University of Nottingham
UK | CHINA | MALAYSIA

Introductory Econometrics BUSI2053

The Simple Linear Regression Model I: Specification and Estimation

Lecture Outline

- An Economic Model
- An Econometric Model
- The Nature of Regression Analysis
- Estimating the Regression Parameters
- Interpretation the estimated coefficient

- **Suggested Reading:**
- Chapter 2, 3 & 4.1, 4.2: Hill, R.C., Griffiths W.E. and Lim, G.C. Principles of Econometrics, fourth edition, Wiley, 2012 (pp. 39-110)
- Chapter 5-6: Westhoff (2013) An introduction to econometrics: a self-contained approach, MIT Press, 2013
- Chapter 1-3: Gujarati, D.N. and Porter D.C. Basic econometrics, 5th ed., McGraw-Hill, 2009 (pp. 34-117)
- Chapter 1 & 2: Dougherty, Christopher. Introduction to econometrics. 4th ed. Oxford University Press, 2011 (pp.83-147)

Before we move on,

- Linear equation
- Random variable
- Conditional probability
- Population parameters (Fixed values)
- Sample statistics (Random variables)
- Expected value rules
- Variance and covariance rules
- Estimator and estimate
- Properties of an estimator

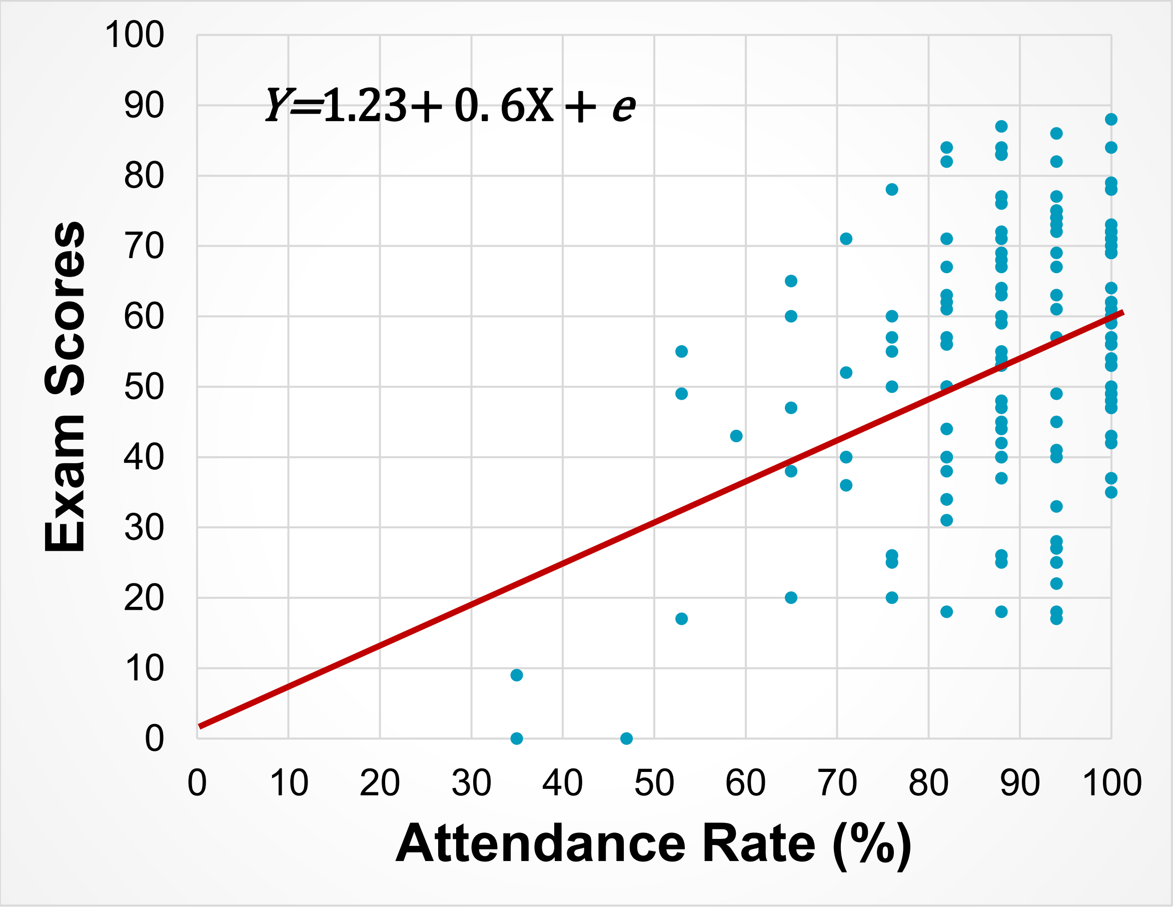
Simple linear regression

The statistical method that uses a straight-line relationship to predict a numerical dependent variable *Y* from a single numerical independent variable *X*.

Inferential analytics:
On average, 1% increase in attendance would lead to 0.6 mark increase in the exam score.

Predictive analytics:
A student with 80% attendance rate is expected to score 49 marks [E(Y)=1.23 + (0.6+80)=49.23]

Prescriptive analytics:
To score the pass mark of 40, a student must attend at least 65% of the classes [E(Y)=1.23 + (0.6+65)=40.23]



$$Exam\ Scoree = f(Attendance)$$

An Economic Model

- As business decision makers, we are usually more interested in studying relationships between variables
  - Economic theory tells us that expenditure on economic goods depends on income

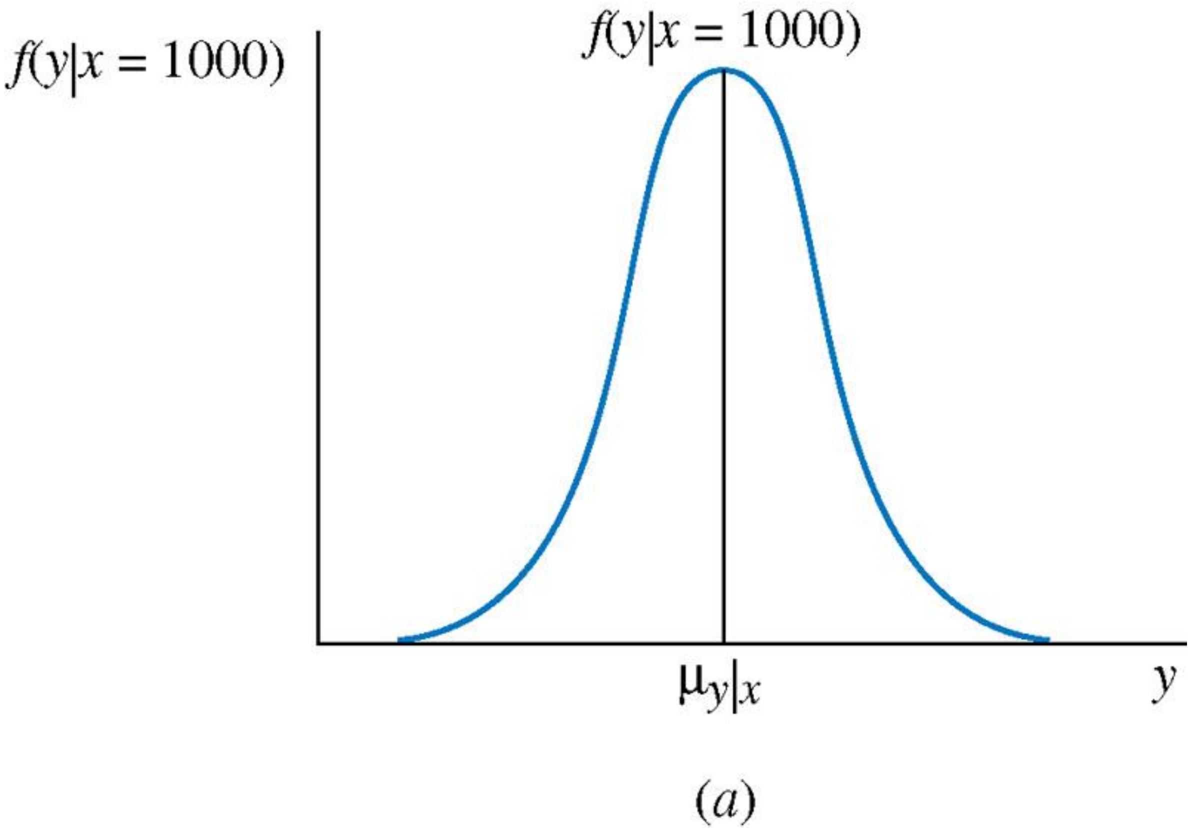
$$Expendidure = f(Income)$$

  - Consequently, *Expenditure* is defined as the ''dependent variable'' and *Income* as the "independent'' or ''explanatory'' variable
  - The dependent variable is denoted as y and the independent variable is denoted as x.

y=Expenditure; x=Income

An Economic Model to Econometric Model

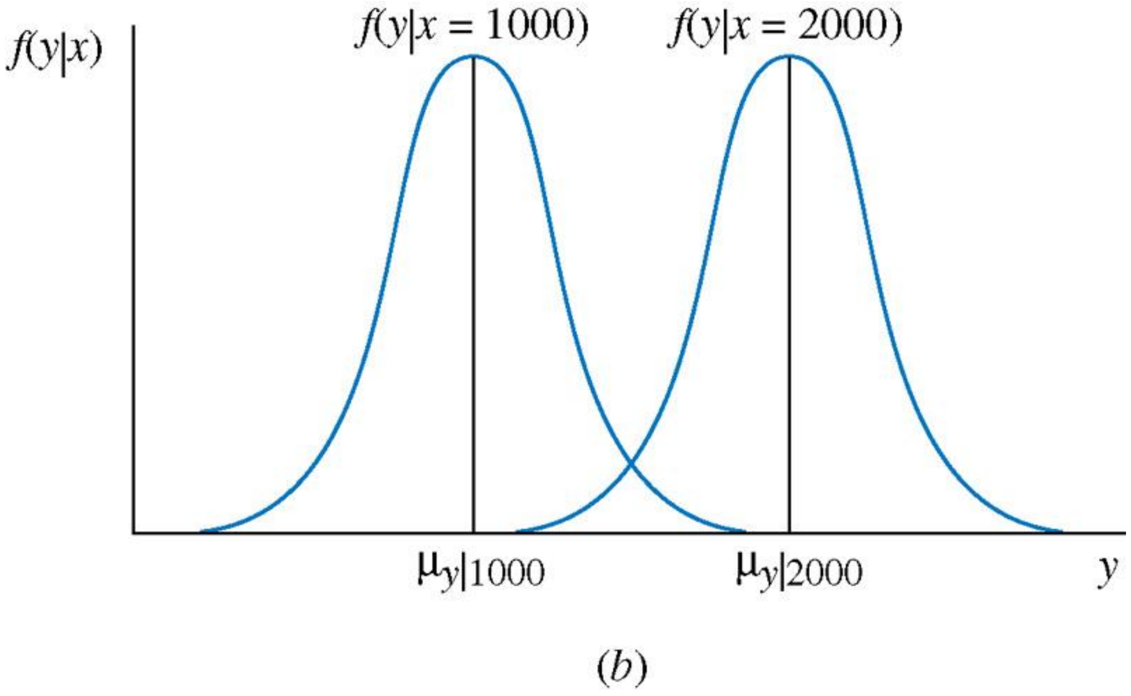
- Suppose that we are interested only in food expenditure of households with an income of $1,000 per week



Probability distribution of food expenditure y given income x = $1000

An Economic Model to Econometric Model

- In econometrics, we recognise that real-world expenditures are **random variables**, and we want to use data to learn about the relationship between income and expenditure.



Regression analysis is largely concerned with estimating and/or predicting the (population) mean value of the dependent variable based on the known or fixed values of the explanatory variable(s).

Probability distributions of food expenditures y given incomes x = \$1000 and x = \$2000

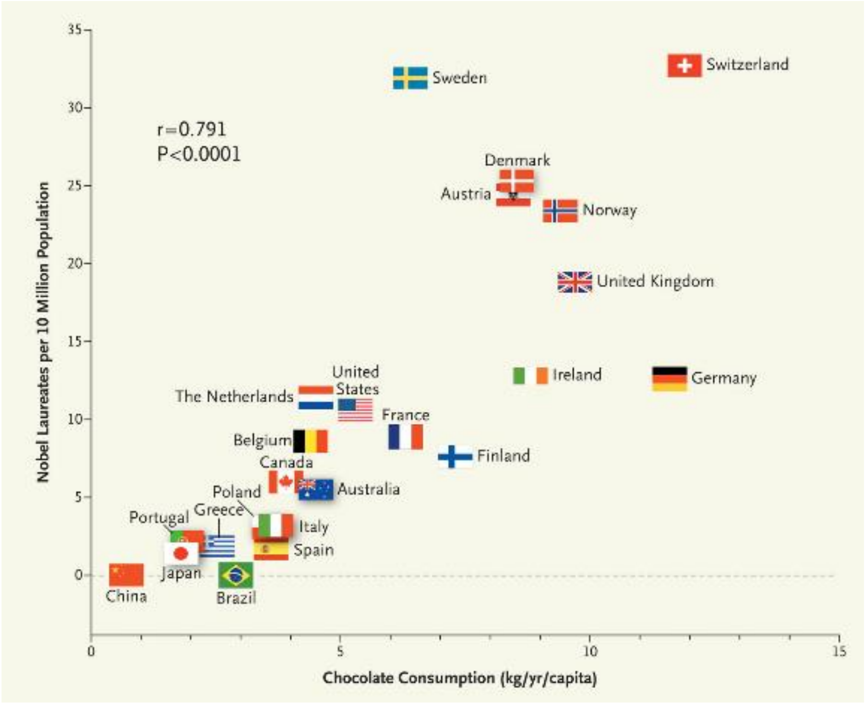
Notes on Regression Analysis 1

Regression versus Correlation

- The correlation analysis measures the strength or degree of linear association between two variables that are assumed to be random
- Regression analysis tries to estimate or predict the average value of a random variable based on the fixed values of other variables
- In regression analysis the dependent variable is assumed to be statistical, random, or stochastic. The explanatory variables, on the other hand, are assumed to have fixed values (in repeated sampling)

Does chocolate make you clever?

By Charlotte Pritchard
BBC News

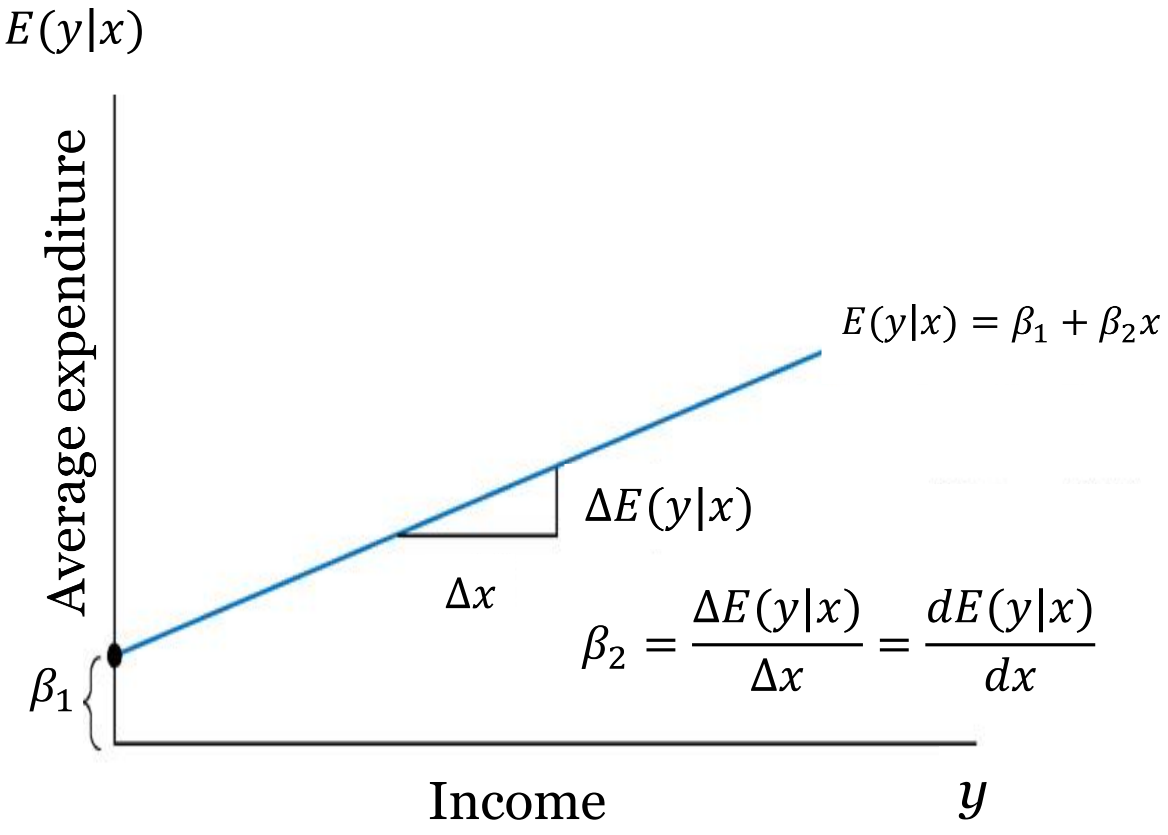


An Econometric Model

- To investigate the relationship between expenditure and income we must build an **economic model** and then a corresponding **econometric model**
- This econometric model is also called a **regression model**
  - The simple regression model can be written as

$$E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x$$

where β_1 is the intercept and β_2 is the slope and they are **population parameters (and constant)**

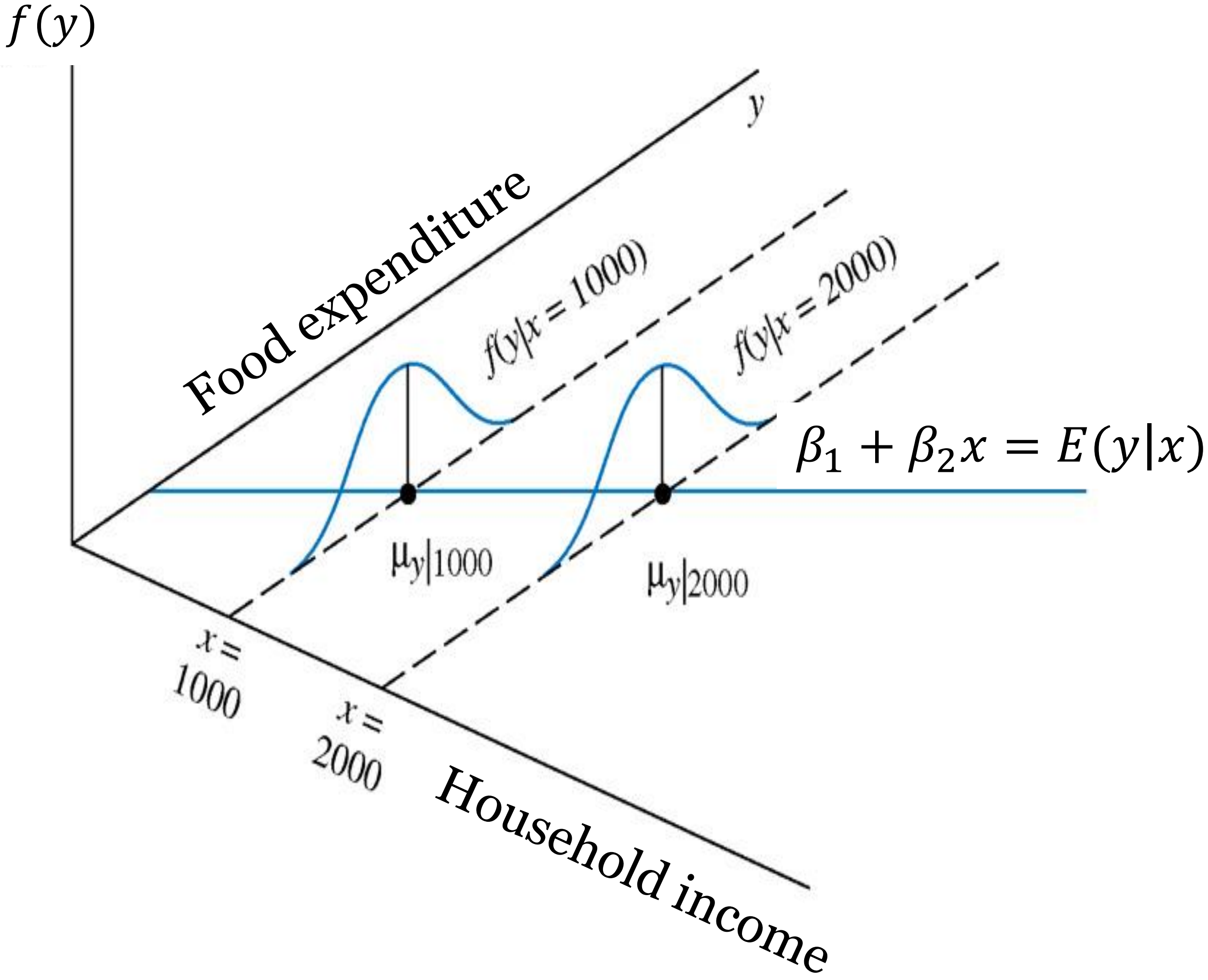


$E(y|x)$
Average expenditure
$E(y|x) = \beta_1 + \beta_2 x$
$\Delta E(y|x)$
Δx
$\beta_2 = \frac{\Delta E(y|x)}{\Delta x} = \frac{dE(y|x)}{dx}$
β_1
Income
y

$dE(y|x)/dx$ denotes the derivative of the expected value of y given an x value

An Econometric Model

- The model $E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x$ describes economic behaviour in the population of our interest
- If we were to sample household expenditures at other levels of income the regression function describes the mean value of household expenditure for each level of income
- Note: $E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x$ represents the model for expected value of y. For the actual value of y, the model can be written as: $y = \beta_1 + \beta_2 x + e$ where e is **random error term**.



Introducing the Error Term

- Any observation on the dependent variable y can be decomposed into two parts: a systematic component and a random component called a "**random error**".

For example, the equation for y_4:
$y_4 = \beta_1 + \beta_2 x_4 + e_4$

- The random error term is defined as

$$e = y - E(y|x) = y - \beta_1 - \beta_2 x$$

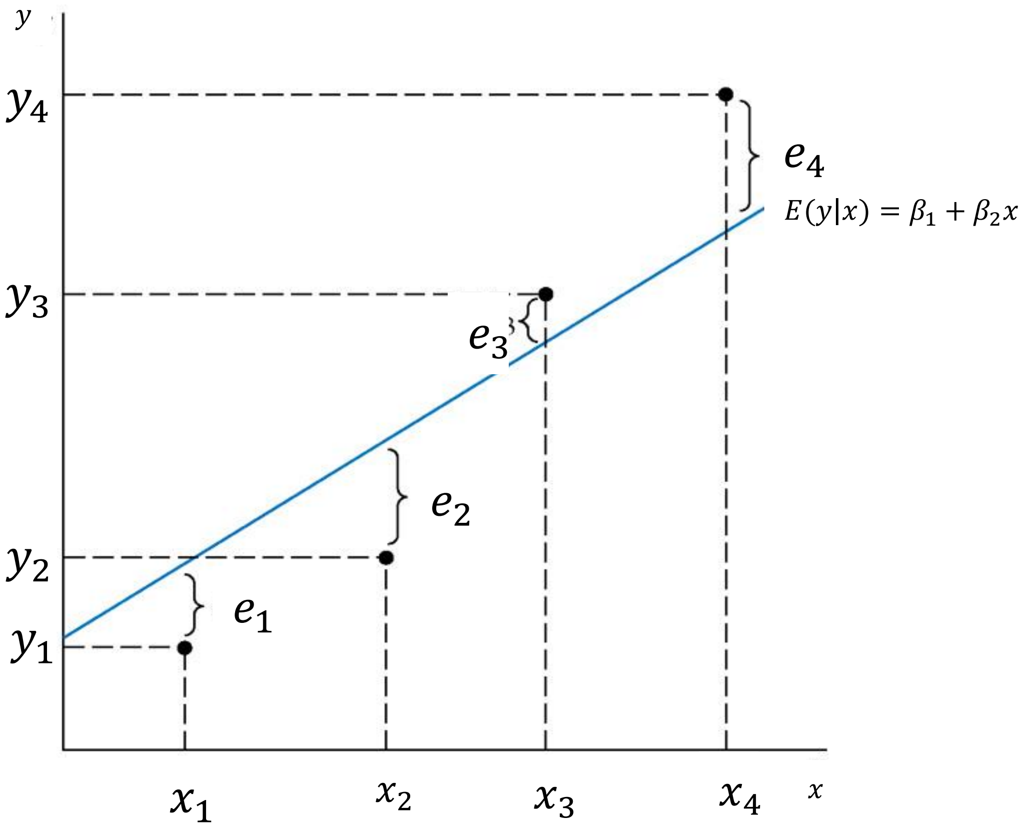
- Rearranging gives a population regression model.

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

- The expected value of the error term, given x, is

$$E(e|x) = E(y|x) - \beta_1 - \beta_2 x = 0$$

Why?



The relationship among y, e and the true regression line

Introducing the Error Term

- The random error term is defined as

$$e = y - E(y|x) = y - \beta_1 - \beta_2 x$$

- Any other economic factors that affect expenditures on food are "collected" in the error term.

- The error term *e* is a "storage bin" for unobservable and/or unimportant factors affecting the dependent variable

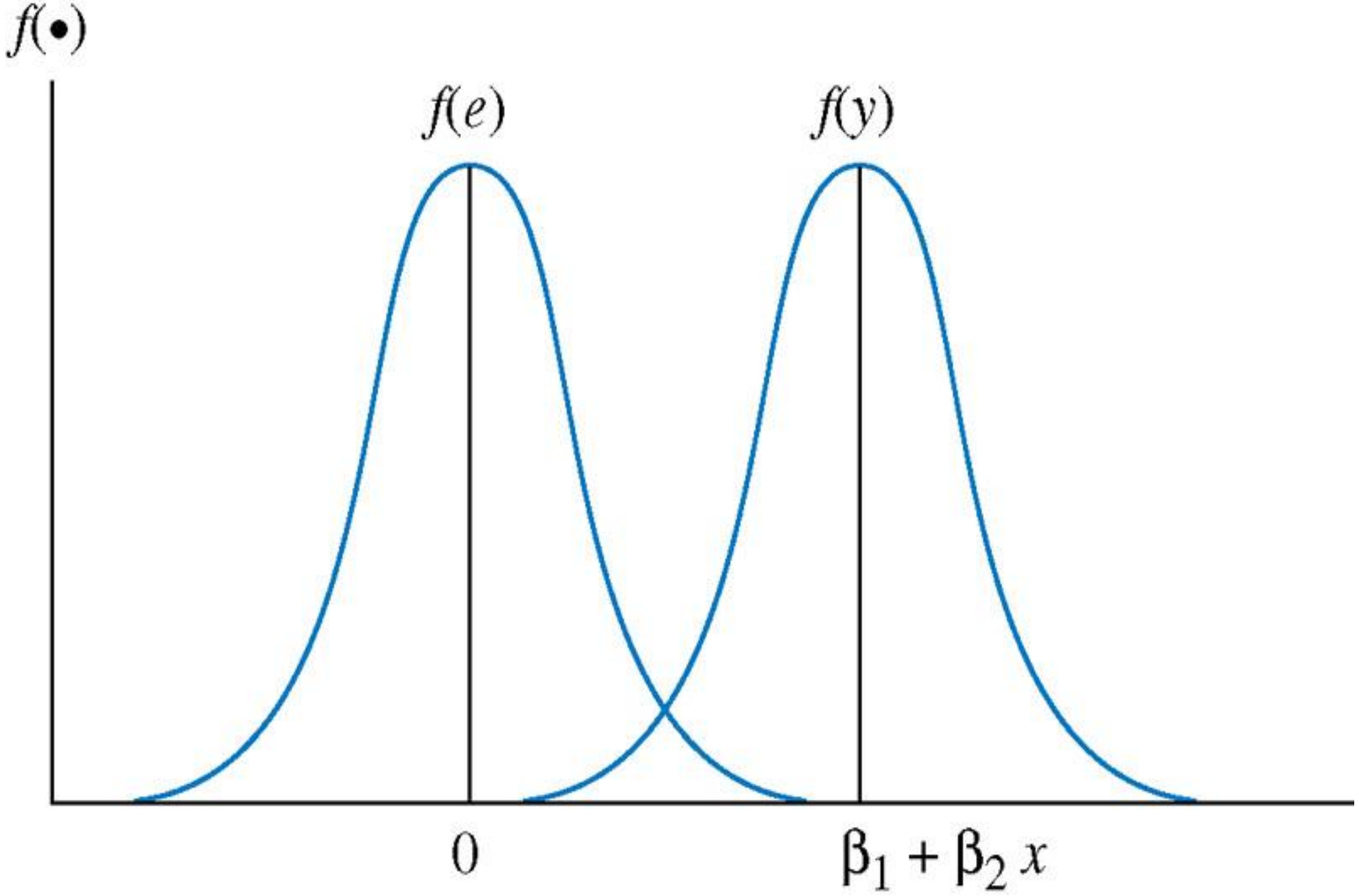
- The error term *e* captures any approximation error that arises because the *linear* functional form we have assumed

- The error term captures any elements of random behaviour that may be present in each individual observation.

Introducing the Error Term

Probability density functions for e and y

Note that $\text{Var}(e) = \text{Var}(y)$

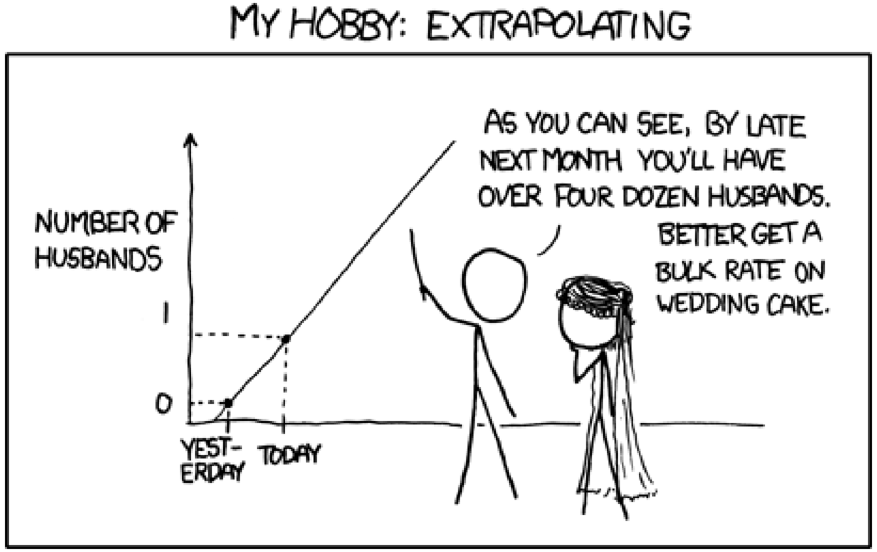


Notes on Regression Analysis 2

- Regression analysis concerned with the statistical, not functional or deterministic, dependence among variables

- The dependence of family food expenditure on income, family size, location, and taste, for example, is statistical in nature

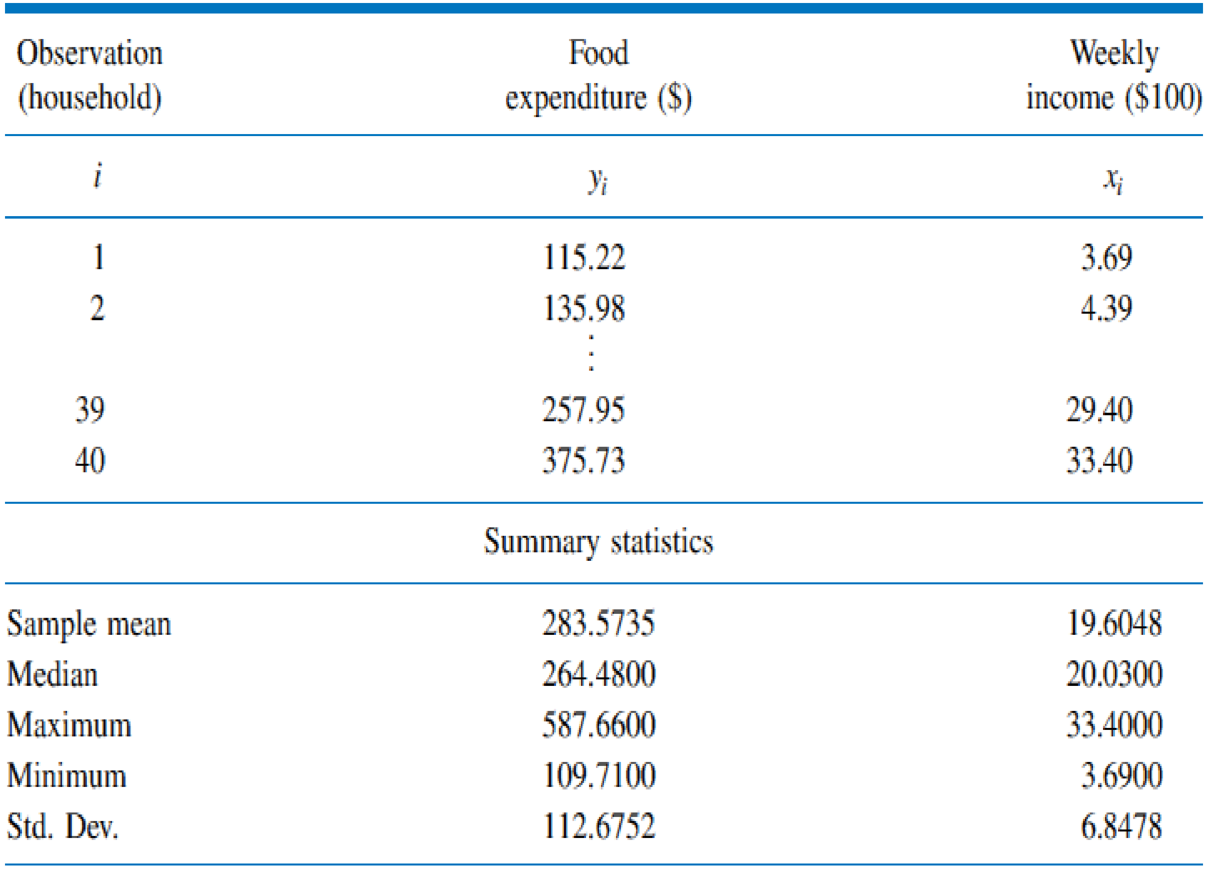
- The explanatory variables, although certainly important, will not enable the economist to predict food expenditure exactly because of measurement errors as well as many other unobservable factors (variables) that collectively affect the expenditure


MY HOBBY: EXTRAPOLATING
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS.
BETTER GET A BULK RATE ON WEDDING CAKE.

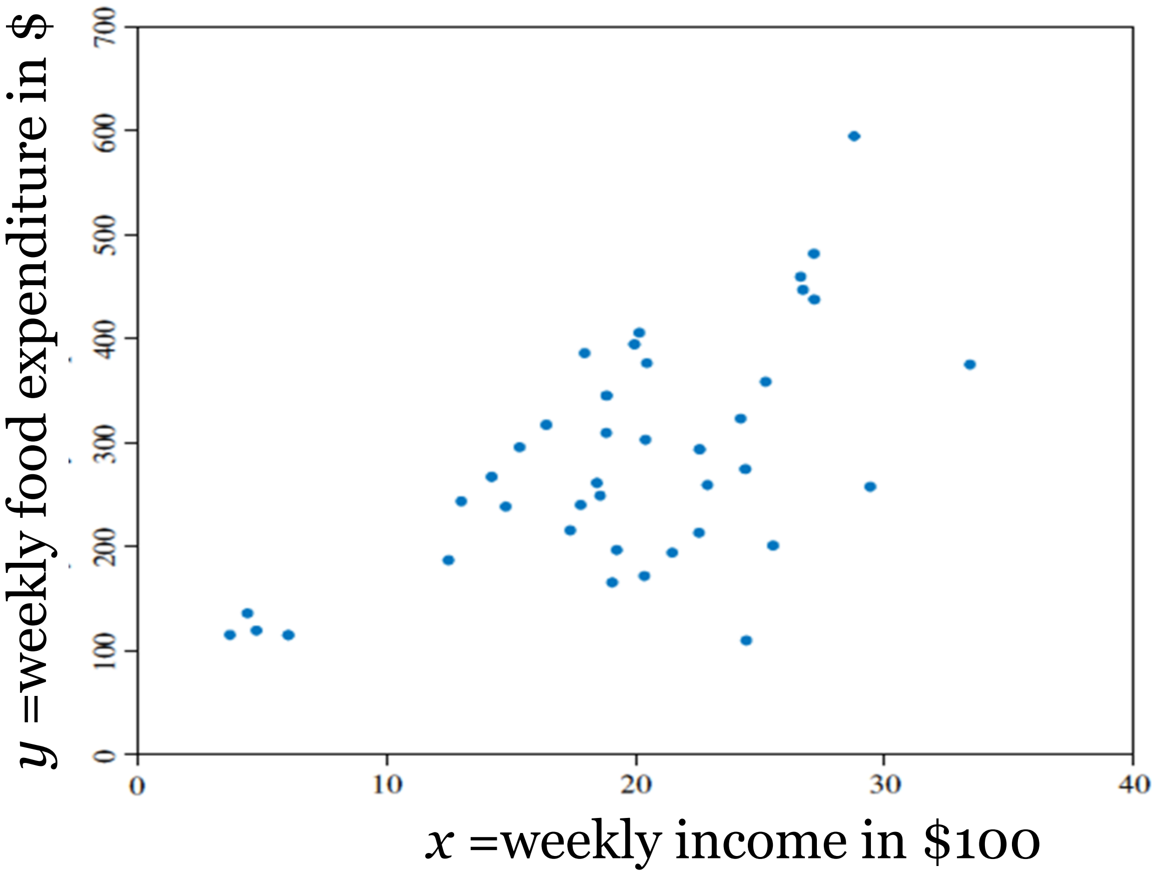
- Thus, there is bound to be some “intrinsic” or random variability in the dependent-variable food expenditure that cannot be fully explained no matter how many explanatory variables we consider.

Estimating the Regression Parameters

Estimates for the Food Expenditure Function



Observation (household)	Food expenditure ($)	Weekly income ($100)
i	y_i	x_i
1	115.22	3.69
2	135.98	4.39
	⋮	
39	257.95	29.40
40	375.73	33.40
	Summary statistics	
Sample mean	283.5735	19.6048
Median	264.4800	20.0300
Maximum	587.6600	33.4000
Minimum	109.7100	3.6900
Std. Dev.	112.6752	6.8478



Food Expenditure and Income Data [food (POE 4th ed.) from gretl]

Estimating the Regression Parameters

Least Squares Principles

- We want to estimate population parameters β_1 and β_2 from

$$y_i = \beta_1 + \beta_2 x_i + e_i$$

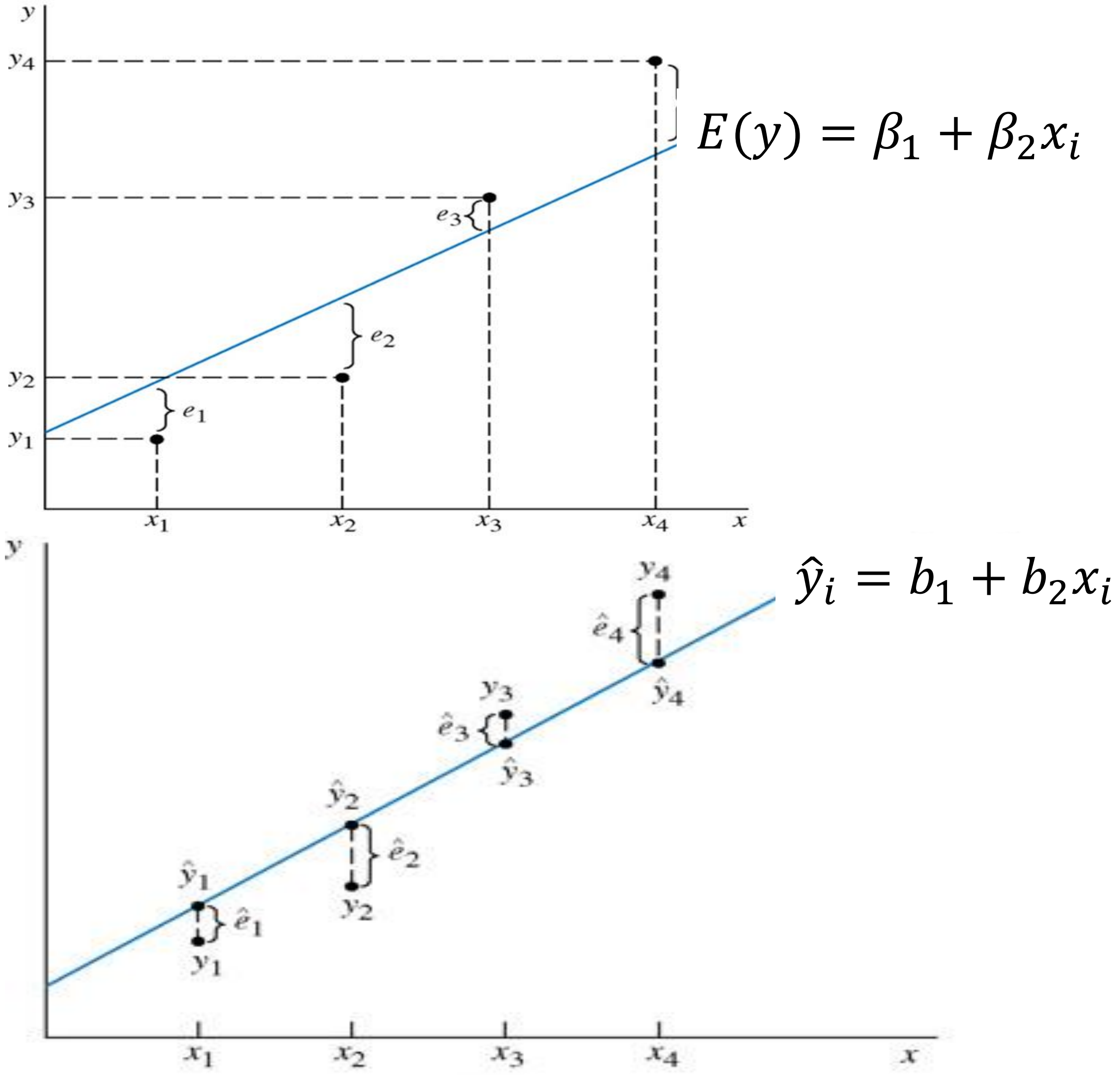
- The fitted regression line is:

$$\hat{y}_i = b_1 + b_2 x_i$$

- The least squares residual is:

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i$$

Be mindful of differences in notations in different textbooks.



Estimating the Regression Parameters

Least Squares Principles

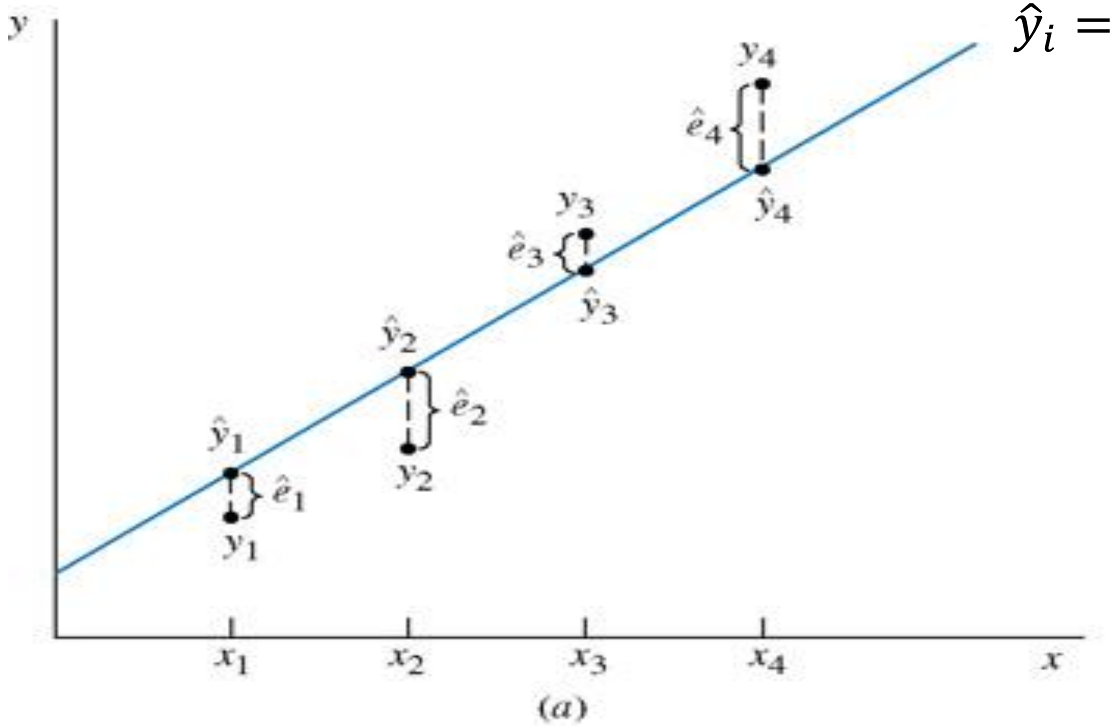
- Least squares estimates for the unknown parameters β₁ and β₂ are obtained my minimizing the sum of squares function:

$$Minimise \square \hat{e}_i^2 = \square (y_i - b_1 - b_2 x_i)^2$$

$$b_2 = \frac{\square (x_i - \bar{x})(y_i - \bar{y})}{\square (x_i - \bar{x})^2}$$

$$b_2 = \frac{\square X_i Y_i - n\bar{X}\bar{Y}}{\square X_i^2 - n\bar{X}^2}$$

$$b_1 = \bar{y} - b_2 \bar{x}$$



$$\hat{y}_i = b_1 + b_2 x_i$$

See Appendix 2A (p83-84) in Hill et el. (2012) or Appendix 3A.1 (p.92) Gujarati and Porter (2009) for derivations.

Estimating the Regression Parameters (exercise)

1	*2*	*3*	*4*	*3*4*	*5*		
x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$		$(x_i - \bar{x})^2$	$\hat{y} = b_1 + b_2 x_i$	$\hat{e}$
2	4	-2	-1	2	4		
3	3	-1	-2	2	1		
5	4	1	-1	-1	1		
6	7	2	2	4	4		
5	4	1	-1	-1	1		
4	7	0	2	0	0		
4	6	0	1	0	0		
3	5	-1	0	0	1		
$\bar{x}$ =4	$\bar{y}$ =5			$\sum$=6	$\sum$=12		$\sum \hat{e} =$

x = Number of children in a household

y = Number of visits to *KFC* in a month

$$b_2 = \frac{\square(x_i - \bar{x})(y_i - \bar{y})}{\square(x_i - \bar{x})^2} =$$

$$b_1 = \bar{y} - b_2 \bar{x} =$$

Estimating the Regression Parameters (exercise)

1	*2*	*3*	*4*	*3*4*	*5*	***6***	***2 – 6***
x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$		$(x_i - \bar{x})^2$	$\hat{y} = b_1 + b_2 x_i$	$\hat{e}$
2	4	-2	-1	2	4	$3 + (0.5 \times 2) = 4$	0
3	3	-1	-2	2	1	$3 + (0.5 \times 3) = 4.5$	-1.5
5	4	1	-1	-1	1	$3 + (0.5 \times 5) = 5.5$	-1.5
6	7	2	2	4	4	$3 + (0.5 \times 6) = 6$	1
5	4	1	-1	-1	1	$3 + (0.5 \times 5) = 5.5$	-1.5
4	7	0	2	0	0	$3 + (0.5 \times 4) = 5$	2
4	6	0	1	0	0	$3 + (0.5 \times 4) = 5$	1
3	5	-1	0	0	1	$3 + (0.5 \times 3) = 4.5$	0.5
$\bar{x}$ =4	$\bar{y}$ =5			$\sum$=6	$\sum$=12		$\sum \hat{e} = 0$

x = Number of children in a household

y = Number of visits to *KFC* in a month

$$y_i = 3 + 0.5x_i + e_i$$

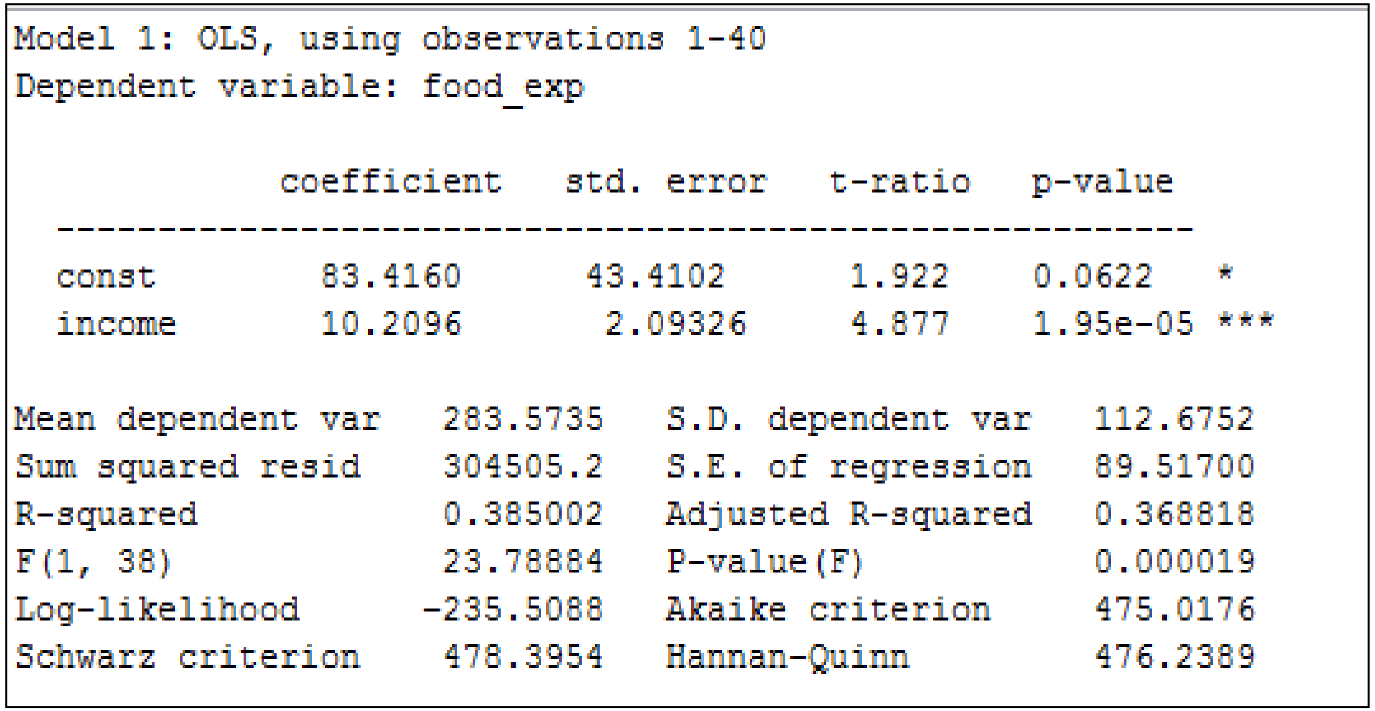
$$b_2 = \frac{\square(x_i - \bar{x})(y_i - \bar{y})}{\square(x_i - \bar{x})^2} = 6/12 = 0.5$$

$$b_1 = \bar{y} - b_2\bar{x} = 5 - (0.5 \times 4) = 3$$

Estimating the Regression Parameters

Estimates for the Food Expenditure Function using GRETL

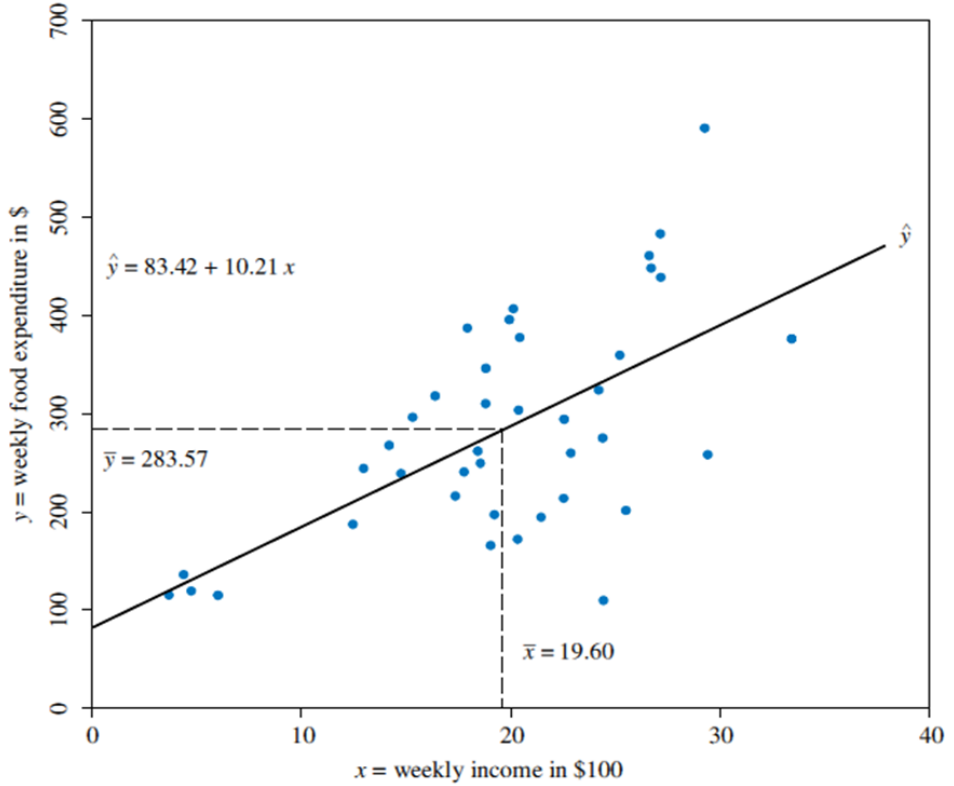
Model...>>Ordinary least squares...



```
Model 1: OLS, using observations 1-40
Dependent variable: food_exp

             coefficient   std. error   t-ratio   p-value
  --------------------------------------------------------
  const        83.4160      43.4102       1.922    0.0622   *
  income       10.2096       2.09326      4.877    1.95e-05 ***

Mean dependent var   283.5735   S.D. dependent var   112.6752
Sum squared resid    304505.2   S.E. of regression   89.51700
R-squared            0.385002   Adjusted R-squared   0.368818
F(1, 38)             23.78884   P-value(F)           0.000019
Log-likelihood      -235.5088   Akaike criterion     475.0176
Schwarz criterion    478.3954   Hannan-Quinn         476.2389
```

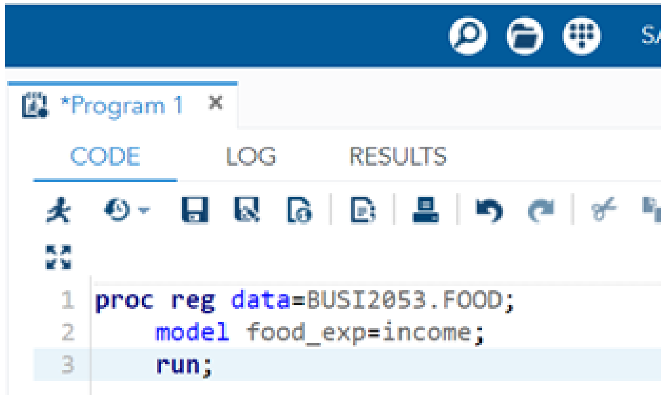


A convenient way to report the values for b_1 and b_2 is to write out the *estimated* or *fitted* regression line: $\hat{y}_i = 83.42 + 10.21x_i$

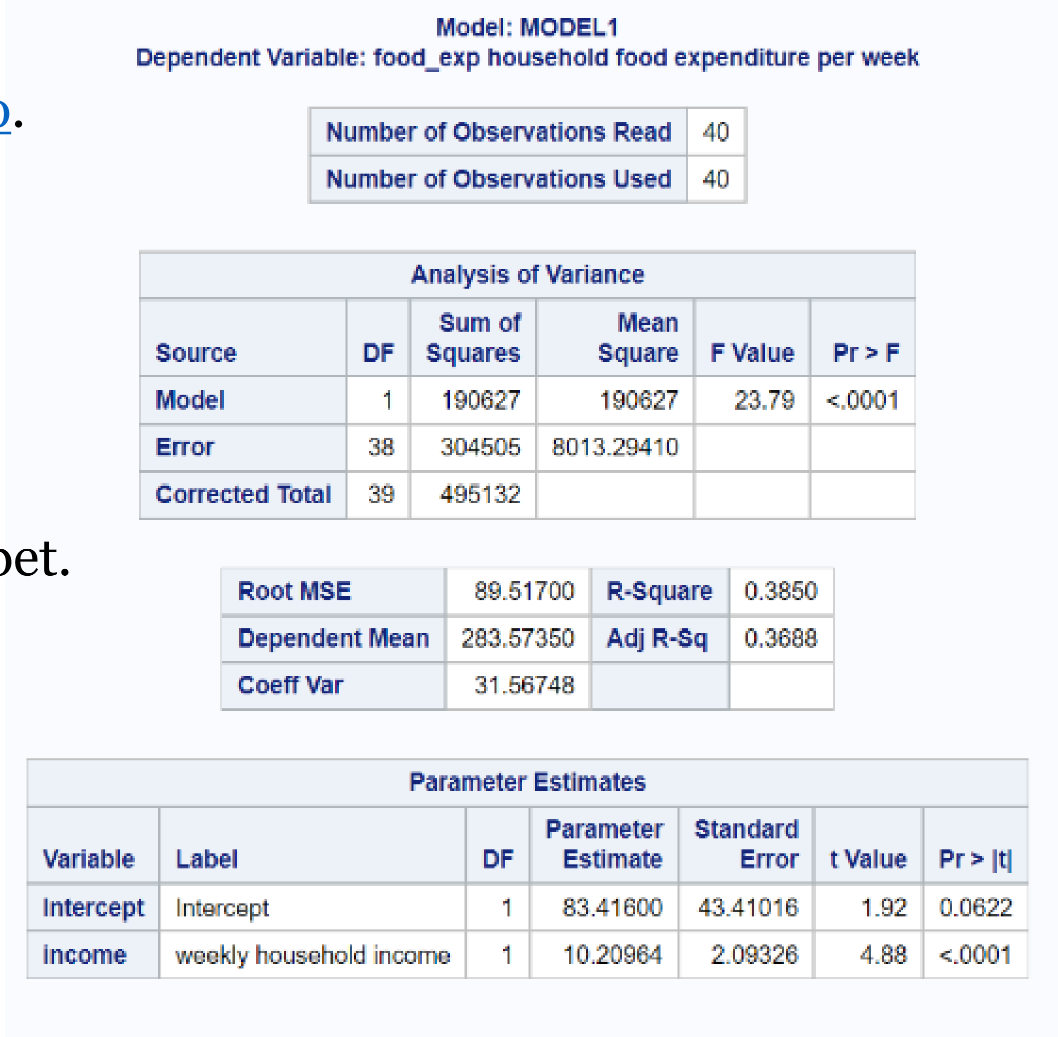
Estimating the Regression Parameters

Estimates for the Food Expenditure Function using SAS®

Use “Linear Regression” under Tasks and Utilities in SAS studio. select “food.sas7bdat” from the “BUSI2053” folder, add “food_exp” as Independent variable and “income” as continuous variable. Then complete the model builder. (or) use the following code snippet.



```
proc reg data=BUSI2053.FOOD;
    model food_exp=income;
    run;
```



Model: MODEL1
Dependent Variable: food_exp household food expenditure per week

Number of Observations Read	40
Number of Observations Used	40

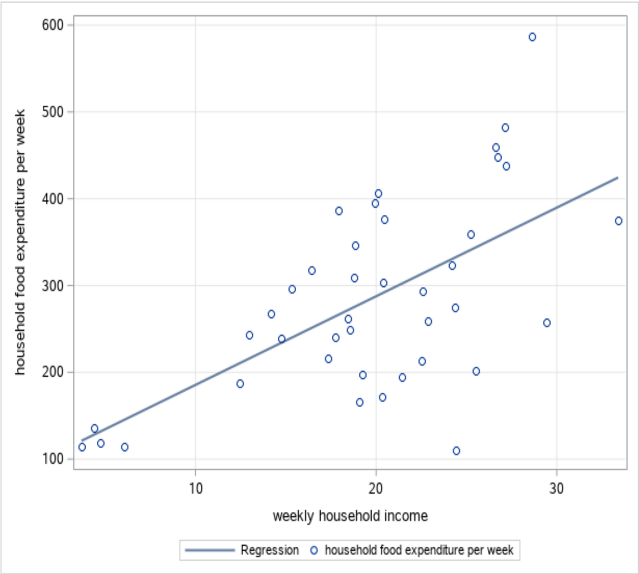
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	190627	190627	23.79	<.0001
Error	38	304505	8013.29410		
Corrected Total	39	495132			

Root MSE	89.51700	R-Square	0.3850
Dependent Mean	283.57350	Adj R-Sq	0.3688
Coeff Var	31.56748		

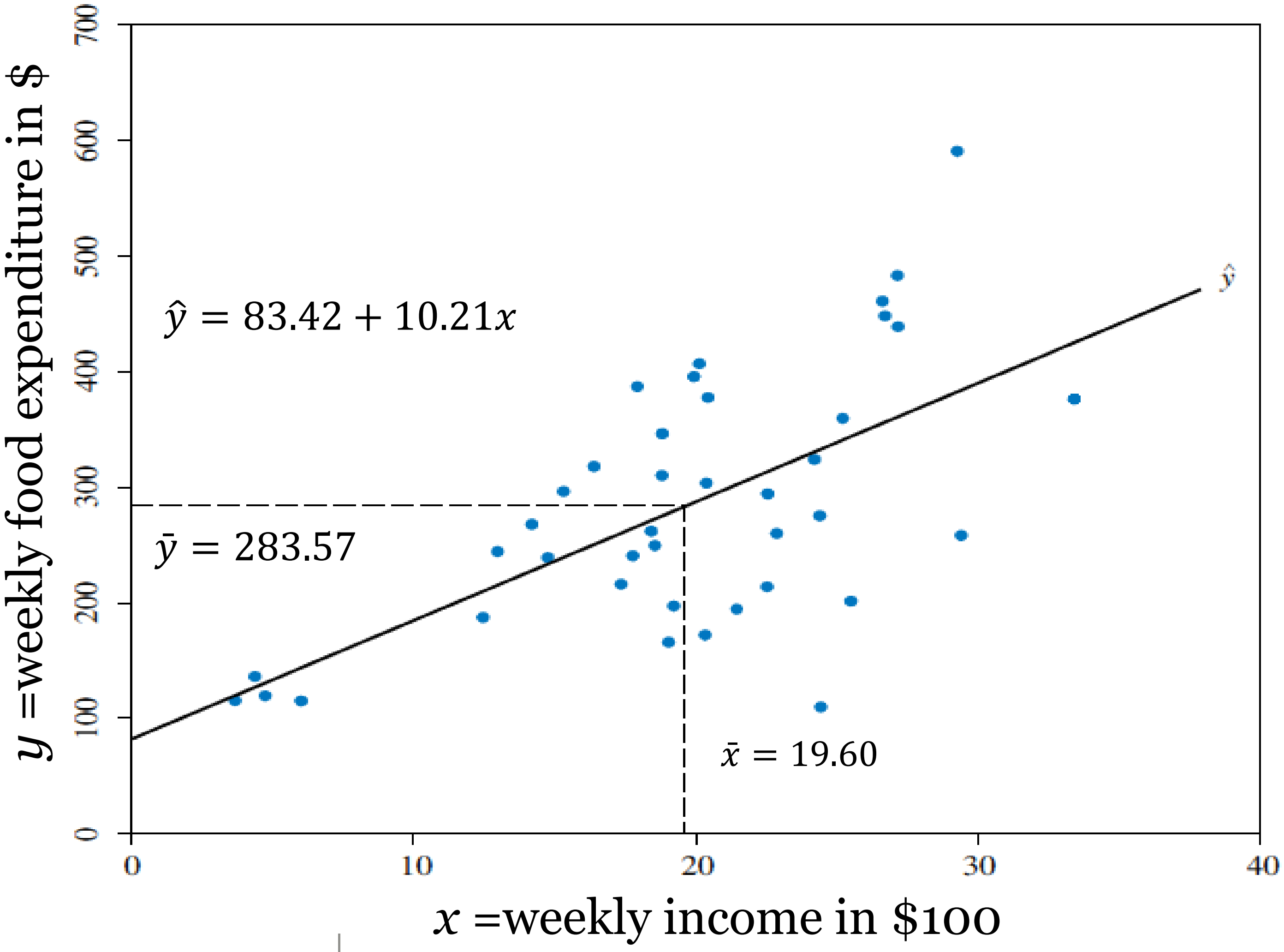
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	83.41600	43.41016	1.92	0.0622
income	weekly household income	1	10.20964	2.09326	4.88	<.0001



Estimating the Regression Parameters

Estimates for the Food Expenditure Function



The population regression model

$$y = \beta_1 + \beta_2 x + e$$

The fitted regression line

$$\hat{y}_i = 83.42 + 10.21x_i$$

The sample regression model

$$y_i = 83.42 + 10.21x_i + \hat{e}_i$$

Interpreting the Estimates

$$\hat{y}_i = 83.42 + 10.21x_i$$

y= weekly food expenditure in \$; x= weekly income in hundred \$

- The value b_2 = 10.21 is an estimate of β_2, the amount by which weekly expenditure on food per household increases when household weekly income increases by \$100.

- Thus, we estimate that if **weekly** income goes up by \$100, the expected (average) weekly expenditure on food will increase by \$10.21

- The intercept estimate b_1 = 83.42 is an estimate of the weekly food expenditure on food for a household with zero income. Does this make sense?

Prediction and Prescription

- Suppose that we wanted to predict weekly food expenditure for a household with a weekly income of $2000. This prediction is carried out by substituting x = 20 into our estimated equation to obtain:

$$\hat{y} = 83.42 + 10.21x_i = 83.42 + 10.21(20) = 287.61$$

- We *predict* that a household with a weekly income of $2000 will spend $287.61 per week on food

- **Prescriptive: How much a family should earn to spend $200 on food expenditure? (Answer: $1142)**

$$200 = 83.42 + 10.21x_i$$

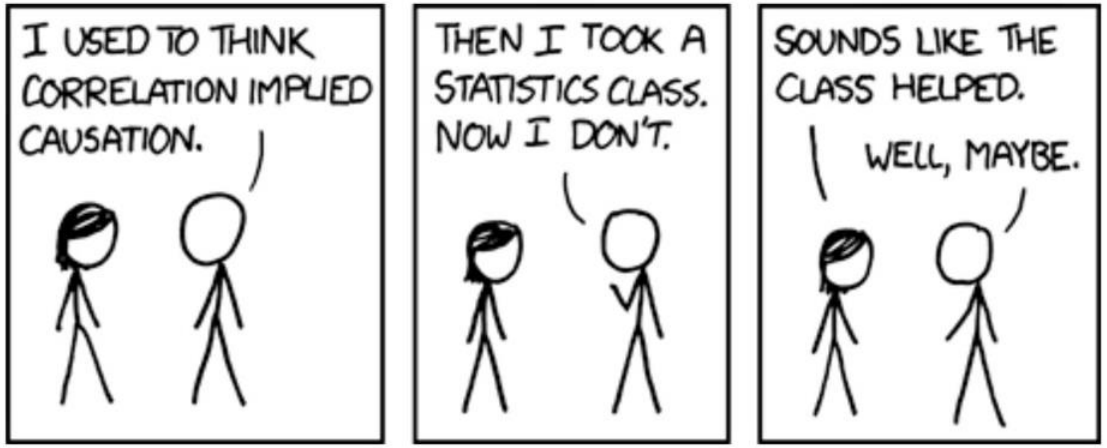
$$x_i = \frac{(200 - 83.42)}{10.21} = 11.418$$

Notes on Regression Analysis 3

Regression versus Causation

- Although regression analysis deals with the dependence of one variable on other variables, it does not necessarily imply causation.

> "A statistical relationship, however strong and however suggestive, can never establish causal connection: our ideas of causation must come from outside statistics, ultimately from some theory or other."


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

https://xkcd.com/552/

- In the expenditure-income example, there is no statistical reason to assume that income does not depends on expenditure. The fact that we treat expenditure as dependent on income (among other things) is due to non-statistical considerations: Common sense suggests that the relationship cannot be reversed, or we cannot control income by varying food expenditure.

Summary

After this lecture, you should be able to:

- differentiate between an economic model and an econometric model
- explain the meaning of random error term
- understand the underlying concept of the least squares method in regression analysis

Preparation for workshops

- Workshop 1: Introduction, Probability and Statistics for Econometrics
  - MCQ and short-answer questions in probability and statistics,
  - You need to do calculations to identify the correct answer

- Workshop 2 to 5
  - MCQ and short-answer questions (8 questions in 30 minutes)
  - You need to run regressions using gretl and SAS® and answer workshop questions based on the regressing results
- All workshop tests are open-book. Bring lecture notes and stationary for calculation.
- Students may work in groups, but it is important that the workshop exercises are undertaken independently. It is important to keep up with the exercises and to attempt practice questions and to study the expositions in the textbooks and lecture notes.

Workshop regulations

- BUSI2053 Workshops are formal assessment weighted 20%. **Therefore, all open-book exam rules and regulations will be applied.**

- **You MUST only attend the group for which you are registered.** However, in exceptional circumstances (e.g. illness) and with appropriate documentation, you may attend any other group subject to space availability and approval of the lecturer or tutor.

- **All workshop sessions are in-person and invigilated.** Attending the group for which you are not registered or attempting the ExamSys assessment outside of the workshop venue will be awarded a mark of zero for that workshop.

- Sharing the password of the workshop assessment with other students is an academic offence and will be investigated.

Thank you.
Any question?