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}

University of Strathclyde Glasgow

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
- Population Mean: $\boldsymbol{\mu} = \frac{\Sigma \mathrm{x}}{\mathrm{T}}$; Sample Mean: $\overline{\boldsymbol{x}} = \frac{\Sigma x}{n}$

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

- $\boldsymbol{\sigma}$
- 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

- $\boldsymbol{\sigma}^2$
- The square of the standard deviation – to fairly weight (e.g. discard negatives)
- Therefore, weights higher deviations more

1.3.5: Covariance

- $\boldsymbol{cov_{Rx,Ry} = \sigma_x \sigma_y \rho}$
- Uses same units as the data – variance of members of the data set relative to others

1.3.6: Correlation & Correlation Matrix

- $\boldsymbol{\rho = \frac{cov_{Rx,Ry}}{\sigma_x, \sigma_y}}$
- Where [$\boldsymbol{\rho = 1}$]: Perfect Positive Correlation (Together)
- Where [$\boldsymbol{\rho = -1}$]: Perfect Negative Correlation (Apart)
- Where [$\boldsymbol{\rho = 0}$]: No Correlation
- 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
  - P_t for $t = 1, 2, \dots, T$
  - E.g. monthly profits of a firm between 1999 to 2008
- Cross Section: observations of different units, same time period
  - P_i for $i = 1, 2, \dots, N$
  - E.g. profits of 256 companies over August 2008
- Panels: several units, varying time (e.g. company, country…)
  - $P_{i,t}$ for $\begin{Bmatrix} i = 1, 2, \dots, N \\ t = 1, 2, \dots, T \end{Bmatrix}$
  - E.g. profits of 256 companies in the financial sector from 1999 to 2008

- **General Notation:**
  - X_i for $i = 1, 2, \dots, n$; otherwise: $X_1, X_2, \dots, X_n$
  - Representing n observations of X

2: Regression Analysis

Regression Summary

Recall (from kindergarten): basic regression: $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + \varepsilon_i$

- y is the dependent variable (what effects y?)
- x is the explanatory variable (does it affect y?)
- α y-intercept (level of y when $x = 0$)
- β slope of the line (severity of the relationship)
- ε residuals (random error 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:

1) $\beta = \frac{\Delta y}{\Delta x} = \frac{\partial y}{\partial x}$
   - For each (+) unit on the x axis, expect the y value to change by β
2) $\varepsilon_i = y_i - \hat{y}_i$
   - Each observation will have a predicted $\hat{y}_i$ on the best fit line, directly above or below their real y_i
3) 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 x^2 values
   - Hence: $Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_k X_i^k + \varepsilon_i$

Assumptions:

1) $E(\varepsilon) = 0$
   - "Expected error terms are equal to zero"
   - Therefore, expected dependent variable to lie on line of best fit
2) $var(\varepsilon_i) = E(\varepsilon_i^2) = E(\varepsilon\varepsilon') = \sigma^2 I$
   - "Observations have constant errors"
   - Homoscedasticity: constant errors
   - Heteroscedasticity: non constant errors (must adjust → Robust Model)
     - $var(u) = \sigma^2 \omega_i^2$
     - Where for: $i = \{1, \dots N\}$; i 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) $cov(\varepsilon_i, \varepsilon_j) = 0$ $(\text{for } i \neq j)$; or $E(X\varepsilon') = 0$
   - 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, ΔX associated with $\Delta\varepsilon$
- Endogenous context: variables correlated with error term ε
- Exogenous context: variables uncorrelated with error term ε
- Use the IV (Instrumental Variable) approach, not OLS (Ordinary Least Squares)
- If two explanatory variables have high collinearity, omit one

4) [Extra] $(\boldsymbol{X'X})^{-1}$ Exists
   - Multicollinearity is not present
   - Under multicollinearity, two variables may have high correlation $(1, -1)$
   - 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 $\boldsymbol{k}$ as it's unknown and outside the Explanatory and Response variables ($\boldsymbol{x}$ and $\boldsymbol{y}$)
  - E.g. **hot weather** ($\boldsymbol{k}$) causes **ice cream sales** ($\boldsymbol{x}$) and **seaside deaths** ($\boldsymbol{y}$)
- $\boldsymbol{x}$ axis: Explanatory Variable
- $\boldsymbol{y}$ axis: Response Variable

2.3: The Line

- Straight Line: $\boldsymbol{y = \alpha + \beta x}$
  - $\boldsymbol{y}$ is the dependent variable (what effects $\boldsymbol{y}$?)
  - $\boldsymbol{x}$ is the explanatory variable (does it affect $\boldsymbol{y}$?)
  - $\boldsymbol{\alpha}$ y-intercept (level of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$)
  - $\boldsymbol{\beta}$ slope of the line (severity of the relationship)
- Everyone has one of these lines (variables values will change), we want to aggregate:
  - $\boldsymbol{\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i}$
  - {Alternatively: $\boldsymbol{\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x}$}
    - Recall: $\boldsymbol{i}$ represents $\boldsymbol{N}$ individuals

- Recall: $\widehat{k}$ (hat) represents estimation
- Recall: $\boldsymbol{\beta} = \frac{\Delta y}{\Delta x} = \frac{\partial y}{\partial x}$
  - "For each (+) unit on the $\boldsymbol{x}$ axis, expect the $\boldsymbol{y}$ value to change by $\boldsymbol{\beta}$"

2.4: Residuals

- Calculating error of the line: distance between actual observations and the **best fit** line
- Error term: $\boldsymbol{\varepsilon_i}$ or $\boldsymbol{u}$ w/ subscript of $\boldsymbol{i}$
- $\boldsymbol{\varepsilon_i = y_i - \widehat{y}_i}$
  - {Alternatively: $\boldsymbol{\varepsilon_i = y_i - \widehat{y}_i}$}
- Each individual will have a predicted $\boldsymbol{\widehat{y}_i}$ on the best fit line, directly above or below their real $\boldsymbol{y_i}$
- Choose line which reduces Aggregated Error across all observations
  - This is such that the sum of $\boldsymbol{e}^2$ is minimised:

- We want to minimise the amount of errors:
  - $\boldsymbol{\sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{N}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}$
  - "Finding $\boldsymbol{\widehat{\alpha}}$ and $\boldsymbol{\widehat{\beta}}$ which reduces the number of squared residuals"
    - Where here, and elsewhere: $\boldsymbol{\widehat{\alpha} = \widehat{\beta}_0; \widehat{\beta} = \widehat{\beta}_1}$
  - Reducing the sum of squared residuals
    - Hence, reducing $\sum_{i=1}^{N} \boldsymbol{\varepsilon_i^2}$

2.5: Including Multiple Explanatory Variables

- Fitting a Hyperplane:
  - $\min_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{N} \boldsymbol{\varepsilon_i^2} \equiv \min_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{N}(y_i - \boldsymbol{\widehat{\beta}_0} - \sum_{k=1}^{k} \boldsymbol{\beta_k x_{k,i}})^2$
- This means that $\boldsymbol{\beta_0, \beta_1, \ldots, \beta_k}$ are all the explanatory variables
- Here (in supplementary slideshow): $\boldsymbol{u_i = \varepsilon_i}$

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: R^2 Value:
- The square of the Correlation Coefficient $\{0 < R^2 < 1\}$
- % of variability in dependent variable ($\boldsymbol{y}$) attributed to explanatory variable ($\boldsymbol{x}$)
- R^2 increases when you increase Explanatory Variables (doesn't mean better model)
  - Hence, don't use if Multiple Regression; use Adjusted R^2
  - 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 (R^2)

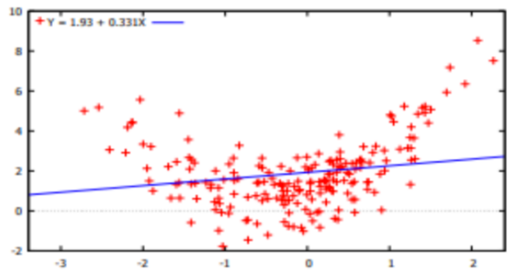
3.1: 'Aggregation' & Adding More Explanatories

- Expand: : $\boldsymbol{y = \alpha + \beta x}$
- For Explanatories: 1 to k
- $\boldsymbol{Y_i = \alpha + \beta_1 X_i + \beta_2 X_i + \cdots + \beta_k X_i + \varepsilon_i}$

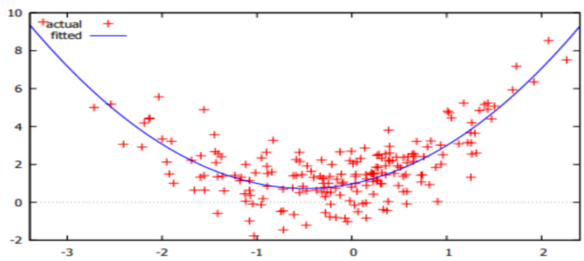
- Note that it may not always be linear (Hyperplane)
  - Not all variables have linear relationships
  - Concavity/Convexity etc.

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: $\boldsymbol{Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_k X_i^k + \varepsilon_i}$
  - Constant remains ($\boldsymbol{\alpha}$ or $\boldsymbol{\beta_0}$)
  - Error term remains ($\boldsymbol{\varepsilon_i}$)
- Thus, keep adding Polynomial Terms (x^2 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 $\boldsymbol{y}$ for any $\boldsymbol{x}$
    - They're purely binary {1 or 0}
        - 1: Observation comes from the group of 'interest'
        - 0: Null response thus, otherwise

- Modify Regression: $\boldsymbol{Y_i = \alpha = \beta_1 X_i + \tau D_i + \varepsilon_i}$
    - Where: $\boldsymbol{\tau} =$ Coefficient of Dummy Variable
    - If: $\boldsymbol{D_i = 0}$; $\boldsymbol{Y_i = \alpha + \beta_1 X_i + \varepsilon_i}$ (Male (Null) in e.g.)
    - If: $\boldsymbol{D_i = 1}$; $\boldsymbol{Y_i = \alpha + \beta_1 X_i + \tau + \varepsilon_i}$ (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 $\boldsymbol{X} \geq \boldsymbol{x}$
    - If: $\boldsymbol{X} \geq \boldsymbol{x}$ then 1
    - If: $\boldsymbol{X} < \boldsymbol{x}$ then 0

- Modify Regression: $\boldsymbol{Y_i = \alpha + \beta_1 X_i + \beta_2 (X_i \bullet D_i(x)) + \varepsilon_i}$
    - Where: $\boldsymbol{D_i(x)} =$ Dummy Variable
    - If: $\boldsymbol{D_i(x) = 0}$; $\boldsymbol{Y_i = \alpha + \beta_1 X_i + \varepsilon_i}$
    - If: $\boldsymbol{D_i(x) = 1}$; $\boldsymbol{Y_i = \alpha + \beta_1 X_i + \beta_2 X_i + \varepsilon_i}$
        - Hence, intercept alters

3.5: Interpretation

- Simple Regression: If the Explanatory Variable changes by 1 unit, how much does the Reliant Variable change?
    - $\boldsymbol{\beta_j}$ is the marginal effect of $\boldsymbol{x}$ on $\boldsymbol{y}$
- 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
    - $\boldsymbol{\beta_j}$ is the marginal effect of $\boldsymbol{X_i}$ on $\boldsymbol{Y}$

3.6: Hypothesis Testing

- Null Hypothesis $\mathbf{H_0}$: $R^2 = 0$; ***X*** doesn't have any explanatory power for ***Y***
- Alternative Hypothesis $\mathbf{H_1}$: $R^2 \neq 0$; 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

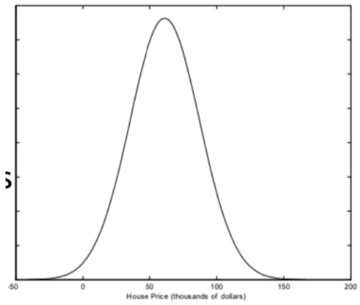
- A variable through which we don't know the outcome (e.g. $\boldsymbol{Y}$ on a regression)
- Probability reflects likelihood of an event
  - E.g. just knowing what income is doesn't allow to you know expenditure
    - Income = 1000; Consumption **not** > 1000
  - E.g. Probability of A occurring denoted:
  - $\boldsymbol{Pr(A)}$

- **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: $Pr(A=1) = Pr(A=2) = \cdots = [A = 1/6]$
  - Realisation of random variable is value which actually arises
  - Independence: events A and B are Independent so:
    - $\Pr(A) = \Pr(B) = \Pr(A, B)$
  - Conditional: event A may be Conditional upon B so:
    - Probability of A occurring given B occurs
    - $\Pr(A|B)$

- o With Continuous Random Variables use notation: $\mathrm{p}(A|B); \mathrm{p}(A,B); \mathrm{p}(\mathrm{B})$

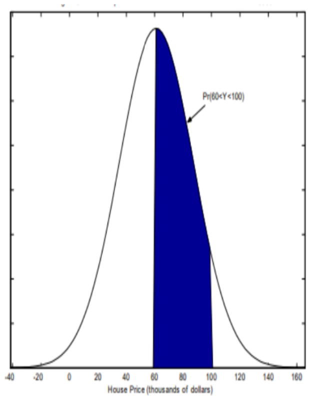
4.3: Probability in Regression

- 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 $\boldsymbol{y}$ can take given the set $\boldsymbol{x}$ value
  - At the highest point, we see the most plausible values

- The shape of the distribution depends on the Mean and the Variance
- "$\boldsymbol{Y}$ has a Normal Probability Density Function"
  - Mean = $\boldsymbol{\mu}$
  - Variance = $\boldsymbol{\sigma^2}$
  - Normal p.d.f. = $\boldsymbol{y \sim N(\mu, \sigma^2)}$
- Recall House Price Example:
  - $\boldsymbol{\mu = 61.153}$ (Mean value of a house of lot size > 5000)
  - $\boldsymbol{\sigma^2 = 683.812}$ (Not really any intuitive value)
- Defined areas under the p.d.f. curve are the Probabilities
- "Probability of price being between 60k and 100k":
  - $\boldsymbol{N(\mu, \sigma^2)}$
  - $= \mathbf{Pr}(\boldsymbol{min \leq k \leq max})$
  - $= \mathbf{Pr}(\mathbf{60} \leq \boldsymbol{Y} \leq \mathbf{100})$
  - $= \mathbf{Pr}\left(\frac{\mathbf{60}-\boldsymbol{\mu}}{\boldsymbol{\sigma}} \leq \frac{\boldsymbol{Y}-\boldsymbol{\mu}}{\boldsymbol{\sigma}} \leq \frac{\mathbf{100}-\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right)$
  - $= \mathbf{Pr}\left(\frac{\mathbf{60-61.153}}{\sqrt{\mathbf{683.812}}} \leq \frac{\boldsymbol{Y}\mathbf{-61.153}}{\sqrt{\mathbf{683.812}}} \leq \frac{\mathbf{100-61.153}}{\sqrt{\mathbf{683.812}}}\right)$
  - $-------------------$
  - $= \mathbf{Pr}(\mathbf{0 \leq Z \leq 0.04}) \rightarrow \mathbf{0.016}$
  - $= \mathbf{Pr}(\mathbf{0 \leq Z \leq 1.49}) \rightarrow \mathbf{0.4319}$
  - $*\,\textit{These are the two independent areas}\,*$
  - $---------------------$
  - (+ Together) = $\mathbf{0.4479}\ \therefore \mathbf{45\%}$



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 $\boldsymbol{df}$
  - Skewness decreases with the raising Degrees of Freedom
- Not bell-shaped like Normal Distribution
  - Only for the positive values of $\boldsymbol{x}$

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) $\boldsymbol{E(\varepsilon_i) = 0 \; Mean}$ Expect dependent variable to lie on the best fit
2) $\boldsymbol{var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2}$ all observations should have constant errors
   - Homoscedasticity: constant errors
   - Heteroscedasticity: non constant errors (must adjust model)
3) $\boldsymbol{cov(\varepsilon_i, \varepsilon_j) = 0 \text{ for } i \neq j}$ 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”
2) **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

3) Bottom (only interested in a few)
    - **Mean Dependent**: for interest
    - **R^2** 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

Model 1: OLS, using observations 1-5144
Dependent variable: P550tpr

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	200.942	8.33305	24.11	<0.0001	***
P344pr	0.468970	0.0110996	42.25	<0.0001	***
ownsHouse	4.55590	6.41697	0.7100	0.4778	
oneAdult	−96.6313	7.06104	−13.69	<0.0001	***
DA094r_1	55.8807	7.26578	7.691	<0.0001	***

Mean dependent var	479.7584	S.D. dependent var	292.3652
Sum squared resid	2.11e+08	S.E. of regression	202.6383
R-squared	0.519986	Adjusted R-squared	0.519613
F(4, 5139)	1391.736	P-value(F)	0.000000
Log-likelihood	−34618.48	Akaike criterion	69246.95
Schwarz criterion	69279.68	Hannan-Quinn	69258.41

5: Hypothesis Testing

5.1: What is a Hypothesis

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

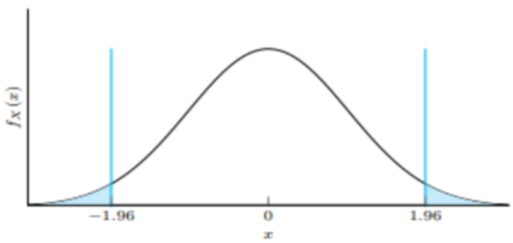
- Suppose $\hat{\boldsymbol{\beta}} = \mathbf{0.47}$ (estimated coefficient); is that significantly different from 0.5?
- We must know distribution/density function of $\hat{\boldsymbol{\alpha}}$ & $\hat{\boldsymbol{\beta}}$

- There are two Hypotheses (e.g. yes/no)
- We want to test if this variable is significant or insignificant
- **H_0:** Null Hypothesis ($\boldsymbol{\beta} = \mathbf{0}$)
  - Unable to reject Null Hypothesis
  - "explanatory variable is insignificant in explanation of dependent variable"
  - If p-value > 0.1: Unable to reject
- **H_A:** Alternative Hypothesis ($\boldsymbol{\beta} \neq \mathbf{0}$)
  - 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: $\hat{t}$ close to 0
- If Null Hypothesis Rejected: $\hat{t} < -1.96$ or $\hat{t} > 1.96$

5.2.2: p-values

- p-value is the probability that, under H_0, the test value is at least as large as $\hat{t}$
- If probability (Significance Level) > p-value: Fail to Reject Null
- If probability (Significance Level) < p-value: Reject Null in favour of Alternative

- **Example:**
- $\hat{t} = 2$; show: $P(t \geq 2) = 0.022$ (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: H_0: $R^2 = 0$
- 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: H_0: Gamma = $\text{Gamma}_2 = 0$

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*
    - $y = \beta x + u$; 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
- $x_{1..k}$: Endogenous Variables
- $w_{1..k}$: Exogenous Variables
- $z_{1..k}$: Instrumental Variables

- *"For every x you expect to be Endogenous, you require an Instrumental Variable"*

1) Regress x on $z_{1..k}$ values for $x_{1..k}$ and obtain **predicted** values: $\hat{x}$
2) Regress y on $w_{1..k}$ (don't need Instrumental Variables)
3) Look for high correlation between Instrumental Variables and Explanatory Variables

6.5: Testing for Endogeneity

- "Hausman Test"
- H_0: 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?**
  - R^2 shows this integrity
  - f-test shows validity of the set of instruments as a whole (like OLS)

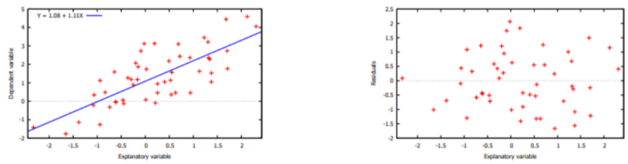
7: Robust Estimation

- 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:
  - $var(u) = \sigma^2$
  - All regression errors have equal variance

7.1: Heteroscedasticity & Homoscedasticity

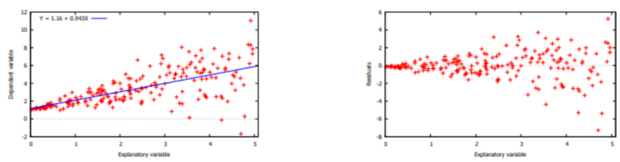
- <u>Heteroscedastic</u>: Non-Constant Error Variance
- <u>Homoscedastic</u>: Constant Error Variance
  - $var(u) = \sigma^2$

- **Example:**



  - Hence, <u>Homoscedastic</u> (no pattern)

- **Example:**



  - Hence, <u>Heteroscedastic</u> (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

- $var(u) = \sigma^2$
- House price dataset:
    - Dependent variable: house price
    - Explanatory variables: bedrooms, bathrooms etc.
- u 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

- $var(u) = \sigma^2 \omega_i^2$
    - For: $i = \{1, \dots N\}$; i 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**)
    - H_0: Homoscedastic Constant Error Variance
    - H_A: 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**