down and what are potential remedies?
- 3 How to calculate and interpret estimates.
- 4 How to do hypothesis testing.
- 5 How to account for different functional forms of the relationship between your outcomes and covariates and what this implies.
- 6 When and how to use dummy variables.

## What you should know by now

- 1 What is OLS and what are the underlying assumptions?
- 2 What happens when the assumptions break down and what are potential remedies?
- 3 How to calculate and interpret estimates.
- 4 How to do hypothesis testing.
- 5 How to account for different functional forms of the relationship between your outcomes and covariates and what this implies.
- 6 When and how to use dummy variables.

## What you should know by now

- 1 What is OLS and what are the underlying assumptions?
- 2 What happens when the assumptions break down and what are potential remedies?
- 3 How to calculate and interpret estimates.
- 4 How to do hypothesis testing.
- 5 How to account for different functional forms of the relationship between your outcomes and covariates and what this implies.
- 6 When and how to use dummy variables.

## What you should know by now

- 1 What is OLS and what are the underlying assumptions?
- 2 What happens when the assumptions break down and what are potential remedies?
- 3 How to calculate and interpret estimates.
- 4 How to do hypothesis testing.
- 5 How to account for different functional forms of the relationship between your outcomes and covariates and what this implies.
- 6 When and how to use dummy variables.



Sorry.....  
Gifted people, please  
draw something on  
heteroscedasticity.

# Multiple Regression Analysis: Heteroskedasticity



## What is heteroskedasticity?

- Recall the assumption of homoskedasticity implied that conditional on the explanatory variables, the variance of the unobserved error,  $u$ , was constant.
- If this is not true, that is if the variance of  $u$  is different for different values of the  $x$ es, then the errors are heteroskedastic.
- Example:  
Low income families will spend little on vacations and the variation in holiday expenditures across these families will be small. In families with large incomes, the amount of discretionary income will be higher. The mean amount spent on vacations will be higher, and there will also be greater variability among such families, resulting in heteroskedasticity.

## What is heteroskedasticity?

- Recall the assumption of homoskedasticity implied that conditional on the explanatory variables, the variance of the unobserved error,  $u$ , was constant.
- If this is not true, that is if the variance of  $u$  is different for different values of the  $x$ es, then the errors are heteroskedastic.
- Example:  
Low income families will spend little on vacations and the variation in holiday expenditures across these families will be small. In families with large incomes, the amount of discretionary income will be higher. The mean amount spent on vacations will be higher, and there will also be greater variability among such families, resulting in heteroskedasticity.

## What is heteroskedasticity?

- Recall the assumption of homoskedasticity implied that conditional on the explanatory variables, the variance of the unobserved error,  $u$ , was constant.
- If this is not true, that is if the variance of  $u$  is different for different values of the  $x$ es, then the errors are heteroskedastic.

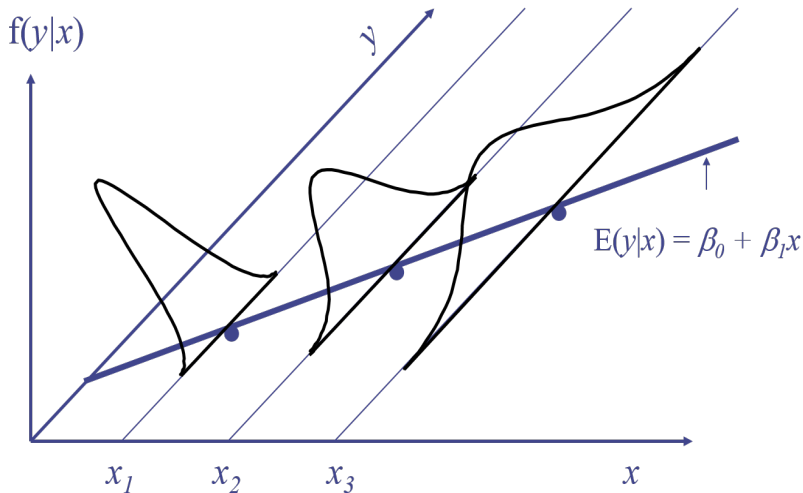
- Example:

Low income families will spend little on vacations and the variation in holiday expenditures across these families will be small. In families with large incomes, the amount of discretionary income will be higher. The mean amount spent on vacations will be higher, and there will also be greater variability among such families, resulting in heteroskedasticity.

## What is heteroskedasticity?

- Recall the assumption of homoskedasticity implied that conditional on the explanatory variables, the variance of the unobserved error,  $u$ , was constant.
- If this is not true, that is if the variance of  $u$  is different for different values of the  $x$ es, then the errors are heteroskedastic.
- Example:  
Low income families will spend little on vacations and the variation in holiday expenditures across these families will be small. In families with large incomes, the amount of discretionary income will be higher. The mean amount spent on vacations will be higher, and there will also be greater variability among such families, resulting in heteroskedasticity.

## Graphically



## When might heteroskedasticity occur?

- When the errors increase in either direction or in the extremes.
- Measurement error can cause heteroskedasticity, e.g. when some respondents provided more accurate responses than others.
- Heteroskedasticity can also occur if there are sub-population differences or other interaction effects (e.g. the effect of income on expenditures differs by ethnicity).
- Model misspecification can produce heteroskedasticity, i.e.
  - instead of using  $y$ , you should be using the log of  $y$ ;
  - instead of using  $x$ , you should be using  $x^2$  or both; and
  - important variables may be omitted from the model.

⇒ If the model were correctly specified, you might find that the patterns of heteroskedasticity disappear.

## When might heteroskedasticity occur?

- When the errors increase in either direction or in the extremes.
- Measurement error can cause heteroskedasticity, e.g. when some respondents provided more accurate responses than others.
- Heteroskedasticity can also occur if there are sub-population differences or other interaction effects (e.g. the effect of income on expenditures differs by ethnicity).
- Model misspecification can produce heteroskedasticity, i.e.
  - instead of using  $y$ , you should be using the log of  $y$ ;
  - instead of using  $x$ , you should be using  $x^2$  or both; and
  - important variables may be omitted from the model.

⇒ If the model were correctly specified, you might find that the patterns of heteroskedasticity disappear.

## When might heteroskedasticity occur?

- When the errors increase in either direction or in the extremes.
- Measurement error can cause heteroskedasticity, e.g. when some respondents provided more accurate responses than others.
- Heteroskedasticity can also occur if there are sub-population differences or other interaction effects (e.g. the effect of income on expenditures differs by ethnicity).
- Model misspecification can produce heteroskedasticity, i.e.
  - instead of using  $y$ , you should be using the log of  $y$ ;
  - instead of using  $x$ , you should be using  $x^2$  or both; and
  - important variables may be omitted from the model.

⇒ If the model were correctly specified, you might find that the patterns of heteroskedasticity disappear.



## When might heteroskedasticity occur?

- When the errors increase in either direction or in the extremes.
- Measurement error can cause heteroskedasticity, e.g. when some respondents provided more accurate responses than others.
- Heteroskedasticity can also occur if there are sub-population differences or other interaction effects (e.g. the effect of income on expenditures differs by ethnicity).
- Model misspecification can produce heteroskedasticity, i.e.
  - instead of using  $y$ , you should be using the log of  $y$ ;
  - instead of using  $x$ , you should be using  $x^2$  or both; and
  - important variables may be omitted from the model.

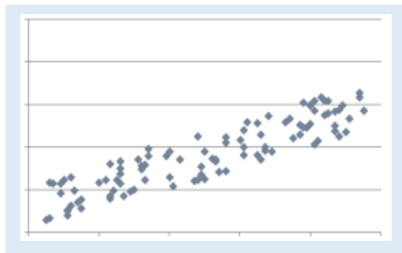
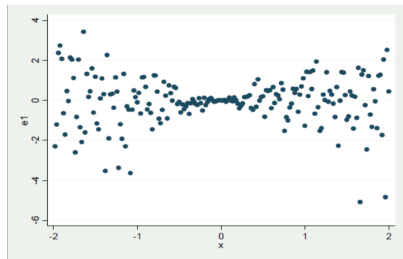
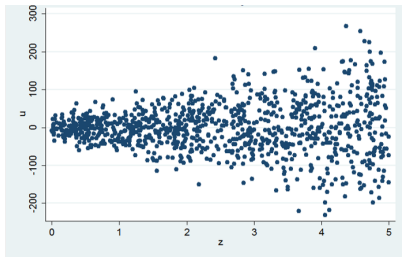
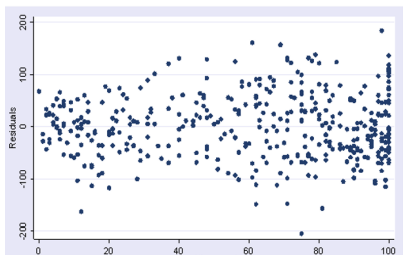
⇒ If the model were correctly specified, you might find that the patterns of heteroskedasticity disappear.

## When might heteroskedasticity occur?

- When the errors increase in either direction or in the extremes.
- Measurement error can cause heteroskedasticity, e.g. when some respondents provided more accurate responses than others.
- Heteroskedasticity can also occur if there are sub-population differences or other interaction effects (e.g. the effect of income on expenditures differs by ethnicity).
- Model misspecification can produce heteroskedasticity, i.e.
  - instead of using  $y$ , you should be using the log of  $y$ ;
  - instead of using  $x$ , you should be using  $x^2$  or both; and
  - important variables may be omitted from the model.

⇒ If the model were correctly specified, you might find that the patterns of heteroskedasticity disappear.

## So, what does heteroskedasticity look like?



## Why bother about heteroskedasticity?

The OLS is still unbiased and consistent, even if we do not assume homoskedasticity.

### But,...

- It is no longer BLUE. It is only LUE as it does not provide the estimate with the smallest variance anymore.  
⇒ OLS is no longer optimal.  
Why?
- The standard errors of the estimates are biased if we have heteroskedasticity.
- If the standard errors are biased, we can not use the usual  $t$ -statistics or  $F$ -statistics for drawing inferences.

## Why bother about heteroskedasticity?

The OLS is still unbiased and consistent, even if we do not assume homoskedasticity.

### But,...

- It is no longer BLUE. It is only LUE as it does not provide the estimate with the smallest variance anymore.  
⇒ OLS is no longer optimal.  
Why?
- The standard errors of the estimates are biased if we have heteroskedasticity.
- If the standard errors are biased, we can not use the usual  $t$ -statistics or  $F$ -statistics for drawing inferences.

## Why bother about heteroskedasticity?

The OLS is still unbiased and consistent, even if we do not assume homoskedasticity.

### But,...

- It is no longer BLUE. It is only LUE as it does not provide the estimate with the smallest variance anymore.  
⇒ OLS is no longer optimal.  
Why?
- The standard errors of the estimates are biased if we have heteroskedasticity.
- If the standard errors are biased, we can not use the usual  $t$ -statistics or  $F$ -statistics for drawing inferences.

## Why bother about heteroskedasticity?

The OLS is still unbiased and consistent, even if we do not assume homoskedasticity.

### But,...

- It is no longer BLUE. It is only LUE as it does not provide the estimate with the smallest variance anymore.

⇒ OLS is no longer optimal.

Why?

- The standard errors of the estimates are biased if we have heteroskedasticity.
- If the standard errors are biased, we can not use the usual  $t$ -statistics or  $F$ -statistics for drawing inferences.

## Variance with heteroskedasticity

In the simple case

$$\hat{\beta}_1 = \beta_1 + \frac{\sum(x_i - \bar{x})u_i}{\sum(x_i - \bar{x})^2} \Rightarrow \text{Var}(\hat{\beta}_1) = \frac{\sum(x_i - \bar{x})^2 \sigma_i^2}{(\sum(x_i - \bar{x})^2)^2} \quad (1)$$

For the general multiple regression model, a valid estimator of  $\text{Var}(\hat{\beta}_j)$  with heteroskedasticity is

$$\text{Var}(\hat{\beta}_j) = \frac{\sum \hat{r}_{ij}^2 \hat{u}_i^2}{SSR_j^2} \quad (2)$$

Though  $\hat{u}_i^2$  is not a consistent estimator of  $\sigma_j^2$ , where  $r_{ij}$  is the  $i$ th residual from regressing  $x_j$  on all other independent variables, and  $SSR_j$  is the sum of squared residuals from this regression.



# Detecting heteroscedasticity

## 1 Visual inspection:

- Plot the residuals against the fitted values.  
In Stata use the commands **rvfplot** or **rvpplot**.

## 2 Test:

- Breusch-Pagan test
- White test

## Testing for heteroskedasticity

Essentially we want to test

$H_0 : \text{Var}(u|x_1, x_2, \dots, x_k) = \sigma^2$ , which is equivalent to

$H_0 : E(u^2|x_1, x_2, \dots, x_k) = E(u^2) = \sigma^2$

- If we assume the relationship between  $u^2$  and  $x_j$  will be linear, we can test this as linear restriction.
- So, for  $u^2 = \delta_0 + \delta_1 x_1 + \dots + \delta_k x_k + v$  this means testing

$H_0 : \delta_1 = \delta_2 = \dots = \delta_k = 0$

In the following we will look at three test procedures which follow this basic idea.

## The Breusch-Pagan test

- We do not observe the error, but we can estimate it with the residuals from the OLS regression.
- After regressing the residuals squared on all of the  $x$ es, we can use the  $R^2$  to form an  $F$ -test
- The  $F$ -statistic is just the reported  $F$ -statistic for overall significance of the regression,

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \quad (3)$$

which is distributed  $F_{k, n-k-1}$

# The Breusch-Pagan test: Example

reg income educ jobexp

Source	SS	df	MS			
Model	1538.22521	2	769.112605		Number of obs =	20
Residual	282.200265	17	16.6000156		F( 2, 17) =	46.33
Total	1820.42548	19	95.8118671		Prob > F =	0.0000
					R-squared =	0.8450
					Adj R-squared =	0.8267
					Root MSE =	4.0743

income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	1.933393	.2099494	9.21	0.000	1.490438	2.376347
jobexp	.6493654	.1721589	3.77	0.002	.2861417	1.012589
_cons	-7.096855	3.626412	-1.96	0.067	-14.74791	.5542051

▶ RobustSE

▶ WLS

## The Breusch-Pagan test: Example (cont.)

### **estat hettest**

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
Ho: Constant variance
Variables: fitted values of income

chi2(1)      =    0.12
Prob > chi2  =    0.7238
```

## The White test

- The Breusch-Pagan test will detect any linear forms of heteroskedasticity.
- The White test allows for non-linearities by using squares and crossproducts of all the  $x$ es.
- We are still just using an  $F$  to test whether all the  $x_j$ ,  $x_j^2$ , and  $x_j x_h$  are jointly significant.
- This can get to be unwieldy pretty quickly.

# The White test: Example

## estat imtest, white

```
White's test for Ho: homoskedasticity
  against Ha: unrestricted heteroskedasticity

      chi2(5)      =      8.98
      Prob > chi2  =      0.1100
```

Cameron & Trivedi's decomposition of IM-test

Source	chi2	df	p
Heteroskedasticity	8.98	5	0.1100
Skewness	2.39	2	0.3022
Kurtosis	0.98	1	0.3226
Total	12.35	8	0.1363

## Alternate form of the White test

- Consider that the fitted values from OLS,  $\hat{y}$ , are a function of all the  $x$ es.
- Thus,  $\hat{y}^2$  will be a function of the squares and crossproducts and  $\hat{y}$  and  $\hat{y}^2$  can proxy for all of the  $x_j$ ,  $x_j^2$ , and  $x_j x_h$ .
- So, we can regress the residuals squared on  $\hat{y}$  and  $\hat{y}^2$  and use the  $R^2$  to form an  $F$ -statistic.
- Note, we are only testing for 2 restrictions now.



## Alternate form of the White test: Example

### reg logwage female educ exper expsq

```
. predict yhat
(option xb assumed; fitted values)
. predict v, residual
. gen yhatsq=yhat*yhat
. gen vsq=v*v

. * the White test
. reg vsq yhat yhatsq
```

Source	SS	df	MS			
Model	.605241058	2	.302620529	Number of obs =	526	
Residual	40.003265	523	.076488078	F( 2, 523) =	3.96	
Total	40.608506	525	.077349535	Prob > F =	0.0197	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yhat	-.187119	.2637874	-0.709	0.478	-.7053321	.331094
yhatsq	.0871914	.0812258	1.073	0.284	-.0723774	.2467603
_cons	.2334829	.2096599	1.114	0.266	-.1783961	.645362

```
. test yhat yhatsq

( 1) yhat = 0.0
( 2) yhatsq = 0.0

F( 2, 523) = 3.96
Prob > F = 0.0197
```

## Dealing with heteroscedasticity

- 1 Respecify the model/transform the variables
- 2 Use robust standard errors
- 3 Use Weighted Least Squares

## Robust standard errors

- If we have consistent estimates of the variance, the square root can be used as a standard error for inference.
- Need estimates robust against heteroskedasticity.
- Sometimes the estimated variance is corrected for degrees of freedom by multiplying by  $\frac{n}{n-k-1}$  but as  $n \rightarrow \infty$  it is all the same.
- It is important to remember that these robust standard errors only have asymptotic justification - with small sample sizes the  $t$ -statistic formed with robust standard errors will not have a distribution close to the  $t$  and inferences will not be correct.
- In Stata, robust standard errors are easily obtained using the robust option of **reg**, i.e.  
**reg y x, robust**

## Robust standard errors: Example

### reg income educ jobexp, robust

Regression with robust standard errors

Number of obs = 20  
F( 2, 17) = 48.15  
Prob > F = 0.0000  
R-squared = 0.8450  
Root MSE = 4.0743

income	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
educ	1.933393	.2006214	9.64	0.000	1.510119	2.356667
jobexp	.6493654	.1701214	3.82	0.001	.2904407	1.00829
_cons	-7.096855	3.365609	-2.11	0.050	-14.19767	.0039603

▶ Regression

## Weighted Least Squares

- While it is always possible to estimate robust standard errors for OLS estimates, if we know something about the specific form of the heteroskedasticity, we can obtain more efficient estimates than OLS.
- The basic idea is going to be to transform the model into one that has homoskedastic errors called Weighted Least Squares (WLS).

## Weighted Least Squares (cont.)

- WLS is great if we know what  $\text{Var}(u_i|x_i)$  looks like.
- In most cases, however, we will not know the form of heteroskedasticity.
- Example where we do is if data is aggregated, but model is individual level. e.g. no of employees per firm.
- Want to weight each aggregate observation by the inverse of the number of individuals.

# Weighted Least Squares: Example

**gen ineduc=1/educ**

**reg income educ jobexp [aw=ineduc]**

Source	SS	df	MS			
Model	1532.21449	2	766.107244			
Residual	151.090319	17	8.88766581			
Total	1683.30481	19	88.5949898			

Number of obs	=	20
F( 2, 17)	=	86.20
Prob > F	=	0.0000
R-squared	=	0.9102
Adj R-squared	=	0.8997
Root MSE	=	2.9812

income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
educ	1.795724	.1555495	11.54	0.000	1.467544	2.123905
jobexp	.4587992	.1628655	2.82	0.012	.115183	.8024155
_cons	-3.159669	1.94267	-1.63	0.122	-7.258345	.9390065

► Regression

## WLS wrap-up

- When doing  $F$ -tests with WLS, form the weights from the unrestricted model and use those weights to do WLS on the restricted model as well as the unrestricted model.
- Remember we are using WLS just for efficiency OLS is still unbiased & consistent.
- Estimates will still be different due to sampling error, but if they are very different then it is likely that some other Gauss-Markov assumption is false.



## Review questions

- 1 State with brief reason whether the following statements are true, false, or uncertain:
  - (a) In the presence of heteroskedasticity OLS estimators are biased as well as inefficient.
  - (b) If heteroskedasticity is present, the conventional  $t$  and  $F$ -tests are invalid.
  - (c) In the presence of heteroskedasticity the usual OLS method always overestimates the standard errors of estimators.
  - (d) If residuals estimated from an OLS regression exhibit a systematic pattern, it means heteroscedasticity is present in the data.
  - (e) If a regression model is mis-specified (e.g. an important variable is omitted), the OLS residuals will show a distinct pattern.
  - (f) If a regressor that has non-constant variance is (incorrectly) omitted from the model, the (OLS) residuals will be heteroskedastic.

## Review questions (cont.)

- ② Sex-selective abortion is a concern in India. In trying to understand what drives sex-selective abortion you begin your analysis by looking at the factors that determine the abortion rates across the 29 states of India. The variables that you used in your analysis are as follows:
- State = name of the state
  - ABR = abortion rate per thousand women aged 15-44 in 2012
  - Rel = percent of the population that is Hindu
  - P = average price charged in non-hospital facilities for abortion in USD
  - Law = takes value 1 if state enforces a law to restrict abortion
  - Educ = percent of state population 25 and older with a high school degree
  - Inc = disposable per capita income in USD, 2012
- (a) Do you think heteroskedasticity is a concern in your analysis? Explain.

## Review questions (cont.)

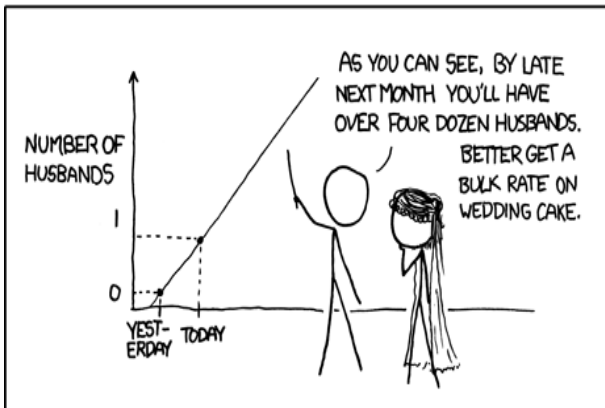
You run two models, the normal OLS and one with robust standard errors (Note: normal standard errors in parenthesis, robust standard errors in square brackets).

$$\hat{A}BR = 14.28 + 0.02Rel - 0.04P - 0.87Law - 0.028Educ + 0.002Inc$$

(15.07)	(0.08)	(0.02)	(2.37)	(0.19)	(0.0004)
[14.90]	[0.08]	[0.02]	[1.79]	[0.17]	[0.0005]

- (b) Interpret your results.
- (c) Which model do you prefer and why?

## MY HOBBY: EXTRAPOLATING



# Multiple Regression Analysis: Time Series

## What is time series data?

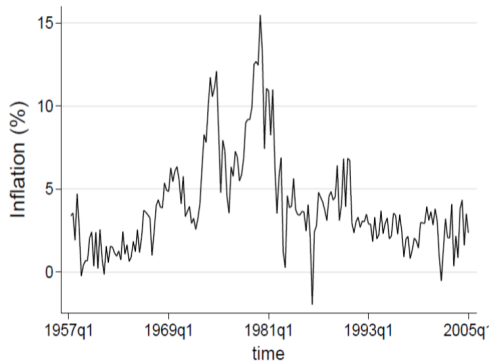
- **Cross-section data** is data collected for multiple entities at one point in time.
- **Panel data** is data collected for multiple entities at multiple points in time.
- **Time series data** is data collected for a single entity at multiple points in time. E.g.
  - Yearly GDP of Chile over the last 20 years.
  - Daily exchange rate CLP/EUR.
  - Quarterly data on inflation and unemployment in Chile from 1957 to 2005.

## What is time series data?

- **Cross-section data** is data collected for multiple entities at one point in time.
- **Panel data** is data collected for multiple entities at multiple points in time.
- **Time series data** is data collected for a single entity at multiple points in time. E.g.
  - Yearly GDP of Chile over the last 20 years.
  - Daily exchange rate CLP/EUR.
  - Quarterly data on inflation and unemployment in Chile from 1957 to 2005.

## What is time series data? (cont.)

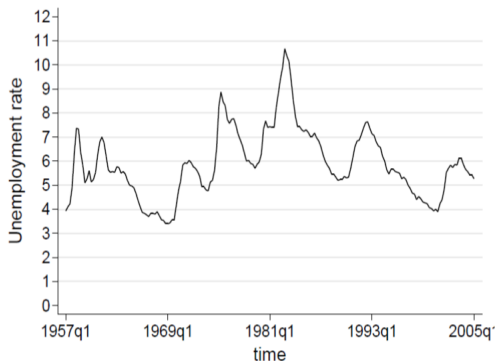
Quarterly time series data on inflation in Chile from 1957 to 2005





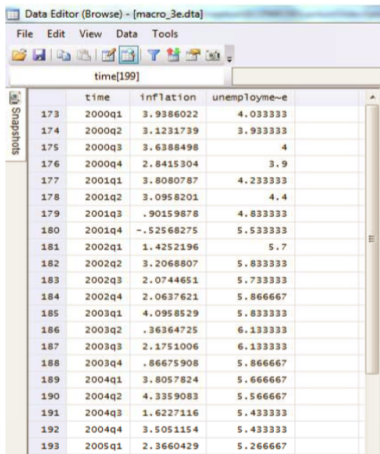
## What is time series data (cont.)

Quarterly time series data on unemployment in Chile from 1957 to 2005



## What is time series data (cont.)

Quarterly data on inflation and unemployment in Chile 1957-2005



	time	inflation	unemployme-e
173	2000q1	3.9386022	4.033333
174	2000q2	3.1231739	3.933333
175	2000q3	3.6388498	4
176	2000q4	2.8415304	3.9
177	2001q1	3.8080787	4.233333
178	2001q2	3.0958201	4.4
179	2001q3	.90159878	4.833333
180	2001q4	-.52568275	5.533333
181	2002q1	1.4252196	5.7
182	2002q2	3.2068807	5.833333
183	2002q3	2.0744651	5.733333
184	2002q4	2.0637621	5.866667
185	2003q1	4.0958529	5.833333
186	2003q2	.36364725	6.133333
187	2003q3	2.1751006	6.133333
188	2003q4	.86675908	5.866667
189	2004q1	3.8057824	5.666667
190	2004q2	4.3359083	5.566667
191	2004q3	1.6227116	5.433333
192	2004q4	3.5051154	5.433333
193	2005q1	2.3660429	5.266667

In Stata: `tsset time`



## What is time series data? (cont.)

Unlike cross-section data, time series data has a temporal ordering.

⇒ For estimation, we need to alter some of our assumptions to take into account that we no longer have a random sample of individuals.

# What do we use time series data for?

Time series regression models can be used for:

- Forecasting;
- Estimating (dynamic) causal effects.

## What do we use time series data for?

Time series regression models can be used for:

- Forecasting;
- Estimating (dynamic) causal effects.

## Some notation

- A particular observation  $y_t$  is indexed by the subscript  $t$ .
- $y_t$  is the value of the current period; the value of previous periods is  $y_{t-1} \rightarrow$  we also call this the first lag.
- In general  $y_{t-j}$  is called the  $j$ th lag and similarly  $y_{t+j}$  is the  $j$ th future value.
- The first difference  $y_t - y_{t-1}$  is the change in  $y$  from period  $t - 1$  to period  $t$ .
- The basic (static) regression relates variables from the same time period e.g.

$$y_t = \beta_0 + \beta_1 z_t + u_t \quad (1)$$

- Finite distributed lag models allow one or more variables to affect  $y$  with a lag:

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t \quad (2)$$

## Some notation

- A particular observation  $y_t$  is indexed by the subscript  $t$ .
- $y_t$  is the value of the current period; the value of previous periods is  $y_{t-1} \rightarrow$  we also call this the first lag.
- In general  $y_{t-j}$  is called the  $j$ th lag and similarly  $y_{t+j}$  is the  $j$ th future value.
- The first difference  $y_t - y_{t-1}$  is the change in  $y$  from period  $t - 1$  to period  $t$ .
- The basic (static) regression relates variables from the same time period e.g.

$$y_t = \beta_0 + \beta_1 z_t + u_t \quad (1)$$

- Finite distributed lag models allow one or more variables to affect  $y$  with a lag:

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t \quad (2)$$

## Some notation

- A particular observation  $y_t$  is indexed by the subscript  $t$ .
- $y_t$  is the value of the current period; the value of previous periods is  $y_{t-1} \rightarrow$  we also call this the first lag.
- In general  $y_{t-j}$  is called the  $j$ th lag and similarly  $y_{t+j}$  is the  $j$ th future value.
- The first difference  $y_t - y_{t-1}$  is the change in  $y$  from period  $t - 1$  to period  $t$ .
- The basic (static) regression relates variables from the same time period e.g.

$$y_t = \beta_0 + \beta_1 z_t + u_t \quad (1)$$

- Finite distributed lag models allow one or more variables to affect  $y$  with a lag:

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t \quad (2)$$



## Estimating causal effects vs. forecasting

- Time series data is often used for forecasting
  - For example, next year's economic growth is forecasted based on past and current values of growth & other (lagged) explanatory variables.
- Forecasting is quite different from estimating causal effects and is generally based on different assumptions.
- Models that are used for forecasting need not have a causal interpretation.
  - OLS coefficients need not be unbiased & consistent.
- Measures of fit, such as the (adjusted)  $R^2$ 
  - are not very informative when estimating causal effects.
  - are informative about the quality of a forecasting model.

## Autocorrelation

In time series data  $y_t$  is typically correlated with  $y_{t-1}$ . This is called **autocorrelation** or **serial correlation**.

### Consequence:

Autocorrelation will not affect the unbiasedness or consistency of OLS estimators, but it does affect their efficiency. With positive autocorrelation, the OLS estimates of the standard errors will be smaller than the true standard errors. This will lead to the conclusion that the parameter estimates are more precise than they really are.

⇒ We can use the Durbin-Watson test to test for autocorrelation.

## Autocorrelation

In time series data  $y_t$  is typically correlated with  $y_{t-1}$ . This is called **autocorrelation** or **serial correlation**.

Consequence:

Autocorrelation will not affect the unbiasedness or consistency of OLS estimators, but it does affect their efficiency. With positive autocorrelation, the OLS estimates of the standard errors will be smaller than the true standard errors. This will lead to the conclusion that the parameter estimates are more precise than they really are.

⇒ We can use the Durbin-Watson test to test for autocorrelation.

## Autocorrelation: Example

Suppose, we have data on quarterly sales from the tailoring industry in Burkina Faso for 5 years (*isales*), and we would like to use this information to model sales for Tailor C.

### reg csales isales

Source	SS	df	MS			
Model	110.256901	1	110.256901	Number of obs =	20	
Residual	.133302302	18	.007405683	F( 1, 18) =	14888.15	
				Prob > F =	0.0000	
				R-squared =	0.9988	
				Adj R-squared =	0.9987	
				Root MSE =	.08606	
Total	110.390204	19	5.81001072			

csales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
isales	.1762828	.0014447	122.02	0.000	.1732475	.1793181
_cons	-1.454753	.2141461	-6.79	0.000	-1.904657	-1.004849

```
. estat dwatson
```

```
Durbin-Watson d-statistic( 2, 20) = .7347276
```

⇒ The critical  $d$ -statistic is 1.2. In this case, we reject the null hypothesis of no serial correlation.

## Autocorrelation: Example

Suppose, we have data on quarterly sales from the tailoring industry in Burkina Faso for 5 years (*isales*), and we would like to use this information to model sales for Tailor C.

### reg csales isales

Source	SS	df	MS			
Model	110.256901	1	110.256901	Number of obs =	20	
Residual	.133302302	18	.007405683	F( 1, 18) =	14888.15	
				Prob > F =	0.0000	
				R-squared =	0.9988	
				Adj R-squared =	0.9987	
				Root MSE =	.08606	
Total	110.390204	19	5.81001072			

csales	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
isales	.1762828	.0014447	122.02	0.000	.1732475	.1793181
_cons	-1.454753	.2141461	-6.79	0.000	-1.904657	-1.004849

```
. estat dwatson
```

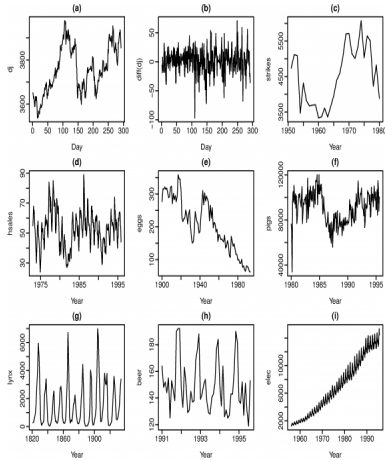
```
Durbin-Watson d-statistic( 2, 20) = .7347276
```

⇒ The critical  $d$ -statistic is 1.2. In this case, we reject the null hypothesis of no serial correlation.

# Stationarity

- A time series  $y_t$  is stationary if its probability distribution does not change over time.
- Stationarity implies that  $y_1$  has the same distribution as  $y_t$  for any  $t = 1, 2, \dots$
- In other words,  $y_1, y_2, \dots, y_t$  are identically distributed. However, they are not necessarily independent.
- If a series is non-stationary, then conventional hypothesis tests, confidence intervals and forecasts can be unreliable.

# Stationarity (cont.)



## First order autoregressive model: AR(1)

- Suppose we want to forecast the change in inflation from this quarter to the next.
- When predicting the future of a time series a good place to start is the immediate past.
- First order autoregressive model (AR(1)):

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t$$

- The forecast next period is based on the AR(1) model:

$$\hat{y}_{t+1} = \hat{\beta}_0 + \hat{\beta}_1 y_t$$

- The forecast error is the mistake made by the forecast

$$\text{Forecast error} = y_{t+1} - \hat{y}_{t+1}$$

### Question:

What is the difference between a forecast and a predicted value and the forecast error and the residual?



## First order autoregressive model: AR(1)

- Suppose we want to forecast the change in inflation from this quarter to the next.
- When predicting the future of a time series a good place to start is the immediate past.
- First order autoregressive model (AR(1)):

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t$$

- The forecast next period is based on the AR(1) model:

$$\hat{y}_{t+1} = \hat{\beta}_0 + \hat{\beta}_1 y_t$$

- The forecast error is the mistake made by the forecast

$$\text{Forecast error} = y_{t+1} - \hat{y}_{t+1}$$

### Question:

What is the difference between a forecast and a predicted value and the forecast error and the residual?

## First order autoregressive model: AR(1)

- Suppose we want to forecast the change in inflation from this quarter to the next.
- When predicting the future of a time series a good place to start is the immediate past.
- First order autoregressive model (AR(1)):

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t$$

- The forecast next period is based on the AR(1) model:

$$\hat{y}_{t+1} = \hat{\beta}_0 + \hat{\beta}_1 y_t$$

- The forecast error is the mistake made by the forecast

$$\text{Forecast error} = y_{t+1} - \hat{y}_{t+1}$$

### Question:

What is the difference between a forecast and a predicted value and the forecast error and the residual?

## AR(1): Example

$$\Delta inflation_t = \beta_0 + \beta_1 \Delta inflation_{t-1} + u_t$$

**gen Dinflation=D1.inflation**

**reg Dinflation L1.Dinflation**

Linear regression

Number of obs = 172  
F( 1, 170) = 6.08  
Prob > F = 0.0146  
R-squared = 0.0564  
Root MSE = 1.664

d_inflation	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
d_inflation L1.	-.2380471	.0965017	-2.47	0.015	-.4285431	-.047551
_cons	.0171008	.1268849	0.13	0.893	-.2333721	.2675736

In the 4<sup>th</sup> quarter in 2004 inflation increased by 1.7.

- Forecast how inflation will change in the next quarter.

In the 1<sup>st</sup> quarter 2005 it reduced by 1.14.

- How big is your forecasting error?

## AR(1): Example (cont.)

**reg** Dinflation L1.Dinflation L2.Dinflation L3.Dinflation  
**L4.Dinflation**

Linear regression

Number of obs = 172  
F( 4, 167) = 7.93  
Prob > F = 0.0000  
R-squared = 0.2038  
Root MSE = 1.5421

d_inflation	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
d_inflation						
L1.	-.2579426	.0925934	-2.79	0.006	-.4407471	-.0751381
L2.	-.3220312	.0805465	-4.00	0.000	-.4810518	-.1630106
L3.	.1576089	.0841017	1.87	0.063	-.0084307	.3236484
L4.	-.0302511	.0930471	-0.33	0.746	-.2139512	.153449
_cons	.0224294	.1176344	0.19	0.849	-.2098127	.2546715

## AR(1): Example (cont.)

Is the AR(4) model better than the AR(1) model?

```
. test L2.d_inflation=L3.d_inflation=L4.d_inflation=0
( 1)  L2.d_inflation - L3.d_inflation = 0
( 2)  L2.d_inflation - L4.d_inflation = 0
( 3)  L2.d_inflation = 0

F( 3, 167) = 6.71
Prob > F = 0.0003
```

How should we choose the lag length?

- One approach is to start with a model with many lags and to perform a hypothesis test on the final lag.
- Delete the final lag if insignificant and perform an hypothesis test on the new final lag,..., continue until all included lags are significant.

## AR(1): Example (cont.)

- Drawback of this approach is that it can produce too large a model.
  - At a 5% significance level: if the true lag length is 5 it will estimate  $p$  to be 6 in 5% of the time.

Alternative test: Akaike information criterion (AIC)

In Stata: **estat ic**

## Autoregressive distributed lag model (ADL)

- Economic theory often suggests other variables could help to forecast the variable of interest.
- When we add other variables and their lags, we have an autoregressive distributed lag model.
- Form?

## ADL: Example

When predicting future changes in inflation economic theory suggests that lagged value of unemployment might be a good predictor (Remember the Philips curve?).

**reg Dinflation L1.Dinflation L2.Dinflation L1.unemployment  
L2.unemployment**

Linear regression

Number of obs = 172  
F( 4, 167) = 15.41  
Prob > F = 0.0000  
R-squared = 0.3514  
Root MSE = 1.3918

d_inflation	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
d_inflation						
L1.	-.4685035	.0771115	-6.08	0.000	-.6207425	-.3162645
L2.	-.4251441	.0821394	-5.18	0.000	-.5873096	-.2629787
unemployment_rate						
L1.	-2.243865	.4020402	-5.58	0.000	-3.037602	-1.450129
L2.	2.044221	.3875693	5.27	0.000	1.279054	2.809388
_cons	1.193032	.4344711	2.75	0.007	.3352683	2.050796



## Granger 'causality' test

- Do the included lags of unemployment have useful predictive content conditional on the included lags of the change in inflation?
- The claim that a variable has no predictive content corresponds to the null hypothesis that the coefficients on all lags of the variable are zero.
- The  $F$ -statistic of this test is called the **Granger causality statistic**.
- If the null hypothesis is rejected the variable  $x$  is said to Granger-cause the dependent variable  $y$ .
- This does not mean that we have estimated the causal effect of  $x$  on  $y$ !
- It means that  $x$  is a useful predictor of  $y$  (Granger predictability would be a better term).

## Granger 'causality' test: Example

reg Dinflation L1.Dinflation L2.Dinflation L1.unemployment  
L2.unemployment

d_inflation	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
d_inflation						
L1.	-.4685035	.0771115	-6.08	0.000	-.6207425	-.3162645
L2.	-.4251441	.0821394	-5.18	0.000	-.5873096	-.2629787
unemployment_rate						
L1.	-2.243865	.4020402	-5.58	0.000	-3.037602	-1.450129
L2.	2.044221	.3875693	5.27	0.000	1.279054	2.809388
_cons	1.193032	.4344711	2.75	0.007	.3352683	2.050796

```
. test L1.unemployment=L2.unemployment=0
( 1)  L.unemployment_rate - L2.unemployment_rate = 0
( 2)  L.unemployment_rate = 0
      F( 2, 167) =    16.13
      Prob > F =    0.0000
```

- Null hypothesis that coefficients on the 2 lags of unemployment are zero is rejected at a 1% level.
- Unemployment is a useful predictor for the change in the inflation rate.

# Trends

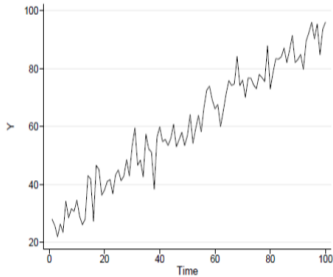
- A time series  $y_t$  is stationary if its probability distribution does not change over time.
- If a time series has a trend, it is non-stationary.
- A trend is a persistent long-term movement of a variable over time.

We consider two types of trends:

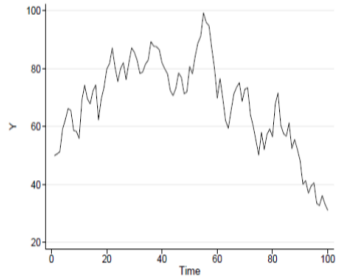
- 1 Deterministic trend:  $y_t = \beta_0 + \lambda t + u_t$ 
  - Series is a non-random function of time.
- 2 Stochastic trend:  $y_t = \beta_0 + y_{t-1} + u_t$ 
  - Series is a random function of time.  $\Rightarrow$  random walk model

## Trends (cont.)

Deterministic trend



Stochastic trend



## Detecting stochastic trends: Dickey-Fuller test for unit root

Test for a unit root in Chilean inflation (Note: we test for stochastic trend in *inflation* and not  $\Delta inflation$ )

**reg Dinflation L1.inflation**

d_inflation	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
inflation L1.	-.1643553	.0415311	-3.96	0.000	-.2463383	-.0823722
_cons	.7219345	.217599	3.32	0.001	.2923904	1.151479

**dfuller**

## Dickey-Fuller test for unit root (cont.)

- Under  $H_0$ ,  $y_t$  is non-stationary and the DF-statistic has a non-normal distribution, we therefore use the following critical values:

	10%	5%	1%
Intercept only	-2.57	-2.86	-3.43

- In our case the DF-statistic is  $-3.96$ .
- $DF = -3.96$  is more negative than  $-2.86$ , so we reject the null hypothesis of a stochastic trend at a 5% significance level.

## Dickey-Fuller test for unit root (cont.)

- Under  $H_0$ ,  $y_t$  is non-stationary and the DF-statistic has a non-normal distribution, we therefore use the following critical values:

	10%	5%	1%
Intercept only	-2.57	-2.86	-3.43

- In our case the DF-statistic is  $-3.96$ .
- $DF = -3.96$  is more negative than  $-2.86$ , so we reject the null hypothesis of a stochastic trend at a 5% significance level.

## Dickey-Fuller test for unit root (cont.)

DF test for a unit root using 4 lags (AR(4)-model)

d_inflation	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
inflation						
L1.	-.1134169	.0422344	-2.69	0.008	-.1968029	-.030031
d inflation						
L1.	-.1864426	.0805144	-2.32	0.022	-.3454068	-.0274783
L2.	-.2563879	.081463	-3.15	0.002	-.4172251	-.0955507
L3.	.1990491	.0793514	2.51	0.013	.0423811	.3557171
L4.	.0099994	.0779921	0.13	0.898	-.1439849	.1639837
_cons	.5068158	.2141807	2.37	0.019	.0839466	.9296851

- In this case the DF-statistic is  $-2.69$ .
- So?
- $DF = -2.69$  is less negative than  $-2.86$ , so we do not reject the null hypothesis of a stochastic trend at a 5% significance level.



## Dickey-Fuller test for unit root (cont.)

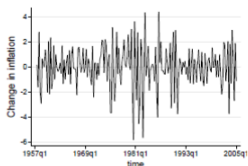
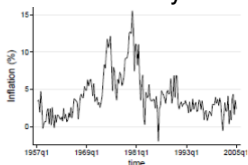
DF test for a unit root using 4 lags (AR(4)-model)

d_inflation	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
inflation						
L1.	-.1134169	.0422344	-2.69	0.008	-.1968029	-.030031
d inflation						
L1.	-.1864426	.0805144	-2.32	0.022	-.3454068	-.0274783
L2.	-.2563879	.081463	-3.15	0.002	-.4172251	-.0955507
L3.	.1990491	.0793514	2.51	0.013	.0423811	.3557171
L4.	.0099994	.0779921	0.13	0.898	-.1439849	.1639837
_cons	.5068158	.2141807	2.37	0.019	.0839466	.9296851

- In this case the DF-statistic is  $-2.69$ .
- So?
- $DF = -2.69$  is less negative than  $-2.86$ , so we do not reject the null hypothesis of a stochastic trend at a 5% significance level.

## Avoiding problems caused by stochastic trends

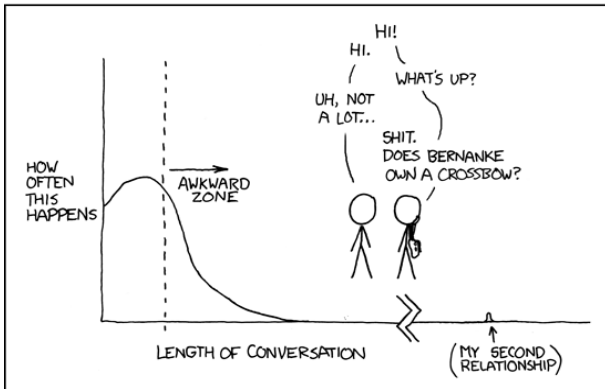
- Best way to deal with a trend is to transform the series such that it does not have a trend.
- If a series,  $y_t$  does have a trend, then the first difference does not have a trend.
- That is why we started the lecture by looking at  $\Delta inflation$ .



## Review questions

- 1 Decide if you agree or disagree with the following statements and give a brief explanation of your decision:
  - (a) Like cross-sectional observations, we can assume that most time series observations are independently distributed.
  - (b) A trending variable cannot be used as the dependent variable in multiple regression analysis.
  - (c) Seasonality is an issue when using annual time series observations.
- 2 Suppose you have quarterly data on new housings, interest rates, and real per capita income in Brazil. Specify a model for housing starts that accounts for possible trends and seasonality in the variables.

SOMETIMES MY CONVERSATIONS WITH STRANGERS GO ON FOR A WHILE BEFORE I REALIZE THEY'RE TALKING ON THEIR PHONES.



# Multiple Regression Analysis: Panel Data

## What is panel data?

- Often loosely use the term panel data to refer to any data set that has both a cross-sectional dimension and a time-series dimension.
- More precisely it is only data following the same cross-section units over time.
- Otherwise it is a pooled cross-section.

## What is panel data?

- Often loosely use the term panel data to refer to any data set that has both a cross-sectional dimension and a time-series dimension.
- More precisely it is only data following the same cross-section units over time.
- Otherwise it is a pooled cross-section.

# What is panel data? (cont.)

Data Editor (Edit) - [SriLanka\_9198.dta]

File Edit View Data Tools

nh[1] 11054

nh	year	v1111d	thanaid	agehead	sexhead	educhead
1	11054	0	1	72	1	0
2	11054	1	1	79	1	0
3	11061	0	1	35	1	5
4	11061	1	1	43	1	6
5	11081	0	1	54	1	3
6	11081	1	1	52	0	0
7	11101	0	1	44	1	5
8	11101	1	1	48	1	0
9	12021	0	2	28	1	8
10	12021	1	2	35	1	10
11	12035	0	2	25	1	8
12	12035	1	2	33	1	5
13	12051	0	2	63	1	6
14	12051	1	2	26	1	2
15	12054	0	2	27	1	5
16	12054	1	2	38	1	5
17	12081	0	2	26	1	0
18	12081	1	2	35	1	0
19	12121	0	2	43	1	3
20	12121	1	2	48	1	0
21	13014	0	3	38	1	7
22	13014	1	3	45	1	9
23	13015	0	3	21	1	0
24	13015	1	3	35	1	0
25	13021	0	3	45	1	10
26	13021	1	3	55	1	10
27	13025	0	3	41	1	4
28	13025	1	3	55	1	1
29	13035	0	3	37	1	0
30	13035	1	3	40	1	5
31	13041	0	3	47	1	0
32	13041	1	3	68	1	0
33	13054	0	3	60	1	5
34	13054	1	3	28	1	0

Ready Vars: 23 Order: Dataset Obs: 1652



## Pooled cross section

- We may want to pool cross sections just to get bigger sample sizes.
- We may want to pool cross sections to investigate the effect of time.
- We may want to pool cross sections to investigate whether relationships have changed over time.

## Pooled cross section (cont.)

Data Editor (Edit) - [Malawi.dta]

File Edit View Data Tools

caseid[1] 467 9 2

caseid	mtdx	v000	v001	v003	v007	v190
1	2	467	9	2	2004	poorer
2	1	47927	1	1	2004	middle
3	2	2624	2	1	2004	richer
4	1	8640	1	2	2004	richer
5	2	46718	2	2	2004	richer
6	2	100514	1	1	2000	.
7	2	10016	2	2	2000	.
8	2	100428	1	1	2000	.
9	2	10058	2	2	2000	.
10	2	100249	1	1	2000	.
11	2	100112	2	2	2000	.
12	2	100241	1	1	2000	.
13	2	2626	2	1	2004	richer
14	2	100213	2	3	2000	.
15	2	2634	2	2	2004	richer
16	2	10059	2	1	2000	.
17	2	4796	1	1	2004	richer
18	2	4834	2	2	2004	poorest
19	2	46725	1	1	2004	richer
20	2	47935	1	1	2004	richer
21	2	100211	1	1	2000	.
22	2	10041	1	1	2000	.
23	2	46736	1	1	2004	richer
24	3	2626	1	1	2004	richer
25	3	46725	2	3	2004	richer
26	2	10049	2	2	2000	.
27	2	46734	1	1	2004	richest
28	2	4677	1	1	2004	middle
29	2	46711	1	1	2004	richer
30	4	100232	1	1	2000	.
31	1	46722	1	1	2004	richest
32	2	46744	3	3	2004	middle
33	2	47922	1	1	2004	poorest
34	2	100225	1	1	2000	.

Ready Length: 15 Vars: 23 Order: Dataset Obs: 22,840

## Difference-in-Difference (DID)

- Take a random assignment into treatment and control groups, like in a medical experiment.
- One can simply compare the change in outcomes across the treatment and control groups to estimate the treatment effect.
- If we have two time periods (1, 2) and two groups (A, B) the DID is:

$$(y_{2,B} - y_{2,A}) - (y_{1,B} - y_{1,A}) = (y_{2,B} - y_{1,B}) - (y_{2,A} - y_{1,A}) \quad (1)$$

- A regression framework using time and treatment dummy variables can calculate this DID as well.
- Consider the model:

$$y_{it} = \beta_0 + \beta_1 \text{treat}_{it} + \beta_2 \text{after}_{it} + \beta_3 \text{treat}_{it} * \text{after}_{it} + u_{it} \quad (2)$$

- The estimated  $\beta_3$  will be the DID in group means.

## DID (cont.)

- When we do not have a truly random assignment the regression form becomes useful.
- Additional  $x$ es can be added to the regression to control for differences across the treatment and control groups.
- Sometimes referred to as a ‘natural’ experiment especially when a policy change is being analysed.

## DID: Example

You are interested in analysing how house prices are affected by an incinerator built in a 5 km vicinity. You have data collected in 1978 before the incinerator was built and in 1981 just after the incinerator was inaugurated.

You are running the following regression:

```
reg rprice nearinc if year==1981
```

Source	SS	df	MS
Model	2.7059e+10	1	2.7059e+10
Residual	1.3661e+11	140	975815048
Total	1.6367e+11	141	1.1608e+09

Number of obs = 142  
F( 1, 140) = 27.73  
Prob > F = 0.0000  
R-squared = 0.1653  
Adj R-squared = 0.1594  
Root MSE = 31238

rprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
nearinc	-30688.27	5827.709	-5.27	0.000	-42209.97 -19166.58
_cons	101307.5	3093.027	32.75	0.000	95192.43 107422.6

- What do you conclude?
- Are you sure that is a 'good' conclusion?

## DID: Example

You are interested in analysing how house prices are affected by an incinerator built in a 5 km vicinity. You have data collected in 1978 before the incinerator was built and in 1981 just after the incinerator was inaugurated.

You are running the following regression:

```
reg rprice nearinc if year==1981
```

Source	SS	df	MS			
Model	2.7059e+10	1	2.7059e+10	Number of obs =	142	
Residual	1.3661e+11	140	975815048	F( 1, 140) =	27.73	
Total	1.6367e+11	141	1.1608e+09	Prob > F =	0.0000	
				R-squared =	0.1653	
				Adj R-squared =	0.1594	
				Root MSE =	31238	

rprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
nearinc	-30688.27	5827.709	-5.27	0.000	-42209.97 -19166.58
_cons	101307.5	3093.027	32.75	0.000	95192.43 107422.6

- What do you conclude?
- Are you sure that is a 'good' conclusion?

## DID: Example

You are interested in analysing how house prices are affected by an incinerator built in a 5 km vicinity. You have data collected in 1978 before the incinerator was built and in 1981 just after the incinerator was inaugurated.

You are running the following regression:

```
reg rprice nearinc if year==1981
```

Source	SS	df	MS	
Model	2.7059e+10	1	2.7059e+10	Number of obs = 142
Residual	1.3661e+11	140	975815048	F( 1, 140) = 27.73
Total	1.6367e+11	141	1.1608e+09	Prob > F = 0.0000
				R-squared = 0.1653
				Adj R-squared = 0.1594
				Root MSE = 31238

rprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
nearinc	-30688.27	5827.709	-5.27	0.000	-42209.97 -19166.58
_cons	101307.5	3093.027	32.75	0.000	95192.43 107422.6

- What do you conclude?
- Are you sure that is a 'good' conclusion?

## DID: Example (cont.)

Here is the relationship in 1978:

```
reg rprice nearinc if year==1978
```

Source	SS	df	MS	
Model	1.3636e+10	1	1.3636e+10	Number of obs = 179
Residual	1.5332e+11	177	866239953	F( 1, 177) = 15.74
Total	1.6696e+11	178	937979126	Prob > F = 0.0001
				R-squared = 0.0817
				Adj R-squared = 0.0765
				Root MSE = 29432

rprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
nearinc	-18824.37	4744.594	-3.97	0.000	-28187.62	-9461.117
_cons	82517.23	2653.79	31.09	0.000	77280.09	87754.37

- What do you conclude, now?
- What is the DID estimate?



## DID: Example (cont.)

Here is the relationship in 1978:

```
reg rprice nearinc if year==1978
```

Source	SS	df	MS	
Model	1.3636e+10	1	1.3636e+10	Number of obs = 179
Residual	1.5332e+11	177	866239953	F( 1, 177) = 15.74
Total	1.6696e+11	178	937979126	Prob > F = 0.0001
				R-squared = 0.0817
				Adj R-squared = 0.0765
				Root MSE = 29432

rprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
nearinc	-18824.37	4744.594	-3.97	0.000	-28187.62	-9461.117
_cons	82517.23	2653.79	31.09	0.000	77280.09	87754.37

- What do you conclude, now?
- What is the DID estimate?

## DID: Example (cont.)

Check:

```
reg rprice nearinc y81 y81nrinc
```

Source	SS	df	MS
Model	6.1055e+10	3	2.0352e+10
Residual	2.8994e+11	317	914632739
Total	3.5099e+11	320	1.0969e+09

Number of obs = 321  
F( 3, 317) = 22.25  
Prob > F = 0.0000  
R-squared = 0.1739  
Adj R-squared = 0.1661  
Root MSE = 30243

rprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
nearinc	-18824.37	4875.322	-3.86	0.000	-28416.45	-9232.293
y81	18790.29	4050.065	4.64	0.000	10821.88	26758.69
y81nrinc	-11863.9	7456.646	-1.59	0.113	-26534.67	2806.867
_cons	82517.23	2726.91	30.26	0.000	77152.1	87882.36

- And, what do you observe?

## DID: Example (cont.)

Check:

```
reg rprice nearinc y81 y81nrinc
```

Source	SS	df	MS
Model	6.1055e+10	3	2.0352e+10
Residual	2.8994e+11	317	914632739
Total	3.5099e+11	320	1.0969e+09

Number of obs = 321  
F( 3, 317) = 22.25  
Prob > F = 0.0000  
R-squared = 0.1739  
Adj R-squared = 0.1661  
Root MSE = 30243

rprice	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
nearinc	-18824.37	4875.322	-3.86	0.000	-28416.45	-9232.293
y81	18790.29	4050.065	4.64	0.000	10821.88	26758.69
y81nrinc	-11863.9	7456.646	-1.59	0.113	-26534.67	2806.867
_cons	82517.23	2726.91	30.26	0.000	77152.1	87882.36

- And, what do you observe?

## Two-period panel data

- It's possible to use a panel just like pooled cross-sections, but with a panel you do more than that.
- Panel data can be used to address some kinds of omitted variable bias.
- If we can think of the omitted variables as being fixed over time, then we can represent this in a model as a composite error.

## Two-period panel data

- It's possible to use a panel just like pooled cross-sections, but with a panel you do more than that.
- Panel data can be used to address some kinds of omitted variable bias.
- If we can think of the omitted variables as being fixed over time, then we can represent this in a model as a composite error.

## Unobserved fixed effects

- Suppose the population model is

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + \dots + \beta_k x_{itk} + a_i + u_{it}$$

- Here, we have added a time-constant component of the error,  
 $v_{it} = a_i + u_{it}$
- if  $a_i$  is correlated with the  $x$ es, OLS will be biased, since  $a_i$  is part of the error term.
- With panel data we can difference-out the unobserved fixed effect.
- In this case we treat  $a_i$  as an additional regressor.

## First differences

- We can subtract one period from the other, to obtain
$$\Delta y_i = \delta_0 + \beta_1 \Delta x_{i1} + \dots + \beta_k \Delta x_{ik} + \Delta u_i$$
- This model has no correlation between the  $x$ es and the error term, so no bias.
- Note: Need to be careful about the organisation of your data to be sure to compute the correct change.

## Differencing with multiple periods

- We can extend this method to more periods.
- Simply difference adjacent periods.
- So if 3 periods, then subtract period 1 from period 2, period 2 from period 3 and have 2 observations per individual.
- Simply estimate by OLS, assuming the  $\Delta u_i$  are uncorrelated over time.



## Fixed effects estimation

- When there is an observed fixed effect, an alternative to first differences is fixed effects estimation.
- Consider the average over time of
$$y_{it} = \beta_1 x_{it1} + \dots + \beta_k x_{itk} + a_i + u_{it}$$
- The average of  $a_i$  will be  $a_i$ , so if you subtract the mean,  $a_i$  will be differenced out just as when doing first differences.
- This method is identical to including a separate intercept or dummy for every individual, which can become a bit tedious...
- ... and we have to be careful with the degrees of freedom...
- Luckily, Stata will do fixed effects estimation for you

⇒ Stata command: **xtreg y x, fe**

## First differences vs. fixed effects

- First differences and fixed effects will be exactly the same when  $t = 2$ .
- For  $t > 2$ , the two methods are different.
- Probably, you will see fixed effects estimation more often than differences - probably more because it is easier not better.
- Fixed effects can be used not only for balanced panels but can also be easily implemented for unbalanced panels.

## Random effects

- Start with the same basic model with a composite error, i.e.  
$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it1} + \dots + \beta_k x_{itk} + a_i + u_{it}$$
- Previously, we have assumed that  $a_i$  was correlated with the  $x$ es, but what if it is not?
- OLS would be consistent in this case, but the composite error will be serially correlated.
- Stata will estimate the random effects model for you.  
⇒ Stata command: **xtreg y x, re**

## Fixed or random effects

- It is more common to use fixed effects, since one of the most frequent issues is that something unobserved is correlated with the  $x$ es.
- If truly need to use random effects, the only problem is the standard errors.
- Can simply just adjust the standard errors for correlation within group.

## Pooled OLS, fixed and random effects: Comparison

Dependent Variable: $\log(wage)$			
Independent Variables	Pooled OLS	Random Effects	Fixed Effects
<i>educ</i>	.091 (.005)	.092 (.011)	————
<i>black</i>	-.139 (.024)	-.139 (.048)	————
<i>hispan</i>	.016 (.021)	.022 (.043)	————
<i>exper</i>	.067 (.014)	.106 (.015)	————
<i>exper</i> <sup>2</sup>	-.0024 (.0008)	-.0047 (.0007)	-.0052 (.0007)
<i>married</i>	.108 (.016)	.064 (.017)	.047 (.018)
<i>union</i>	.182 (.017)	.106 (.018)	.080 (.019)

- Write down the regression function(s) underlying the estimations.
- How do you interpret the coefficients and what do you conclude? Which model would do you prefer and why? Explain.

## Pooled OLS, fixed and random effects: Comparison

Dependent Variable: $\log(wage)$			
Independent Variables	Pooled OLS	Random Effects	Fixed Effects
<i>educ</i>	.091 (.005)	.092 (.011)	————
<i>black</i>	-.139 (.024)	-.139 (.048)	————
<i>hispan</i>	.016 (.021)	.022 (.043)	————
<i>exper</i>	.067 (.014)	.106 (.015)	————
<i>exper</i> <sup>2</sup>	-.0024 (.0008)	-.0047 (.0007)	-.0052 (.0007)
<i>married</i>	.108 (.016)	.064 (.017)	.047 (.018)
<i>union</i>	.182 (.017)	.106 (.018)	.080 (.019)

- Write down the regression function(s) underlying the estimations.
- How do you interpret the coefficients and what do you conclude? Which model would do you prefer and why? Explain.

## Pooled OLS, fixed and random effects: Comparison

Dependent Variable: $\log(wage)$			
Independent Variables	Pooled OLS	Random Effects	Fixed Effects
<i>educ</i>	.091 (.005)	.092 (.011)	————
<i>black</i>	-.139 (.024)	-.139 (.048)	————
<i>hispan</i>	.016 (.021)	.022 (.043)	————
<i>exper</i>	.067 (.014)	.106 (.015)	————
<i>exper</i> <sup>2</sup>	-.0024 (.0008)	-.0047 (.0007)	-.0052 (.0007)
<i>married</i>	.108 (.016)	.064 (.017)	.047 (.018)
<i>union</i>	.182 (.017)	.106 (.018)	.080 (.019)

- Write down the regression function(s) underlying the estimations.
- How do you interpret the coefficients and what do you conclude? Which model would do you prefer and why? Explain.

## Other uses of panel methods

- It's possible to think of models where there is an unobserved fixed effect, even if we do not have true panel data.
- A common example is where we think there is an unobserved family effect.
- We can difference siblings ( $\Rightarrow$  sibling fixed effects).
- We can also estimate family fixed effects models.



## Additional issues

- Many of the things we already know about both cross section and time series data can be applied to panel data.
- Can test and correct for serial correlation in the errors. **How?**
- Can test and correct for heteroskedasticity. **How?**
- Can estimate standard errors robust to both. **How?**

## Review questions

- 1 Why can we not use first differences when we have independent cross sections in two years (as opposed to panel data)?
- 2 Suppose that we want to estimate the effect of several variables on annual saving and that we have a panel data set on individuals collected on January 31, 1990, and January 31, 1992. If we include a year dummy for 1992 and use first differencing, can we also include age in the original model? Explain.
- 3 In order to determine the effect of doping on athletic performance (focus on running for now), you collect data on participants of the Golden League Meetings in 2010, 2012, and 2014.
  - a) What measures of athletic success would you include in an equation? Are there some timing issues?

## Review questions (cont.)

- b) What other measures might you control for in the equation?
- c) Write down an equation that allows you to estimate the effect of athletic success on the percentage change in substance use. How would you estimate this equation? Why would you choose this method.

WHAT ARE YOU DOING?



IMPORTANT SCIENTIFIC RESEARCH!!



DO...OR  
DO NOT.



# Identification: Specification and Data Problems

## Recap: Functional form

- We have seen that a linear regression can fit non-linear relationships:
  - We can use logs on the right hand side, the left hand side or both.
  - We can use quadratic forms of  $x$ .
  - We can use interactions of  $x$ .
- How do we know if we've chosen the right functional form of the model?
  - 1 Economic theory should guide you.
  - 2 Think about the interpretation:
    - Does it make more sense for  $x$  to affect  $y$  in percentage (use logs) or absolute terms?
    - Does it make for sense for the derivative  $y'(x_1)$  to vary with  $x_1$  (quadratic) or with  $x_2$  (interactions) or to be fixed?

## Recap: Functional form

- We have seen that a linear regression can fit non-linear relationships:
  - We can use logs on the right hand side, the left hand side or both.
  - We can use quadratic forms of  $x$ .
  - We can use interactions of  $x$ .
- How do we know if we've chosen the right functional form of the model?
  - 1 Economic theory should guide you.
  - 2 Think about the interpretation:
    - Does it make more sense for  $x$  to affect  $y$  in percentage (use logs) or absolute terms?
    - Does it make for sense for the derivative  $y(x_1)$  to vary with  $x_1$  (quadratic) or with  $x_2$  (interactions) or to be fixed?

## Recap: Functional form

- We have seen that a linear regression can fit non-linear relationships:
  - We can use logs on the right hand side, the left hand side or both.
  - We can use quadratic forms of  $x$ .
  - We can use interactions of  $x$ .
- How do we know if we've chosen the right functional form of the model?
  - 1 Economic theory should guide you.
  - 2 Think about the interpretation:
    - Does it make more sense for  $x$  to affect  $y$  in percentage (use logs) or absolute terms?
    - Does it make for sense for the derivative  $y'(x_1)$  to vary with  $x_1$  (quadratic) or with  $x_2$  (interactions) or to be fixed?



## Recap: Functional form

- We have seen that a linear regression can fit non-linear relationships:
  - We can use logs on the right hand side, the left hand side or both.
  - We can use quadratic forms of  $x$ .
  - We can use interactions of  $x$ .
- How do we know if we've chosen the right functional form of the model?
  - 1 Economic theory should guide you.
  - 2 Think about the interpretation:
    - Does it make more sense for  $x$  to affect  $y$  in percentage (use logs) or absolute terms?
    - Does it make for sense for the derivative  $y'(x_1)$  to vary with  $x_1$  (quadratic) or with  $x_2$  (interactions) or to be fixed?

## Recap: Functional form

- We have seen that a linear regression can fit non-linear relationships:
  - We can use logs on the right hand side, the left hand side or both.
  - We can use quadratic forms of  $x$ .
  - We can use interactions of  $x$ .
- How do we know if we've chosen the right functional form of the model?
  - 1 Economic theory should guide you.
  - 2 Think about the interpretation:
    - Does it make more sense for  $x$  to affect  $y$  in percentage (use logs) or absolute terms?
    - Does it make for sense for the derivative  $y(x_1)$  to vary with  $x_1$  (quadratic) or with  $x_2$  (interactions) or to be fixed?

## Proxy variables

- What if the model is misspecified because no data is available on an important  $x$  variable?
- It may be possible to avoid omitted variable bias by using a proxy variable.
- A proxy variable must be related to the unobservable variable, e.g.  $x_3^* = \delta_3 x_3 + v_3$ , where  $*$  implies unobserved.
- Now, suppose we just substitute  $x_3$  for  $x_3^*$ .
- What do we need for this solution to give us consistent estimates of  $\beta_1$  and  $\beta_2$ ?
- $E(x_3^* | x_1, x_2, x_3) = E(x_3^*) = \delta_0 + \delta_3 x_3$
- That is  $u$  is uncorrelated with  $x_1, x_2$  and  $x_3^*$  and  $v_3$  is uncorrelated with  $x_1, x_2$  and  $x_3$ .

## Proxy variables

- What if the model is misspecified because no data is available on an important  $x$  variable?
- It may be possible to avoid omitted variable bias by using a proxy variable.
- A proxy variable must be related to the unobservable variable, e.g.  $x_3^* = \delta_3 x_3 + v_3$ , where  $*$  implies unobserved.
- Now, suppose we just substitute  $x_3$  for  $x_3^*$ .
- What do we need for this solution to give us consistent estimates of  $\beta_1$  and  $\beta_2$ ?
- $E(x_3^* | x_1, x_2, x_3) = E(x_3^*) = \delta_0 + \delta_3 x_3$
- That is  $u$  is uncorrelated with  $x_1, x_2$  and  $x_3^*$  and  $v_3$  is uncorrelated with  $x_1, x_2$  and  $x_3$ .

## Proxy variables (cont.)

- So what you would run is:  
$$y = (\beta_0 + \beta_3\delta_0) + \beta_1x_1 + \beta_2x_2 + \beta_3\delta_3x_3 + (u + \beta_3v_3)$$
- So this redefines the intercept, error term and the coefficient on  $x_3$  but  $\beta_1$  and  $\beta_2$  remain unaffected.
- Without these assumption you would have biased estimates.
- The direction of the bias will depend on the sign of  $\beta_3$  and  $\delta_j$ .
- With the proxy, there may still be bias but it might be smaller than omitted variable bias though.

## Lagged Dependent Variables

- What if there are unobserved variables, and you cannot find reasonable proxies?
- It may be possible to include a lagged dependent variable to account for omitted variables that contribute to both past and current levels of  $y$ .
- Obviously, you must think past and current  $y$  are related for this to make sense.

## Lagged Dependent Variables

- What if there are unobserved variables, and you cannot find reasonable proxies?
- It may be possible to include a lagged dependent variable to account for omitted variables that contribute to both past and current levels of  $y$ .
- Obviously, you must think past and current  $y$  are related for this to make sense.

## Measurement error

- Sometimes we have the variable we want, but we think it is measured with error.
- **Examples:** A survey asks how many hours you worked over the last year, or how many weeks of child care you used when your child was young.
- Measurement error in the dependent variable ( $y$ ), is different from measurement error in the independent variable ( $x$ ).
- Measurement error in  $y$ : The intercept ( $\beta_0$ ) will be biased.
- Measurement error in  $x$ :
  - If there is no correlation between the error and the observed variable ( $x_1$ ) then the OLS estimate will be unbiased **but** the variances will be larger.
  - If there is correlation between the error and the observed variable ( $x_1$ ) then the OLS estimate will be biased.



## Measurement error

- Sometimes we have the variable we want, but we think it is measured with error.
- **Examples:** A survey asks how many hours you worked over the last year, or how many weeks of child care you used when your child was young.
- Measurement error in the dependent variable ( $y$ ), is different from measurement error in the independent variable ( $x$ ).
- Measurement error in  $y$ : The intercept ( $\beta_0$ ) will be biased.
- Measurement error in  $x$ :
  - If there is no correlation between the error and the observed variable ( $x_1$ ) then the OLS estimate will be unbiased **but** the variances will be larger.
  - If there is correlation between the error and the observed variable ( $x_1$ ) then the OLS estimate will be biased.

## Missing data: Is it a problem?

- If any observation has missing data on one of the variables in the model, it can't be used.
- If data is missing at random, using a sample restricted to observations with no missing values will be fine.
- A problem can arise if the data is missing systematically - say high income individuals refuse to provide income data.

## Non-random samples

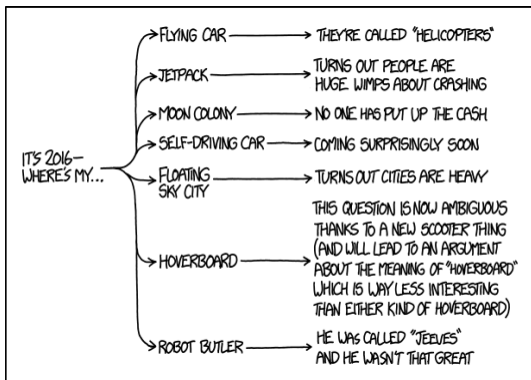
- If the sample is chosen on the basis of an  $x$  variable (exclusively), then estimates may be unbiased.
- But, if the sample is chosen on the basis of the  $y$  variable, then we have sample selection bias.
- Sample selection can be more subtle. Say looking at wages for workers - since people choose to work this isn't the same as wage offers.

# Outliers

- Sometimes an individual observation can be very different from the others, and can have a large effect on the outcome.
- Distinguish outliers in  $y$  from those in  $x$  direction (here multidimensional, problematic).
- Sometimes an outlier will occur due to errors in data entry - one reason why looking at summary statistics is important.
- Sometimes the observation will just truly be very different.
- Not unreasonable to fix observations where it's clear there was just an extra zero entered or left off, etc.
- Not unreasonable to drop observations that appear to be extreme.
- It depends on whether they are influential or not.
- Can use Stata to investigate outliers. (Remember how?)

## Review questions

- 1 The regression model includes a random error or disturbance term for a variety of reasons. Which of the following is NOT one of them?
- a) Measurement errors in the observed variables
  - b) Omitted influences on Y (other than X)
  - c) Linear functional form is only an approximation
  - d) The observable variables do not exactly correspond with their theoretical counterparts
  - e) There may be approximation errors in the calculation of the least squares estimates



# Limited Dependent Variables

## Binary dependent variables

- Recall the linear probability model (LPM), which can be written as

$$P(y = 1|x) = \beta_0 + x\beta \quad (1)$$

- A drawback of the linear probability model is that predicted values are not constrained to be between 0 and 1.
- An alternative is to model the probability as a function,

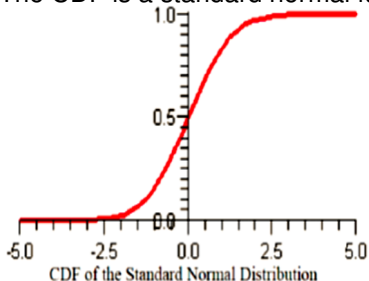
$$G(\beta_0 + x\beta) \quad (2)$$

with  $0 < G(z) < 1$ .



## The probit model

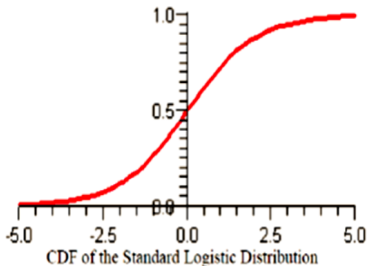
- One choice for  $G(z)$  is the standard normal cumulative distribution function (CDF).
- The CDF is a standard normal function.



- Since the model is non-linear, it cannot be estimated by using our usual methods.  
⇒ We use **maximum likelihood estimation (MLE)**.

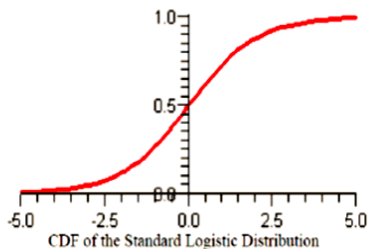
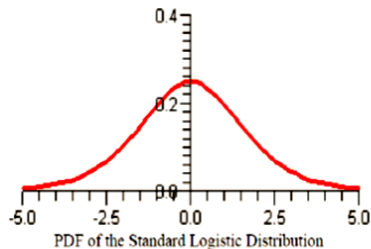
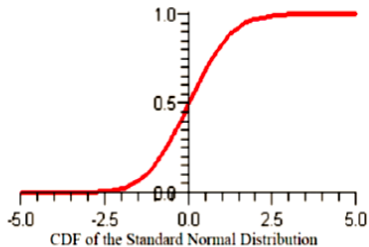
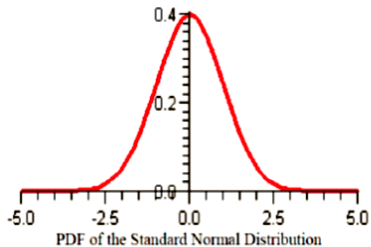
## The logit model

- Another choice for  $G(z)$  is the logistic function, which is the CDF for a standard logistic random variable.



- Both functions have similar shapes, they are increasing in  $z$ , most quickly around 0.
- The logit model is sometimes also referred to as logistic regression.

## Graphically: Probit and logit



## Probits and logits

- Both, the probit and logit are nonlinear and require maximum likelihood estimation.
- There is no real reason to prefer one over the other.
- Traditionally we saw more of the logit, mainly because the logistic function leads to a more easily computed model.
- Today, probit is easy to compute with standard packages, so more popular.

## Interpretation of probits and logits (vs. the LPM)

- In general, we care about the effect of  $x$  on  $P(y = 1|x)$ , that is, we care about  $\frac{\delta p}{\delta x}$ .
- For the linear case, i.e. when using the LPM, this is easily computed as the coefficient on  $x$ .
- For the non-linear probit and logit models, it's more complicated:

$$\frac{\delta p}{\delta x_j} = g(\beta_0 + x\beta)\beta_j \quad (3)$$

- So, you cannot just compare the coefficients across the three models.

## Interpretation of probits and logits (vs. the LPM) (cont.)

- You can compare the sign and significance of the coefficients, though.
- To compare the magnitude of effects, you need to calculate the derivatives, i.e. the marginal effects.
- Stata will do that for you.  
Command: **margins dydx**

## Example

You study the factors influencing the purchase of health insurance. The dependent variable is binary indicating whether the person has health insurance or not. You estimate three different models with the following results:

Have health insurance	Regression coefficient	Logit coefficient	Probit coefficient
Retired (=1)	0.04*	0.19*	0.11*
Age (yrs.)	-0.002	-0.01	-0.008
Good health status (=1)	0.06*	0.31*	0.19*
HH income (USD)	0.0004*	0.002*	0.001*
Education (yrs.)	0.02*	0.11*	0.07*
Married (=1)	0.12*	0.57*	0.36*
Hispanic (=1)	-0.12*	-0.81*	-0.46*
Constant	0.12	-1.71*	-1.06*
<i>R-squared</i>	<i>0.08</i>	<i>0.07</i>	<i>0.07</i>
<i>N</i>	<i>12,367</i>	<i>12,367</i>	<i>12,367</i>

Note: \* indicates significance at 5% level.

**Question:** How do you interpret the results?



## Example (cont.)

**Question:** How do you interpret the results?

Have health insurance	Regression		
	coefficient	ME (Logit)	ME (Probit)
Retired (=1)	0.04*	0.04*	0.04*
Age (yrs.)	-0.002	-0.003	-0.003
Good health status (=1)	0.06*	0.07*	0.06*
HH income (USD)	0.0004*	0.0005*	0.0004*
Education (yrs.)	0.02*	0.02*	0.02*
Married (=1)	0.12*	0.12*	0.12*
Hispanic (=1)	-0.12*	-0.16*	-0.15*



## Example (cont.)

**Question:** How do you interpret the results?

Have health insurance	Regression		
	coefficient	ME (Logit)	ME (Probit)
Retired (=1)	0.04*	0.04*	0.04*
Age (yrs.)	-0.002	-0.003	-0.003
Good health status (=1)	0.06*	0.07*	0.06*
HH income (USD)	0.0004*	0.0005*	0.0004*
Education (yrs.)	0.02*	0.02*	0.02*
Married (=1)	0.12*	0.12*	0.12*
Hispanic (=1)	-0.12*	-0.16*	-0.15*

## Goodness-of-fit

- Unlike the LPM, where we can compute an  $R^2$  to judge the goodness of fit, here we need new measures.
- One possibility is a pseudo- $R^2$  based on the log likelihood. The log likelihood is the equivalent to the explained sum of squares in OLS (kinda...!). It is defined as  $1 - L_{ur}/L_r$ .
- An alternative is also to look at the percent of the predicted values falling within the  $[0; 1]$  interval.

## The tobit model

- We can also have limited dependent variable models that do not involve binary dependent variables.
- If we have censored data, e.g. when the dependent variable starts or includes many zeros, we often use tobit model.
- The tobit model also uses MLE. Hence, the same points apply with respect to the interpretation of the coefficients.

## Review questions

- 1 In the probit model all of the following are lies except: . . .
- (a)  $\beta_0$  cannot be negative since probabilities have to lie between 0 and 1.
  - (b)  $\beta_j$  tells you the effect of a unit increase in  $x_j$  on the probability that  $y = 1$ .
  - (c)  $\beta_j$  does not have a simple interpretation (i.e. cannot be interpreted directly).
  - (d)  $\beta_0$  is the probability of observing  $y$  when all the  $x$  variables are 0.

## Review questions (cont.)

You want to assess which factors influence school choice in India. The data that you have at hand contains the following variables:

```
obs:          902
vars:         9
size:        36,080 (99.7% of memory free)
```

---

variable name	storage type	display format	value label	variable label
numsib	float	%9.0g		Number of siblings
sraven	float	%9.0g		Raven ability score
wealth	float	%9.0g		Index of household asset value
male	float	%9.0g		Gender dummy: male=1, female=0
lowcaste	float	%9.0g		Low caste? yes=1,no=0
muslim	float	%9.0g		Muslim? yes=1,no=0
medyrs	float	%9.0g		Mother's education in years
sikhchr	float	%9.0g		Sikh or Christian? yes=1,no=0
stype	float	%9.0g		School type: 0=govt, 1=private aided, 2=private unaided

---

- Write down the model that you would estimate.
- How would you estimate the model?

## Review questions (cont.)

Running a probit model you get the following results:

```
> probit puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq sikhchr;
```

```
Iteration 0: log likelihood = -606.73067
Iteration 1: log likelihood = -373.10677
Iteration 2: log likelihood = -343.27331
Iteration 3: log likelihood = -340.47774
Iteration 4: log likelihood = -340.43889
Iteration 5: log likelihood = -340.43888
```

Probit estimates

```
Number of obs   =      902
LR chi2(9)      =     532.58
Prob > chi2     =     0.0000
Pseudo R2      =     0.4389
```

Log likelihood = -340.43888

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
puaind					
numsib	-.0998298	.0382835	-2.61	0.009	-.1748641 -.0247956
sraven	.0301986	.0054653	5.53	0.000	.0194869 .0409103
wealth	.0461453	.0043108	10.70	0.000	.0376963 .0545943
male	.8575159	.1199153	7.15	0.000	.6224862 1.092546
lowcaste	-.5496526	.1865875	-2.95	0.003	-.9153575 -.1839478
muslim	-.7229197	.1530685	-4.72	0.000	-1.022929 -.4229109
medyrs	-.1260082	.0373075	-3.38	0.001	-.1991296 -.0528868
medyrsq	.0079365	.0024278	3.27	0.001	.0031781 .0126948
sikhchr	.8875504	.3272338	2.71	0.007	.246184 1.528917
_cons	-1.882662	.287822	-6.54	0.000	-2.446783 -1.318541

## Review questions (cont.)

Running a probit model you get the following results:

```
Marginal effects after probit
      y = Pr(puaind) (predict)
      = .38659838
```

variable	dy/dx	Std. Err.	z	P> z	[	95% C.I.	]	X
numsib	-.0382063	.01465	-2.61	0.009	-.066911	-.009502		3.98891
sraven	.0115574	.00208	5.54	0.000	.007472	.015643		30.5266
wealth	.0176605	.00172	10.27	0.000	.014289	.021032		24.2572
male*	.3167018	.04152	7.63	0.000	.235328	.398076		.532151
lowcaste*	-.1926379	.05762	-3.34	0.001	-.305563	-.079713		.133038
muslim*	-.2513949	.04612	-5.45	0.000	-.341793	-.160997		.218404
medyrs	-.0482251	.01433	-3.37	0.001	-.076308	-.020142		8.66519
medyrsq	.0030374	.00093	3.26	0.001	.001211	.004864		99.6009
sikhchr*	.3401752	.11123	3.06	0.002	.122159	.558191		.031042

(\*) dy/dx is for discrete change of dummy variable from 0 to 1

(c) Interpret the results. What do you conclude?

## Summary

In this course you have seen:

- Different types of data and their characteristics
- Simple and multiple regression using OLS
  - Assumptions and properties
  - Coefficients and interpretation
  - Judging adequacy of a model
  - Sources of biases, implications and solutions
  - Hypothesis testing
  - Large sample characteristics
  - Variable manipulation
  - Incorporating different functional forms
  - Dummy variables
  - Heteroskedasticity (implications and remedies)
  - Measurement Error
  - Outliers



## Summary (cont.)

- Time series analysis
- Panel data analysis
  - First differencing, fixed and random effects

# Introduction to Quantitative Methods for Development



- A binary variable is often called a

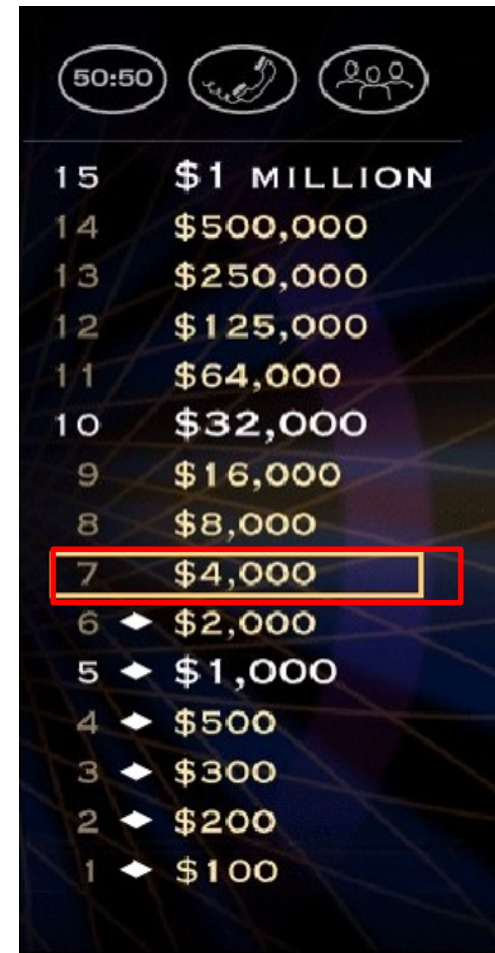
- a) dummy variable
- b) dependent variable
- c) residual
- d) power of a test



A screenshot of a game show wheel with 15 segments. The segments are numbered 1 to 15 and correspond to prize amounts. The wheel is set against a dark background with a grid pattern. At the top, there are three icons: a clock showing 50:50, a hand holding a coin, and two people. The prize amounts are listed on the right side of the wheel.

15	\$1 MILLION
14	\$500,000
13	\$250,000
12	\$125,000
11	\$64,000
10	\$32,000
9	\$16,000
8	\$8,000
7	\$4,000
6	\$2,000
5	\$1,000
4	\$500
3	\$300
2	\$200
1	\$100

- The t-statistic is calculated by dividing
  - a) the OLS estimator by its standard error.
  - b) the slope by the standard deviation of the explanatory variable.
  - c) the estimator minus its hypothesized value by the standard error of the estimator.
  - d) the slope by 1.96.



50:50

15	\$1 MILLION
14	\$500,000
13	\$250,000
12	\$125,000
11	\$64,000
10	\$32,000
9	\$16,000
8	\$8,000
7	\$4,000
6	◆ \$2,000
5	◆ \$1,000
4	◆ \$500
3	◆ \$300
2	◆ \$200
1	◆ \$100

- A fitted regression equation is given by  $\hat{Y} = 20 + 0.75X$ . What is the value of the residual at the point  $X=100$ ,  $Y=90$ ?

- a) 5
- b) -5
- c) 0
- d) 15



15	\$1 MILLION
14	\$500,000
13	\$250,000
12	\$125,000
11	\$64,000
10	\$32,000
9	\$16,000
8	\$8,000
7	\$4,000
6	\$2,000
5	\$1,000
4	\$500
3	\$300
2	\$200
1	\$100

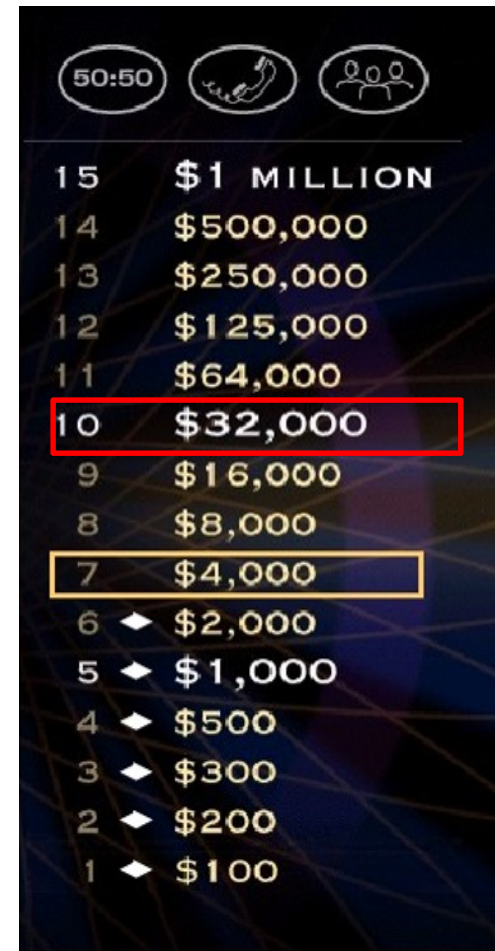
- If you wanted to test, using a 5% significance level, whether or not a specific slope coefficient is equal to one, then you should
  - a) subtract 1 from the estimated coefficient, divide the difference by the standard error, and check if the resulting ratio is larger than 1.96.
  - b) add and subtract 1.96 from the slope and check if that interval includes 1.
  - c) see if the slope coefficient is between 0.95 and 1.05.
  - d) check if the adjusted  $R^2$  is close to 1.



The image shows a vertical ladder of monetary values. At the top, there are three icons: a clock labeled '50:50', a hand holding a coin, and two people. The ladder consists of 15 steps, numbered 1 to 15 from bottom to top. Each step has a corresponding monetary value. Step 9 is highlighted with a red box, and step 7 is highlighted with a yellow box.

15	\$1 MILLION
14	\$500,000
13	\$250,000
12	\$125,000
11	\$64,000
10	\$32,000
9	\$16,000
8	\$8,000
7	\$4,000
6	◆ \$2,000
5	◆ \$1,000
4	◆ \$500
3	◆ \$300
2	◆ \$200
1	◆ \$100

- In the multiple regression model, the adjusted  $R^2$ 
  - a) cannot be negative.
  - b) will never be greater than the regression  $R^2$ .
  - c) equals the square of the correlation coefficient  $r$ .
  - d) cannot decrease when an additional explanatory variable is added.



Item	Value
15	\$1 MILLION
14	\$500,000
13	\$250,000
12	\$125,000
11	\$64,000
10	\$32,000
9	\$16,000
8	\$8,000
7	\$4,000
6	\$2,000
5	\$1,000
4	\$500
3	\$300
2	\$200
1	\$100

- Using 143 observations, assume that you had estimated a simple regression function and that your estimate for the slope was 0.04, with a standard error of 0.01. You want to test whether or not the estimate is statistically significant. Which of the following decisions is the only correct one:
  - a) you decide that the coefficient is small and hence most likely is zero in the population
  - b) the slope is statistically significant since it is four standard errors away from zero
  - c) the response of Y given a change in X must be economically important since it is statistically significant
  - d) since the slope is very small, so must be the regression  $R^2$



Number	Amount
15	\$1 MILLION
14	\$500,000
13	\$250,000
12	\$125,000
11	\$64,000
10	\$32,000
9	\$16,000
8	\$8,000
7	\$4,000
6	◆ \$2,000
5	◆ \$1,000
4	◆ \$500
3	◆ \$300
2	◆ \$200
1	◆ \$100



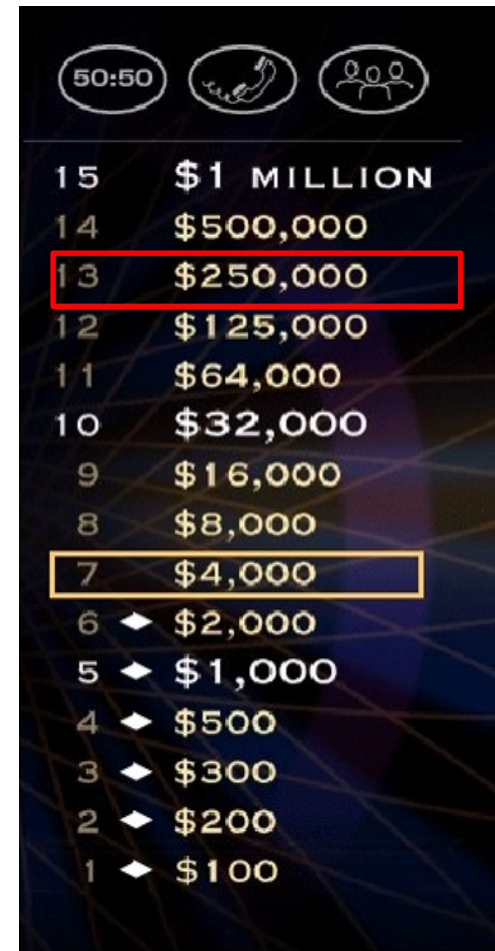
- What is the number of degrees of freedom for a simple bivariate linear regression with 20 observations?

- a) 20
- b) 22
- c) 18
- d) 2



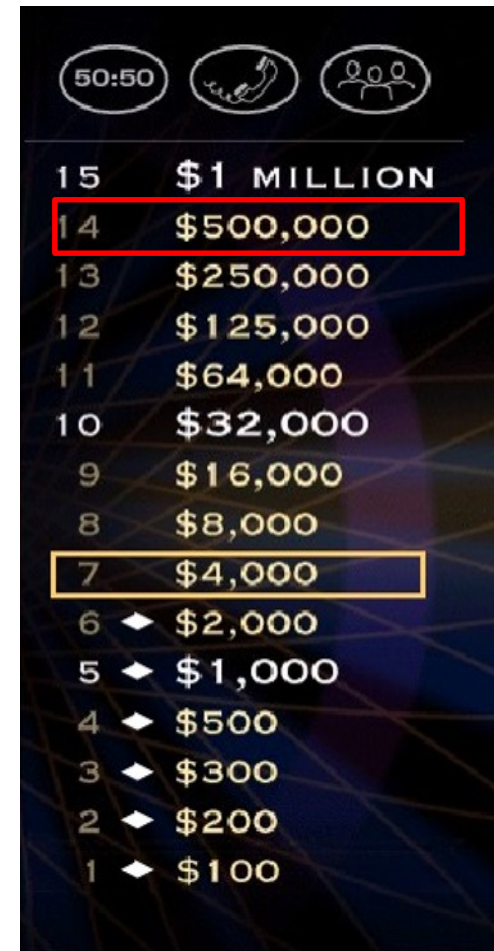
50:50		
15	\$1 MILLION	
14	\$500,000	
13	\$250,000	
12	\$125,000	
11	\$64,000	
10	\$32,000	
9	\$16,000	
8	\$8,000	
7	\$4,000	
6	◆ \$2,000	
5	◆ \$1,000	
4	◆ \$500	
3	◆ \$300	
2	◆ \$200	
1	◆ \$100	

- In a simple linear regression model the slope coefficient measures
  - a) the elasticity of  $Y$  with respect to  $X$
  - b) the change in  $Y$  which the model predicts for a unit change in  $X$
  - c) the change in  $X$  which the model predicts for a unit change in  $Y$
  - d) the value of  $Y$  for any given value of  $X$



Value	Amount
15	\$1 MILLION
14	\$500,000
13	\$250,000
12	\$125,000
11	\$64,000
10	\$32,000
9	\$16,000
8	\$8,000
7	\$4,000
6	▶ \$2,000
5	▶ \$1,000
4	▶ \$500
3	▶ \$300
2	▶ \$200
1	▶ \$100

- When testing joint hypothesis, you should
  - a) use t-statistics for each hypothesis and reject the null hypothesis if all of the restrictions fail
  - b) use the F-statistic and reject all the hypothesis if the statistic exceeds the critical value
  - c) use t-statistics for each hypothesis and reject the null hypothesis once the statistic exceeds the critical value for a single hypothesis
  - d) use the F-statistics and reject at least one of the hypothesis if the statistic exceeds the critical value



Prize	Amount
15	\$1 MILLION
14	\$500,000
13	\$250,000
12	\$125,000
11	\$64,000
10	\$32,000
9	\$16,000
8	\$8,000
7	\$4,000
6	\$2,000
5	\$1,000
4	\$500
3	\$300
2	\$200
1	\$100

- The least squares estimator of the slope coefficient is unbiased means
  - a) the estimated slope coefficient will always be equal to the true parameter value
  - b) the estimated slope coefficient will get closer to the true parameter value as the size of the sample increases
  - c) the estimated slope coefficient will be equal to the true parameter if the sample is large
  - d) if repeated samples of the same size are taken, on average their value will be equal to the true parameter



Value	Amount
15	\$1 MILLION
14	\$500,000
13	\$250,000
12	\$125,000
11	\$64,000
10	\$32,000
9	\$16,000
8	\$8,000
7	\$4,000
6	\$2,000
5	\$1,000
4	\$500
3	\$300
2	\$200
1	\$100

# REVIEW SESSION

## Lecture 2: Multiple Regression Analysis

---

- Which of the following can cause OLS estimators to be biased?
  - (a) Heteroskedasticity.
  - (b) Omitting an important variable.
  - (c) A sample correlation coefficient of 0.95 between two independent variables included in the model.

## Lecture 3: Inference

- Which of the following can cause the usual OLS  $t$ -statistic to be invalid (that is, not to have  $t$ -distributions under  $H_0$ )?
  - (a) Heteroskedasticity.
  - (b) Omitting an important variable.
  - (c) A sample correlation coefficient of 0.95 between two independent variables included in the model.

## Lecture 4: Asymptotic Properties

- What is the difference between unbiasedness and consistency?  
Explain



## Lecture 5: Further Issues

- Clarification: In contrast, you do not want to include a variable that prohibits a sensible interpretation of the variable of interest.

### **Example?**

- The following model allows the returns to education to depend upon the total amount of both parents' education called *par*. The estimated equation is:

$$\log(\hat{w}age) = 5.65 + 0.47edu + 0.00078educ * par + 0.19exper$$

(0.13)    (0.01)    (0.00021)    (0.004)

$n = 722, R^2 = 0.169$

How do you interpret the results?

## Lecture 6: Dummy Variables

We have estimate the following model:

$$\hat{w}age = -1.37 - 1.2female + 0.572educ + 0.025exper$$

(0.72)      (0.80)      (0.049)      (0.012)

How would the interpretation change if the dependent variable was measured in logs?

## Lecture 6 (cont.)

- 1 Suppose you collected data from youth working in the informal sector in Laos (Nigeria). Your survey asks information on wages, education, experience, and gender. In addition, you ask for information about marijuana usage. The original question is: 'On how many separate occasions last month did you smoke marijuana?'
  - (a) Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, 'smoking marijuana five more times per month is estimated to change wage by x%.'
  - (b) Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug usage for men and women?

## Lecture 6 (cont.)

- (c) Suppose you think it is better to measure marijuana usage by putting people into one of four categories: non-user, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana usage on wage.
- (d) Using the model in part (c), explain in detail how to test the null hypothesis that marijuana usage has no effect on wage. Be very specific and include a careful listing of the degrees of freedom.
- (e) What are some potential problems with drawing causal inference using the survey data that you collected?

## Lecture 7: Heteroskedasticity

- State with brief reason whether the following statements are true, false, or uncertain:
  - (a) In the presence of heteroskedasticity OLS estimators are biased as well as inefficient.
  - (b) If heteroskedasticity is present, the conventional  $t$  and  $F$ -tests are invalid.
  - (c) In the presence of heteroskedasticity the usual OLS method always overestimates the standard errors of estimators.
  - (d) If residuals estimated from an OLS regression exhibit a systematic pattern, it means heteroscedasticity is present in the data.
  - (e) If a regression model is mis-specified (e.g. an important variable is omitted), the OLS residuals will show a distinct pattern.
  - (f) If a regressor that has non-constant variance is (incorrectly) omitted from the model, the (OLS) residuals will be heteroskedastic.

## Lecture 8: Time Series

- Clarification: What is the difference between a forecast and a predicted value and the forecast error and the residual?
- Decide if you agree or disagree with the following statements and give a brief explanation of your decision:
  - (a) Like cross-sectional observations, we can assume that most time series observations are independently distributed.
  - (b) A trending variable cannot be used as the dependent variable in multiple regression analysis.
  - (c) Seasonality is an issue when using annual time series observations.
- Suppose you have quarterly data on new housings, interest rates, and real per capita income in Brazil. Specify a model for housing starts that accounts for possible trends and seasonality in the variables.

## Lecture 9: Panel Data

- Suppose that we want to estimate the effect of several variables on annual saving and that we have a panel data set on individuals collected on January 31, 1990, and January 31, 1992. If we include a year dummy for 1992 and use first differencing, can we also include age in the original model? Explain.
- In order to determine the effect of doping on athletic performance (focus on running for now), you collect data on participants of the Golden League Meetings in 2010, 2012, and 2014.
  - a) What measures of athletic success would you include in an equation? Are there some timing issues?
  - b) What other measures might you control for in the equation?
  - c) Write down an equation that allows you to estimate the effect of athletic success on the percentage change in substance use. How would you estimate this equation? Why would you choose this method.

## Lecture 10: Specification & Data Problems

- The regression model includes a random error or disturbance term for a variety of reasons. Which of the following is NOT one of them?
  - a) Measurement errors in the observed variables
  - b) Omitted influences on Y (other than X)
  - c) Linear functional form is only an approximation
  - d) The observable variables do not exactly correspond with their theoretical counterparts
  - e) There may be approximation errors in the calculation of the least squares estimates



## Lecture 12: Limited Dependent Variables

You want to assess which factors influence school choice in India. The data that you have at hand contains the following variables:

```
obs:          902
vars:         9
size:        36,080 (99.7% of memory free)
```

---

variable name	storage type	display format	value label	variable label
numsib	float	%9.0g		Number of siblings
sraven	float	%9.0g		Raven ability score
wealth	float	%9.0g		Index of household asset value
male	float	%9.0g		Gender dummy: male=1, female=0
lowcaste	float	%9.0g		Low caste? yes=1,no=0
muslim	float	%9.0g		Muslim? yes=1,no=0
medyrs	float	%9.0g		Mother's education in years
sikhchr	float	%9.0g		Sikh or Christian? yes=1,no=0
stype	float	%9.0g		School type: 0=govt, 1=private aided, 2=private unaided

---

- Write down the model that you would estimate.
- How would you estimate the model?

## Lecture 12 (cont.)

Running a probit model you get the following results:

```
> probit puaind numsib sraven wealth male lowcaste muslim medyrs medyrsq sikhchr;
```

```
Iteration 0: log likelihood = -606.73067
Iteration 1: log likelihood = -373.10677
Iteration 2: log likelihood = -343.27331
Iteration 3: log likelihood = -340.47774
Iteration 4: log likelihood = -340.43889
Iteration 5: log likelihood = -340.43888
```

Probit estimates

```
Number of obs   =      902
LR chi2(9)      =    532.58
Prob > chi2     =    0.0000
Pseudo R2      =    0.4389
```

Log likelihood = -340.43888

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
puaind					
numsib	-.0998298	.0382835	-2.61	0.009	-.1748641 -.0247956
sraven	.0301986	.0054653	5.53	0.000	.0194869 .0409103
wealth	.0461453	.0043108	10.70	0.000	.0376963 .0545943
male	.8575159	.1199153	7.15	0.000	.6224862 1.092546
lowcaste	-.5496526	.1865875	-2.95	0.003	-.9153575 -.1839478
muslim	-.7229197	.1530685	-4.72	0.000	-1.022929 -.4229109
medyrs	-.1260082	.0373075	-3.38	0.001	-.1991296 -.0528868
medyrsq	.0079365	.0024278	3.27	0.001	.0031781 .0126948
sikhchr	.8875504	.3272338	2.71	0.007	.246184 1.528917
_cons	-1.882662	.287822	-6.54	0.000	-2.446783 -1.318541

## Lecture 12 (cont.)

Running a probit model you get the following results:

```
Marginal effects after probit
      y = Pr(puaind) (predict)
      = .38659838
```

variable	dy/dx	Std. Err.	z	P> z	[	95% C.I.	]	X
numsib	-.0382063	.01465	-2.61	0.009	-.066911	-.009502	3.98891	
sraven	.0115574	.00208	5.54	0.000	.007472	.015643	30.5266	
wealth	.0176605	.00172	10.27	0.000	.014289	.021032	24.2572	
male*	.3167018	.04152	7.63	0.000	.235328	.398076	.532151	
lowcaste*	-.1926379	.05762	-3.34	0.001	-.305563	-.079713	.133038	
muslim*	-.2513949	.04612	-5.45	0.000	-.341793	-.160997	.218404	
medyrs	-.0482251	.01433	-3.37	0.001	-.076308	-.020142	8.66519	
medyrsq	.0030374	.00093	3.26	0.001	.001211	.004864	99.6009	
sikhchr*	.3401752	.11123	3.06	0.002	.122159	.558191	.031042	

(\*) dy/dx is for discrete change of dummy variable from 0 to 1

(c) Interpret the results. What do you conclude?