

**EC315 Summary (3):**

**Topics in Microeconomics With Cross Section Econometrics**

**Lewis Britton {201724452}**

EC315: Topics in Microeconomics With Cross Section Econometrics

Academic Year 2019/2020

Word Count: {N/A}

**EC315: Topics in Microeconomics With Cross Section Econometrics**

**Topic Summary**

**Topics:**

1. Exam Summary
2. Game Theory (Externalities & Consequences)
3. Topics in Public Economics (Government Role & Functions)
4. **Cross-Section Economics (Theory & Real World)**

**Cross-Section Econometrics**

**1: Descriptive Statistics**

**1.1: Variables**

Numerical:

* Continuous: infinite possible values (on real line or in an interval)
* Discrete: set value (number of values it can take on are finite (countable))

Categorical:

* Ordinal: ordered and means something
* Regular: ordered but means nothing

Relationships:

* Correlation ≠ Causation
* Associated: roughly connected
* Independent: not connected
* Dependent: depends on another

**1.2: Data Collection**

* Sample: group you’re analysing
* Population: entire group of something

Sampling Bias:

* Non-Responsive: only fraction respond
* Voluntary Response: people feel too strong
* Convenience: more accessible – easier to answer

Explanatory & Response Variables:

* Observations: rather than asking questions
* Experiment: man-made situations

**1.3: Examining Data**

* Scatterplot: allows to identify relationship (e.g. Linear, Pos/Neg)
  + *x*-axis: explanatory variable
  + *y*-axis: response variable
* Dot Plot: shows volume at ends of sample scale
* Return Distribution Moments:
  + 1: Mean
  + 2: Variance
  + 3: Skewness
  + 4: Kurtosis

***1.3.1: Mean***

* Most common value (useful for predicting values etc. such as stock)
* Influenced by outliers so can be skewed inaccurately
* **;**

***1.3.2: Median***

* Value in the middle of the dataset
* Splits 50%’ile (Quartile 2)
  + Q1: 25%, Q2: 50%; Q3: 75%
  + Interquartile Range: Q1 – Q3
* Use here where we don’t want outliers’ influence (e.g. employee salary. Mean misleads due to the CEO’s etc. salary)

***1.3.3: Standard Deviation***

* How far deviated from the mean, is the data
* Same units as data
* “how many std.devs does the data lie from the mean”

***1.3.4: Variance***

* The square of the standard deviation – to fairly weight (e.g. discard negatives)
* Therefore, weights higher deviations more

***1.3.5: Covariance***

* Uses same units as the data – variance of members of the data set relative to others

***1.3.6: Correlation & Correlation Matrix***

* To what degree the data moves together
* A Correlation Matrix maps all individual values with movement relative to all others in a relative **N**by **N**matrix

***1.3.7: Skewness***

* The degree of asymmetry around the mean
* Symmetric: assume mean is centre
  + {mean ≈ median}; {skewness ≈ 0}
* Left Skewness: {Skewness < 0}; tail to the left
  + {mean > median}; **Positive Distribution**
* Right Skewness: {Skewness > 0}; tail to the right
  + {mean < median}; **Negative Distribution**

***1.3.8: Kurtosis***

* Leptokurtic: **Positive Kurtosis**; above Normal Distribution w/ skinny tails
  + {Excess Kurtosis < 0}
* Platykurtic: **Negative Kurtosis**; below Normal Distribution w/ fat tails
  + {Excess Kurtosis > 0}
* Mesokurtic: Normal Distribution
  + {Excess Kurtosis = 0}
* **Excess Kurtosis**: How peaked the data is relative to the Normal Distribution
  + Excess Kurtosis = *{k – 3}*
  + Generally, EK of 1 is significant
* Measure of the peak of data; likelihood of extreme values
* The higher the value of Kurtosis, the more likely you have outliers

***1.3.9: Modality***

* Unimodal: 1 Peak
* Multimodal: > 2 Peaks
* Uniform: No Peaks (outcomes have equal probabilities)

**1.4: Types of Economic Data**

* Time Series: observations of the same unit, different points in time
  + E.g. monthly profits of a firm between 1999 to 2008
* Cross Section: observations of different units, same time period
  + E.g. profits of 256 companies over August 2008
* Panels: several units, varying time (e.g. company, country…)
  + E.g. profits of 256 companies in the financial sector from 1999 to 2008
* **General Notation:**
  + Representing ***n***observations of ***X***

**2: Regression Analysis**

**Recall** (from kindergarten): basic regression:

* ***y*** is the dependent variable (what effects ***y***?)
* ***x*** is the explanatory variable (does it affect ***y***?)
* y-intercept (level of ***y*** when ***x*** = 0)
* slope of the line (severity of the relationship)
* residuals (randomerror term) (outliers from the best fit)
  + Captures any explanation not contained within the explanatory variables
* *Recall also, when a variable has a ‘hat’ accent it is an* ***estimation***

**Know that**:

* + For each (+) unit on the ***x*** axis, expect the ***y*** value to change by
  + Each observation will have a predicted on the best fit line, directly above or below their real

1. This is a Linear Regression and uses a straight best-fit (hyperplane)
   * For non-linear data, a Polynomial Model can be used to account for concave/convex data using *x2* values
   * Hence:

**1**

* The sum of all residuals in the model are equal to zero

**2** ; orin vector format

* + “Observations have constant errors”
  + Homoscedasticity: constant errors
  + Heteroscedasticity: non constant errors (must adjust Robust Model)
  + Where for: ; denotes that variance of the error can be different for each observation
  + The **White Test** can be used to hypothesise and test for this

**3** ;or

* + Error terms should be uncorrelated
  + Expected observations should be uncorrelated with errors
  + **Instrumental Variable Approach**: where explanatories are correlated with error terms
  + *X* may be **Endogenous**,where factors within the model cause changes in *X* therefore, associated with
  + Endogenouscontext: variables correlated with error term
  + Exogenouscontext: variables uncorrelated with error term
  + Use the IV (Instrumental Variable) approach, not OLS (Ordinary Least Squares)
  + If two explanatory variables have high collinearity, omit one

**4** [Extra] Exists

* + Multicollinearity is not present
  + Under multicollinearity, two variables may have high correlation
  + The model would struggle to understand which one explains *Y*
  + Hence, w/o multicollinearity, each explanatory gives unique information

**2.1: Introduction**

* How variation in one variable effects variation in the other
* Make expectations based on regression projections
* Step 1: Determine correlation for relationship
* Step 2: Is the relationship statistically significant?

**2.2: Graphing**

* Must find the **best fit** line to identify Correlation & Relationship
  + Recall: perfect positive, perfect negative, no correlation
  + This will be seen through the **gradient** of the slope
* Recall: Correlation ≠ Causation
  + With no causation however, there can still be a variable in common. Call it ***k*** as it’s unknown and outside the Explanatory and Response variables (***x*** and ***y***)
  + E.g. **hot weather** (***k***) causes **ice cream sales** (***x***) and **seaside deaths** (***y***)
* ***x*** axis: Explanatory Variable
* ***y*** axis: Response Variable

**2.3: The Line**

* Straight Line:
  + ***y*** is the dependent variable (what effects ***y***?)
  + ***x*** is the explanatory variable (does it affect ***y***?)
  + y-intercept (level of ***y*** when ***x*** = 0)
  + slope of the line (severity of the relationship)
* Everyone has one of these lines (variables values will change), we want to aggregate:
  + {Alternatively: }
    - “For each (+) unit on the ***x*** axis, expect the ***y*** value to change by ”

**2.4: Residuals**

* Calculating error of the line: distance between actual observations and the **best fit** line
* Error term: or ***u*** w/ subscript of ***i***
  + {Alternatively: }
* Each individual will have a predicted on the best fit line, directly above or below their real
* Choose line which reduces Aggregated Error across all observations
  + This is such that the sum of ***e*2** is minimised
* We want to minimise the amount of errors:
  + “Finding and which reduces the number of squared residuals”
    - Where here, and elsewhere: **;**
  + Reducing the sum of squared residuals
    - Hence, reducing

**2.5: Including Multiple Explanatory Variables**

* Fitting a Hyperplane:
* This means that are all the explanatory variables
* Here (in supplementary slideshow):

**2.6: Conditions & Assumptions**

* Assume to be Linear (not quadratic etc.)
* Assume Nearly Normal distribution (closest to line as possible: {Exp. Error 0})

**2.7: R2 Value:**

* The square of the Correlation Coefficient {0 < R2 < 1}
* % of variability in dependent variable (***y***) attributed to explanatory variable (***x***)
* R2 increases when you increase Explanatory Variables (doesn’t mean better model)
  + Hence, don’t use if Multiple Regression; use Adjusted R2
  + Interpret in the same way

**2.8: P-Value**:

* 3(\*\*\*): Certain at 99% (1% Significance Level; P-Value < 0.01)
* 2(\*\*): Certain at 95% (5% Significance Level; 0.01 < P-Value < 0.05)
* 1(\*): Certain at 90% (10% Significance Level; P-Value < 0.10 )
* (No Stars): Insignificant (Statistically Insignificant; 0.10 < P-Value)

**3: Multiple Regression & Goodness of Fit**

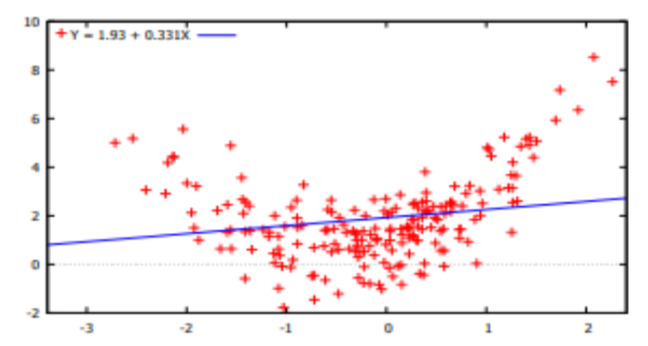
1. This simply adds more Explanatory Variables
2. Explore the extent to which the model explains the data (R2)

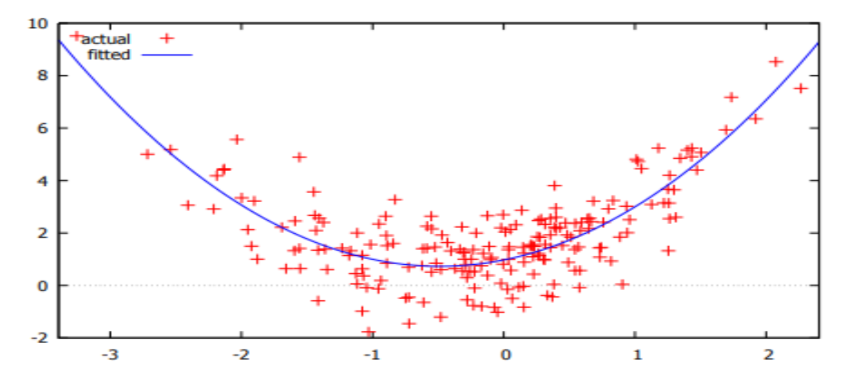
**3.1: ‘Aggregation’ & Adding More Explanatories**

* Expand: :
* For Explanatories:
* Note that it may not always be linear (Hyperplane)
  + Not all variables have linear relationships
  + Concavity/Convexity etc.

**3.2: Nonlinearity**

* Non-linear data can be captured in a linear model
* E.g. the Concave
* Hence, linear model does a poor job of explaining the data



* This can be fixed with Polynomial Models
* Hence:
  + Constant remains ()
  + Error term remains ()
* Thus, keep adding Polynomial Terms (*x2* values) until the model best fits the spread
* ****Hence:

**3.3: Dummy Variables**

* Dummy Variables do not have values
* They are Categorical: e.g. Male/Female or Retired/Employed
  + You expect one group to show different levels of ***y*** for any ***x***
  + They’re purely binary {1 or 0}
    - 1: Observation comes from the group of ‘interest’
    - 0: Null response thus, otherwise
* Modify Regression:
  + Where:
  + If: (Male (Null) in e.g.)
  + If: (Female in e.g.)
    - Hence, intercept alters

**3.4: Changes in Slope**

* E.g. expenditure patters may be at two extremes
* Take a Dummy Variable with criteria to identify all people who spend > 2000
  + 1: Spend > 2000 (e.g.)
  + 0: Spend < 2000 (e.g.)
* Thus: Dummy Variable for
  + If: then 1
  + If: then 0
* Modify Regression:
  + Where:
  + If: **;**
  + If: 
    - Hence, intercept alters

**3.5: Interpretation**

* Simple Regression: If the Explanatory Variable changes by 1 unit, how much does the Reliant Variable change?
* Multiple Regression: If the Explanatory Variable changes by 1 unit, how much does the Reliant Variable change, given all other Explanatory Variables are constant – work your way along all of the Beta values holding each other constant

**3.6: Hypothesis Testing**

* Null Hypothesis **H0**: R2 = 0; ***X*** doesn’t have any explanatory power for ***Y***
* Alternative Hypothesis **H1**: R2 ≠ 0; Reject Null in favour of Alternative

**3.7: Multicollinearity**

* When two variables have a high correlation (close to 1 or -1)
* The model struggles to understand which one is actually explaining ***Y***
* Run a Multicollinearity test for the Matrix
* Drop a highly correlated variable

**3.8: Choosing Explanatory Variables**

1. Use hypothesis testing
2. Test for significance and omit insignificant ones
   * If significant and omitted, Omitted Variables Bias

**3.9: Choosing Models**

1. Schwartz Information Criterion
2. Akaike Information Criterion
3. Hannan-Quinn Information Criterion

* Pick one
* Compare across models
* Select the lowest value

**4: Theory**

* Probability Theory & relationship to Econometrics
  + Expected Values
  + Variance
  + Probability Distribution (Density Functions)
* Problem: taking several Samples from the same population means estimates will change from sample to sample so not represent the Population correctly
* If we only have one sample: how significant are the estimates to the Population?

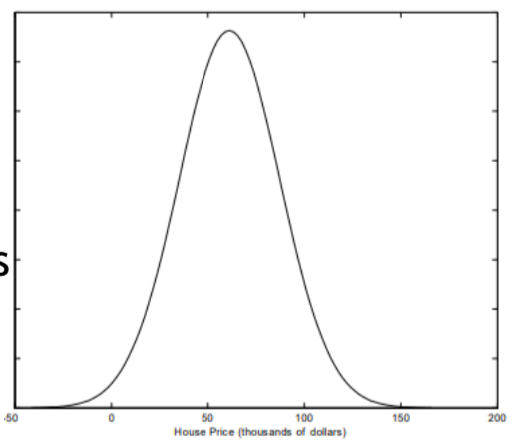
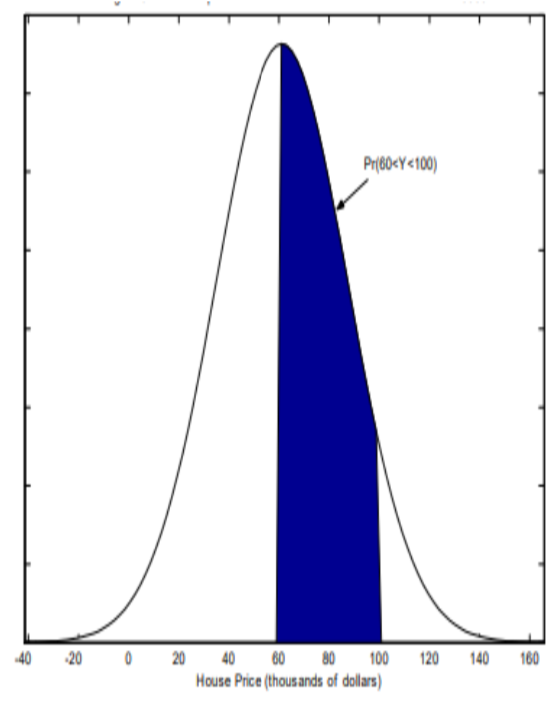
**4.1: Experiments & Events**

* “An outcome unknown in advance”
* Possible outcomes (realisations) of experiments: Events
  + i.e. predict positive relationship
* Set of all possible outcomes: Sample Space
* Variables of Experiments & Events are:
* Recall: Discrete: set value (number of values it can take on are finite (countable))
  + Depends on scale of variability (e.g. Happy? Rate: 0-M&S)
  + If counter-intuitive (e.g. Happy? Rate M&S-0) cant interpret like Continuous
* Recall: Continuous: infinite possible values (on real line or in an interval)

**4.2: Random Variables & Probability**

* A variable through which we don’t know the outcome (e.g. ***Y*** on a regression)
* Probability reflects likelihood of an event
  + E.g. just knowing what income is doesn’t allow to you know expenditure
    - Income = 1000; Consumption **not** > 1000
  + E.g. Probability of A occurring denoted:
* **Example**:
  + Dice, probability of rolling any six options: **Constant** Probability
  + Sample Space: {1,2,3,4,5,6}
  + The Discrete Random Variable (A): {1,2,3,4,5,6}
    - Same probability of rolling any face
  + Hence, Probabilities:
  + Realisation of random variable is value which actually arises
  + Independence: events A and B are Independent so:
  + Conditional: event A may be Conditional upon B so:
    - Probability of A occurring given B occurs
  + With Continuous Random Variables use notation:

**4.3: Probability in Regression**

* Regression provides description of the probable values of the dependent variable
* Hence, we use Probability Density Functions (p.d.f.)
  + Used with Continuous Normal Variables
  + Probabilities are the number under the Normal Distribution function
* **Example:**
  + ****
  + Tells you which plausible values that ***y*** can take given the set ***x*** value
  + At the highest point, we see the most plausible values
* The shape of the distribution depends on the Mean and the Variance
* “***Y*** has a Normal Probability Density Function”
  + Mean =
  + Variance =
  + Normal p.d.f. =
* Recall House Price Example:
  + (Mean value of a house of lot size > 5000)
  + (Not really any intuitive value)
* Defined areas under the p.d.f. curve are the Probabilities
* “Probability of price being between 60k and 100k”:

**4.4: Other Distributions**

***4.4.1: Chi-Distribution***

* Distribution depending on the Degrees of Freedom (accounts for number of observations and variables)
  + Higher the better 🡪 more flexibility
  + Denoted by
  + Skewness decreases with the raising Degrees of Freedom
* Not bell-shaped like Normal Distribution
  + Only for the positive values of

***4.4.1: t-Distribution***

* How we calculate the p-Value
* Shows how significant values are
* Symmetric
* Compare from (Critical Value) -1.96 to 1.96
  + If 0 sits in the centre, Normally Distributed
* “If **t-Value** is > **Critical Value**, explanatory variables are statistically significant”

**4.5: Assumptions of a Regression (OLS)**

1. Expect dependent variable to lie on the best fit
2. all observations should have constant errors
   * Homoscedasticity: constant errors
   * Heteroscedasticity: non constant errors (must adjust model)
3. Expecting observations to be uncorrelated
   * If two explanatory variables have high collinearity, omit one (run corr. matrix)
4. Expect errors are normally distributed (not a lot of outliers)
5. Explanatories are fixed

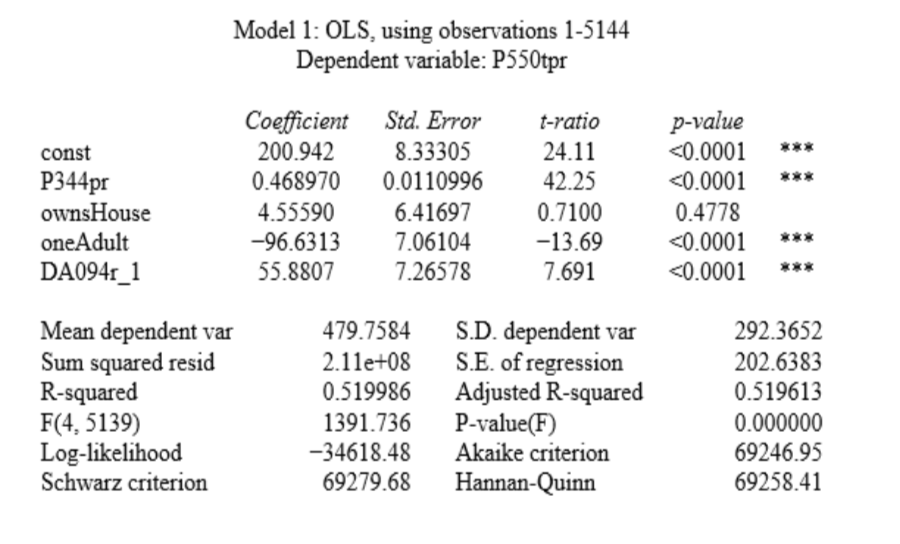
**Rough Notes on Interpretation \*\*TEST\*\*:**

* Don’t interpret constant
* Four explanatory variables follow below, for each:
* Dummies (oneAdult, ownsHouse)
  + If you are not included in the criteria: 0
  + If you are included in the criteria: 0
    1. **P-Vale**: is it significant?
  + Here, {P < 0.01 @ 0.0001} so “Statistically Significant at explaining the dependant variable at the 1% Significance Level”

1. **Coefficients**:
   * As significant, look at coefficients. Don’t analyse if statistically insignificant
   * If Positive Corr. (Converse): as explanatory increases, reliant increases
   * If Negative Corr. (Inverse): as explanatory increases, reliant decreases

* Examples:
  + A One Adult Household is significant at 1% significance level and spends £96.63 less (-96.6313)
  + A higher managerial occupied man is significant at 1% and spends %55.88 more

1. Bottom (only interested in a few)
   * **Mean Dependent**: for interest
   * **R2** or: goodness of fit (% of variability in ***y*** which can be explained by ***x***)
     + E.g. 52% of the variability in expenditure can be explained by the dependent variables
     + But! Use adjusted as there are multiple explanatory variables



**5: Hypothesis Testing**

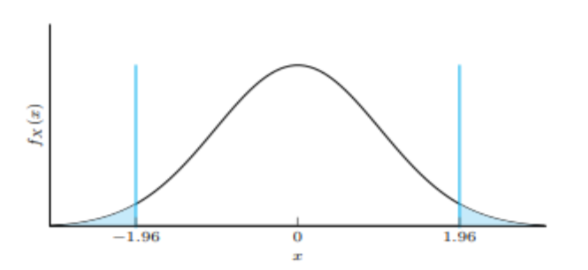
**5.1: What is a Hypothesis**

1. t-test
2. f-test
3. RESET Test

* Suppose (estimated coefficient); is that significantly different from 0.5?
* We must know distribution/density function of
* There are two Hypotheses (e.g. yes/no)
* We want to test if this variable is significant or insignificant
* **H0:** Null Hypothesis ()
  + Unable to reject Null Hypothesis
  + “explanatory variable is insignificant in explanation of dependent variable”
  + If p-value > 0.1: Unable to reject
* **HA:** Alternative Hypothesis ()
  + Reject Null Hypothesis
  + “explanatory variable is significant in explanation of dependent variable”
  + If p-value < 0.1: Reject in favour of alternative
* Type I Error: Reject Null when it’s in fact true
* Type II Error: Fail to Reject Null when it is in fact false
* \*\*As long as p-value < 0.1, these won’t occur\*\*

**5.2: t-test**

***5.2.1: t-ratio***



* If Null Hypothesis Failed Rejection:
* If Null Hypothesis Rejected:

***5.2.2: p-values***

* p-value is the probability that, under H0, the test value is at least as large as
* If probability (Significance Level) > p-value: Fail to Reject Null
* If probability (Significance Level) < p-value: Reject Null in favour of Alternative
* **Example:**
* ; show: (Significance Level) @ defined p-value
  + If p-value = 0.05: Reject as 0.022 < 0.05
  + If p-value = 0.01: Fail to Reject as 0.022 > 0.01

**5.3: f-test**

* A joint test for the whole regression: H0: R2 = 0
* Must reject in favour of the Alternative Hypothesis
  + Observe f-ratio: like t-ratio but for regression as a whole
  + Observe p-value: in the same way as the individual p-values
* This is tested against the 5% level
  + If p-value < 0.05: Reject Null so some significance
  + If p-value > 0.05: Fail to Reject Null
* If single regression: p-value (f-test) = p-value (t-test)

**5.4: RESET Testing**

* Is your model well specified (do not needing logs or polynomials)
* Hypothetically adds gamma coefficients to hypothetical log and polynomial values
* Hence: H0: Gamma = Gamma2 = 0

***5.4.1: RESET in Gretl***

* After the Regression;
* Tests;
* Ramsay’s RESET;
* Squares and Cubes;
* Don’t Interpret Coefficients;
* F-test for Gamma Polynomials;
* P-value at Bottom;
* Use 5% level;
* If p-value < 0.05: model mis-specified (needs extra like polynomials and logs etc.)
* If p-value > 0.05: model well-specified (does not need polynomials etc.)
* Opposite of what we conclude about p-values in general
* If mis-specified: try logs and polynomials

**6: Instrumental Variables**

* Drop assumption that explanatory variables are fixed, they **aren’t**
* Random explanatories don’t cause problems unless correlated with the error (*u*)
* Don’t use OLS (Ordinary Least Squares), use alternative IV (Instrumental Variables) estimator 2SLS (Two Stage Least Squares)
  + This incorporates everything that the model doesn’t include (unknown coefficients where variable correlated with *u*)
  + E.g. all else when your using income to explain consumption (age etc.)
* **Example:**
  + Earnings: dependent; Schooling: explanatory; Error: *u*
  + ; where *u* captures all explanation not done by schooling (which is usually higher with people of **higher ability**) 🡪 E.g. ability can also effect
  + If error value is high: “high un-associated explanation”
  + Endogeneity: “factors within the model cause *x* to change, so changes in *x* are also associated with changes in *u*”
  + What would *x* have been if not measures with error *u*, as *x* is higher than it should be as correlated to *u*. 🡪 **2SLS**

**6.1: Introducing Instrumental Variables**

* Solution to the above problem
* ***z*** = Instrumental Variable
* Isolates movement in *x* which is uncorrelated to the error *u* (e.g. ability)
* Hence, coefficient will no longer be inflated
* Endogenous: variables correlated with error term *u*
* Exogenous: variables uncorrelated with error term *u*

**6.2: Variation of *x***

* 2 Parts: one is correlated with *u* and second is uncorrelated with *u*
* Isolate the uncorrelated with *u*
* The uncorrelated parts are included in 1 to N *z* values for each explanatory
* An IV only influences *y* through an explanatory, it wouldn’t hold up as an explanatory itself
  + E.g. ability effects schooling (*x*) but not directly income (*y*)

**6.3: Instrumental Variable Satisfactions**

* *z* is correlated with (Endogenous) *x*
* *z* is uncorrelated with *y* (*z 🡪 x 🡪 y*; **not** *z 🡪 y*)

**6.4: Two Stage Least Squares (2SLS)**

* *y*: Dependent Variable
* *x1…k*: Endogenous Variables
* *w1…k*: Exogenous Variables
* *z1…k*: Instrumental Variables
* *“For every x you expect to be Endogenous, you require an Instrumental Variable”*

1. Regress *x* on *z1…k* values for *x1…k* and obtain **predicted** values:
2. Regress *y* on *w1…k* (don’t need Instrumental Variables)
3. Look for high correlation between Instrumental Variables and Explanatory Variables

**6.5: Testing for Endogeneity**

* “Hausman Test”
* H0: Explanatory Variable uncorrelated with error term
  + Fail to Reject: use OLS
  + Reject: use Instrumental Variables for 2SLS
* p-value > 0.05 Means no **Endogeneity** problem (Fail to Reject)
* p-value < 0.05 Means **Endogeneity** problem (Reject)
* *Like RESET Test*

***6.5.1: Strength of Instruments***

* High Correlation with (Endogenous) *x*: Strong Instrument – use **2SLS**
* Low Correlation with (Endogenous) *x*: Weak Instrument – might as well use **OLS**
* **Relevant?**
  + R2 shows this integrity
  + f-test shows validity of the set of instruments as a whole (like OLS)

**7: Robust Estimation**

* Characteristics:
  + Heteroscedasticity (as opposed to Homoscedasticity)
  + Cross-Sectional Correlation
* As possible results show:
  + Affect reliability of hypothesis tests
  + Don’t introduce significant bias in estimates
* Recall:
  + All regression errors have equal variance

**7.1: Heteroscedasticity & Homoscedasticity**

* Heteroscedastic: Non-Constant Error Variance
* Homoscedastic: Constant Error Variance
* A close up of a map

  Description automatically generated**Example:**
  + Hence, Homoscedastic (no pattern)
* A close up of a map

  Description automatically generated**Example:**
  + Hence, Heteroscedastic (pattern)
  + OLS regression (left) does good job when income is low but lacks explanatory value as income increases
  + E.g.: when income increases, expenditure may only increase a little and savings may take place instead

***7.1.1: Homoscedasticity***

* House price dataset:
  + Dependent variable: house price
  + Explanatory variables: bedrooms, bathrooms etc.
* measures whether a house is under or over-priced relative to similar houses
* Homoscedasticity doesn’t say all errors are same for every house but, that they’re from the same distribution
  + “Magnitudes of under or over-pricing tend to be the same for all kinds of houses”

***7.1.2: Heteroscedasticity***

* + For: ; denotes that variance of the error can be different for each observation
* **Implications**:
  1. Least squares estimates are unbiased and/or inefficient
  2. Variances and covariances need reconstrained
  3. t-tests and f-tests lose validity so don’t represent good p-values

**7.2: Test for Heteroscedasticity in Gretl**

* Solving problem (3)
* Making standard errors ‘robust’
* **White Test** 
  + Using ‘White’, ‘Robust’, ‘Heteroscedasticity Consistent (HC)’ standard errors
  + New t-ratios, p-values
  + If Heteroscedasticity is not present, OLS fine (**BLUE**)
  + If Heteroscedasticity is present, use robust std. errors (**HCE**)
  + H0: Homoscedastic Constant Error Variance
  + HA: Heteroscedastic Non-Constant Error Variance
  + p-value < 0.05: Reject Null Hypothesis (Reject Homoscedasticity)
  + p-value > 0.05: Fail to Reject Null Hypothesis (Accept Homoscedasticity)
  + In response to Heteroscedasticity, tick **Robust Standard Errors**

1. Run Model
2. Use White Test
3. Analyse p-value
4. Tick **Robust Standard Errors** in OLS Window

**7.3: Cross-Sectional Dependence**

* Samples are completed in ‘clusters’
* Clusters from different classes, clusters from different countries, different firms etc.
* Data can therefore be highly Correlated due to similar traits
* Use **Cluster Robust Standard Errors**