ECMT 461 Lecture Slides for Exam 3 Material Fall 2023

Craig Schulman

The following slide decks cover material that will be included on Exam 3.

I reserve the right to amend/expand this material as needed.

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Craig T. Schulman

Lecture 13

1

AGENDA

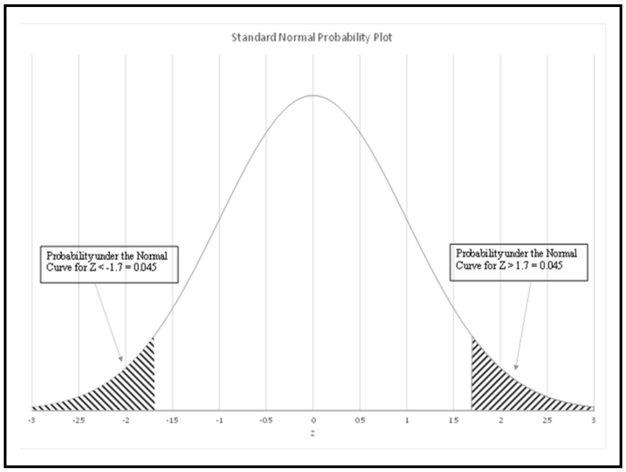
- P-Values
- Term Project Examples
  - Descriptive Statistics
  - Frequency Distributions
  - Single Sample Confidence Intervals
  - Single Sample Hypothesis Tests

2

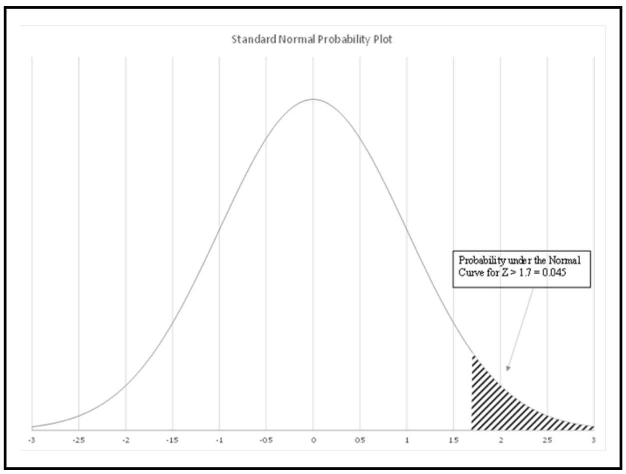
P-VALUES

- For a given ***calculated test statistic,*** the P-Value is the probability in the "tails" of the appropriate probability distribution
- Hypothesis Testing Decision Rules
  - Reject H_0 if the P-Value < the significance level α
  - Fail to Reject H_0 if the P-Value > the significance level α
- For Two-Tailed alternative hypotheses
  - P-Value is TWO TIMES the probability in the Upper or Lower tail
- For One-Tailed alternative hypotheses
  - P-Value is the probability in the tail

3



4



5

TERM PROJECT EXAMPLES

6

EXCEL "ARRAY" FUNCTIONS

- Allow you to include conditional IF statements in a wide array of standard functions to make your coding more efficient

Group	Sample Size	Minimum	Mean	Median	Max
Overall	3195	12.176	27.275	27.332	45.8
A	710	14.703	28.157	28.069	45.8
B	732	14.427	=AVERAGE(IF($A:$A=$F7,$B:$B))		

- Type your formula then CTRL-SHIFT-ENTER

- The "braces" around the formula in the formula bar show that it is an array formula -- THESE CANNOT BE INPUT MANUALLY

7

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Lecture 14

Craig T. Schulman

1

ANALYSIS OF VARIANCE: ANOVA

- A method for testing differences in means across multiple dimensions in a given variable.
- Three versions we will cover:
  - One-Way ANOVA
  - Two-Way ANOVA without Replication
  - Two-Way ANOVA with Replication

2

ONE-WAY ANOVA

- With a Random Variable X, for which we can identify 2 or more sub-groups.
  - For your Term Project, this is your Criterion variable and your identified sub-groups.
- The question is whether the means (averages) are the same across all the sub-groups.

3

ONE-WAY ANOVA

- The Null Hypothesis is that the sub-groups all have the same mean:

$$H_0:\ \mu_1 = \mu_2 = \cdots = \mu_k$$

- Versus the Alternative Hypothesis that at least one of the sub-group means is different.

$$H_A:\ \mu_i \neq \mu_j\ for\ at\ least\ one\ pair, \mu_i, \mu_j$$

4

ONE-WAY ANOVA

- Our Test Statistics will be based on measures of how much error is observed under the Null Hypothesis versus the Alternative Hypothesis – measures of error based on **Sums of Squares:**

$$SS_i = \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2$$

5

ONE-WAY ANOVA

- Under the Alternative Hypothesis when we let each sub-group have a different mean, we have the *within group sum of squares*
  - *SSW with Degrees of Freedom (DF):* $n - K$
    - Where n is the total number of observations across all sub-groups and K is the number of sub-groups

$$SSW = \sum_{i=1}^{K} SS_i = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2$$

6

ONE-WAY ANOVA

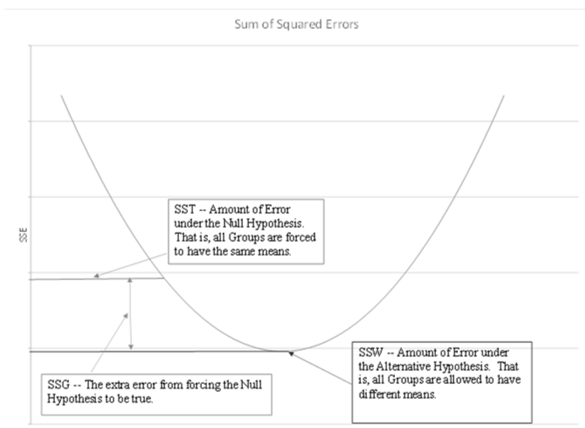
- Under the Null Hypothesis, when we force all the sub-groups to have the same means, we have the *total sum of squares, SST.*

$$SST = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{\bar{x}}\right)^2$$

- The "extra" error associated with forcing the null hypothesis to be true is the *between group sum of squares, SSG*
  - *SSG with DF: $K-1$*

$$SSG = \sum_{i=1}^{K} n_i(\bar{x}_i - \bar{\bar{x}})^2$$

7



8

ONE-WAY ANOVA

- Both SSW and SSG are distributed as Chi-Square. Our test of the Null that all the sub-groups have the same mean will then use an F statistic:

$$F_{calc} = \frac{SSG/(K-1)}{SSW/(n-K)} \sim F_{(K-1),(n-K)}$$

- Excel Examples

9

TWO-WAY ANOVA

- With Two-Way ANOVA, the mean of our variable of interest can vary across two different dimensions – Groups and Blocks.
- Two different types of Two-Way ANOVA
  - Without "Replication"
  - With Replication

10

TWO-WAY ANOVA WITHOUT REPLICATION

- With K different Groups (Columns in Excel), H different Blocks (Rows in Excel), and one observation for each Group-Block combination, for a total of $N = HK$ observations.
- This gives us two potential Null Hypotheses
  - Means are the same across Groups
  - Means are the same across Blocks

11

TWO-WAY ANOVA WITHOUT REPLICATION

- Under the Alternative Hypothesis, where we allow the means to vary over both Groups and Blocks, we have the *error sum of squares*, SSE
  - SSE with DF: $(K - 1)(H - 1)$
- The "extra" error created when we force the Group means to be the same gives us the *between Groups sum of squares, SSG*
  - SSG with DF: $K - 1$
- The "extra" error created when we force the Block means to be the same gives us the *between Blocks sum of squares, SSB*
  - SSB with DF: $\mathrm{H} - 1$

12

TWO-WAY ANOVA WITHOUT REPLICATION

- Sums of Squares Equations

$$between\ Groups:\ SSG = H\sum_{i=1}^{K}(\bar{x}_{i.} - \bar{\bar{x}})^2 \quad with\ DF_G = K - 1$$

$$between\ Blocks:\ SSB = K\sum_{j=1}^{H}\left(\bar{x}_{.j} - \bar{\bar{x}}\right)^2 \quad with\ DF_B = H - 1$$

$$error:\ SSE = \sum_{i=1}^{K}\sum_{j=1}^{H}\left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} - \bar{\bar{x}}\right)^2 \quad with\ DF_E = (K-1)(H-1)$$

13

TWO-WAY ANOVA WITHOUT REPLICATION

- To test the Null Hypothesis that the means are the same across Groups we use:

$$F_{calc} = \frac{SSG/(K-1)}{SSE/(K-1)(H-1)}\ vs.\ F_{\alpha,(K-1),(K-1)(H-1)}$$

- To test the Null Hypothesis that the means are the same across Blocks we use:

$$F_{calc} = \frac{SSB/(H-1)}{SSE/(K-1)(H-1)}\ vs.\ F_{\alpha,(H-1),(K-1)(H-1)}$$

Excel Example

14

TWO-WAY ANOVA WITH REPLICATION

- In Two-Way ANOVA with Replication, we again have K different Groups (Columns in Excel), H different Blocks (Sample in Excel), and multiple (m) observations for each Group-Block combination.
- This allows us to test three different null hypotheses:
  - Means the same across Groups
  - Means the same across Blocks
  - No "Interaction" effects between Groups and Blocks

15

TWO-WAY ANOVA WITH REPLICATION

- For the Null of equal Group means
  $$between\ Groups:\ SSG\ with\ DF_G = K - 1$$
- For the Null of equal Block means
  $$between\ Blocks:\ SSB\ with\ DF_B = H - 1$$
- For the Null of no Interaction effects
  $$Interaction:\ SSI\ with\ DF_I = (K-1)(H-1)$$
- And the SS under the Alternative
  $$error:\ SSE\ with\ DF_E = HK(m-1)$$

16

TWO-WAY ANOVA WITH REPLICATION

- To test the Null of equal Group means
  $$F_{calc} = \frac{SSG/(K-1)}{SSE/HK(m-1)}\ vs. F_{\alpha,(K-1),HK(m-1)}$$
- To test the Null of equal Block means
  $$F_{calc} = \frac{SSB/(H-1)}{SSE/HK(m-1)}\ vs. F_{\alpha,(H-1),HK(m-1)}$$
- And for the Null of no Interaction Effects
  $$F_{calc} = \frac{SSI/(K-1)(H-1)}{SSE/HK(m-1)}\ vs. F_{\alpha,(K-1)(H-1),HK(m-1)}$$

Excel Examples

17

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Lecture 15

Craig T. Schulman

1

LINEAR REGRESSION

- Sample Correlation r_{xy} provides a measure of the degree of association between two variables X and Y
- Linear Regression is a method of "explaining" the observed variation in some variable Y that is believed to be related in some way to the variable X
- Y is denoted the "Dependent" variable
- X is denoted the "Explanatory" variable
- Express Y as a function of X: Y = f(X)

2

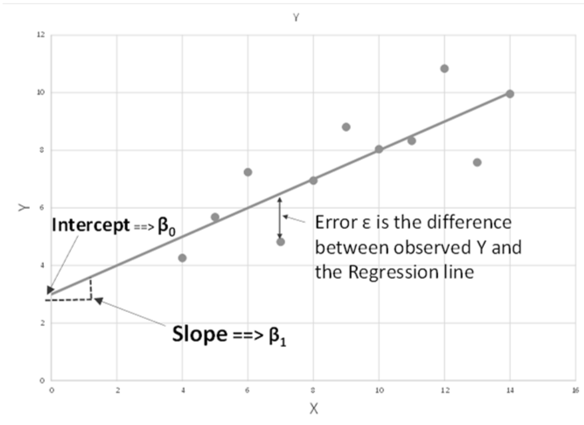
BIVARIATE REGRESSION MODEL

- "Bivariate" because there are two variables X and Y
- Assume the relationship between X and Y is linear in the population parameters β_0 *and* β_1 *and the random error term* ε

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Where β_0 *is the intercept and* β_1 *is the slope* of the regression line
- Error ε is the difference between the observed Y and the regression line

3



4

MODEL ASSUMPTIONS

- The values of X are either fixed or a random variable that is independent of the error term ε
- The random error term ε is distributed with a mean of zero and a constant variance $\varepsilon \sim (0, \sigma^2)$
- The sample variance of X is not zero

5

LEAST SQUARES RULE

- The Least Squares Rule provides a set of formulas – ***estimators*** – to get sample estimates for the population parameters
- This is like using the sample mean $\overline{X}$ as the estimator for the population mean μ
- The task is to find an estimator b_0 for the population intercept β_0, and b_1 for the population slope β_1 that minimize the amount of squared error between the observed data and the regression line – hence, Least Squares

6

LEAST SQUARES RULE

- SSE is the sum of squared errors
- We want to find b_0 *and* b_1 to minimize SSE
- Use calculus to take derivatives of SSE with respect to b_0 *and* b_1 set these equal to zero and solve

$$SSE = \sum e_i^2$$
$$= \sum (Y_i - (b_0 + b_1 X_i))^2$$
$$= \sum (Y_i - \hat{Y}_i)^2$$

7

LEAST SQUARES ESTIMATORS

$$b_0 = \bar{Y} - b_1 \bar{X}$$
$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$e_i = Y_i - b_0 - b_1 X_i = Y_i - \hat{Y}_i$$

$$S^2 = \frac{SSE}{n-2}$$

8

ESTIMATOR PROPERTIES

- Under the model assumptions, the Least Squares estimators b_0 *and* b_1 are ***Unbiased***

$$E(b_0) = \beta_0$$

$$E(b_1) = \beta_1$$

- Among all linear unbiased estimators, the Least Squares estimators have the smallest variance – they are the Best Linear Unbiased Estimator or BLUE

9

MEASURE OF "FIT"

- Regression is used to "explain" the observed variation in Y

$$\sum (Y_i - \bar{Y})^2 = \sum \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum e_i^2$$

$$SST = SSR + SSE$$

- SST: The Total Sum of Squares is the total variation of Y about its mean
- SSR: The Regression Sum of Squares is the variation in Y explained by the regression
- SSE: The Error Sum of Squares is the variance left unexplained by the regression

10

MEASURE OF "FIT"

- A measure of the degree of variation of Y explained by the regression is given by the Coefficient of Determination or Regression R^2

$$R^2 = 1 - \frac{SSE}{SST}$$

- If the regression equation includes the intercept term b_0, then R^2 is bound on $[0,1]$ and for the bivariate model

$$R^2 = r_{XY}^2$$

11

ANOTHER ASSUMPTION

- If we add the assumption that the error term ε is distributed as a Normal random variable: $\varepsilon \sim N(0, \sigma^2)$ then the Least Squares estimators b_0 and b_1 are also have a Normal distribution

$$b_0 \sim N\left(\beta_0, \left\{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)\sum (X_i - \bar{X})^2}\right\} \sigma^2\right)$$

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum (X_i - \bar{X})^2}\right)$$

- For sample estimators, replace σ^2 *with* S^2 and denote the variances of b_0 and b_1 as $S_{b_0}^2$ and $S_{b_1}^2$

12

HYPOTHESIS TESTS

- Given the Normal Distribution assumption we can test hypotheses about the Least Squares parameters as follows:

$$H_0:\ \beta_1 = k$$

$$t_{calc} = \frac{b_1 - k}{S_{b_1}}$$

$$versus\ t_{n-2,\frac{\alpha}{2}}\ or\ t_{n-2,\alpha}$$

- Excel Example

13

FUNCTIONAL FORM

- The assumption that the model is linear in the parameters β_0 *and* β_1 ***DOES NOT*** mean the model involves a linear relation between the two variables of interest
- Consider the Demand equation

$$Q = AP^{\beta_1}v$$

- Taking logarithms, we can write

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Where Y is the log of Q, X is the log of P, and ε is the log of v

14

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Lecture 16

Craig T. Schulman

1

LINEAR REGRESSION

- Regression R^2 – Measure of Fit
- Regression "Outliers"
- Prediction
- Correlation Analysis

2

MEASURE OF "FIT"

- Regression is used to "explain" the observed variation in Y

$$\sum(Y_i - \bar{Y})^2 = \sum\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum e_i^2$$

$$SST = SSR + SSE$$

- SST: The Total Sum of Squares is the total variation of Y about its mean
- SSR: The Regression Sum of Squares is the variation in Y explained by the regression
- SSE: The Error Sum of Squares is the variance left unexplained by the regression

3

MEASURE OF "FIT"

- A measure of the degree of variation of Y explained by the regression is given by the Coefficient of Determination or Regression R^2

$$R^2 = 1 - \frac{SSE}{SST}$$

- If the regression equation includes the intercept term b_0, then R^2 is bound on [0,1] and for the bivariate model

$$R^2 = r_{XY}^2$$

4

MEASURE OF "FIT"

- If the regression equation includes the intercept term b_0, then R^2 is bound on [0,1]
- The R^2 (when an intercept is included in the regression equation) measures the proportion of variation in the dependent variable that is explained by the regression
- An R^2 of 0.45 says the regression explains 45% of the observed variation in the dependent variable
- Not unusual with cross-sectional data to have R^2 values that seem quite 'low' – e.g. 5% or 10%

5

REGRESSION OUTLIERS

- An "Outlier" is a data point that deviates substantially from the predicted value (the regression line)
- Can identify the points by computing "standardized" residuals
- Excel "Standard Residuals" are not technically correct but are a reasonably close approximation
- A standardized residual greater than 2 in absolute value is a potential outlier
- Check for data errors or unusual circumstances

6

PREDICTION

- Can use a regression equation to predict or forecast the dependent variable for an assumed value of the independent variable, X_0

$$\hat{Y} = b_0 + b_1 X_0$$

- Of interest is a prediction interval

$$\hat{Y} \pm t_{n-2,\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{\Sigma\left(X - \overline{X}\right)^2}}$$

7

CORRELATION ANALYSIS

- For the Bivariate Regression model (that includes an intercept)

$$b_1 = \frac{S_Y}{S_X} r_{XY}$$

- So the test of the hypothesis $H_0\colon\ \beta_1 = 0$ is a direct test of the hypothesis $H_0\colon\ \rho_{XY} = 0$.

8

CORRELATION ANALYSIS

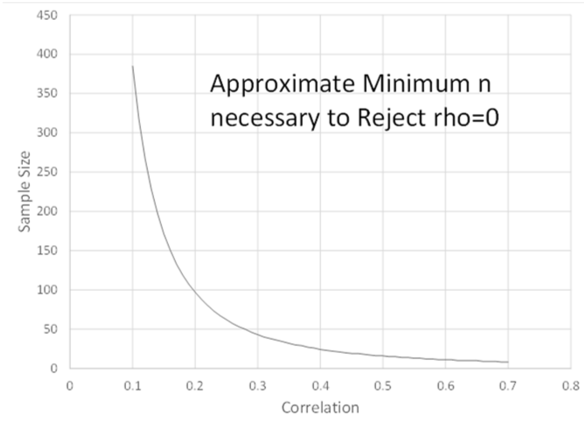
- We can also construct a test statistic based on the sample correlation coefficient r_{XY} to test the hypothesis $H_0\colon\ \rho_{XY} = 0$

$$t_{calc} = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1 - r_{XY}^2}}$$

$$versus\ t_{n-2,\alpha}\ or\ t_{n-2,\alpha/2}$$

- This is the form of the correlation test that I suggest you use for your term projects

9



10

CORRELATION ANALYSIS

- For the null hypothesis $H_0\colon\ \rho_{XY} = 0$ the t-distribution provides a robust test
- To test whether the correlation is any number other than zero, the sampling distribution of t_{calc} becomes highly skewed

$$t_{calc} = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}}$$

$$versus\ t_{n-2,\alpha}\ or\ t_{n-2,\alpha/2}$$

11

CORRELATION ANALYSIS

- For the null hypothesis $H_0\colon\ \rho_{XY} = k,$ where k is any number on the interval (-1,1) we use **Fisher's z_r transformation**

$$z_r = \frac{1}{2}\ln\left[\frac{1+r_{XY}}{1-r_{XY}}\right]$$

- Where ln is the natural logarithm and z_r has a Normal distribution with

$$Var(z_r) = \frac{1}{n-3}$$

12

CORRELATION ANALYSIS

- A confidence interval for z_r is

$$z_r \pm Z_{\alpha/2}\sqrt{\frac{1}{n-3}}$$

- To test the null hypothesis $H_0\colon\ \rho_{XY} = k$ use

$$Z_{calc} = \frac{z_r - z_k}{\sqrt{\frac{1}{n-3}}}$$

13

CORRELATION ANALYSIS

- Suppose n=45 and $r_{XY} = 0.68$. To test the null hypothesis $H_0\colon\ \rho_{XY} = 0.5$

$$z_r = \frac{1}{2}\ln\left[\frac{1+0.68}{1-0.68}\right] = 0.8921$$

$$z_k = \frac{1}{2}\ln\left[\frac{1+0.5}{1-0.5}\right] = 0.5493$$

$$Z_{calc} = \frac{0.8921 - 0.5493}{\sqrt{\frac{1}{45-3}}} = 1.813$$

$$for\ \alpha = 5\%\ Z_{\alpha/2} = 1.96 \quad for\ \alpha = 10\%\ Z_{\alpha/2} = 1.645$$

14

CORRELATION ANALYSIS

- Two methods to test for differences in correlations between two sub-groups from a sample
- Using Linear Regression (a bit ‘messy’)
- Using Fisher’s z_r transformation (much more straightforward)

15

REGRESSION WITH 2 GROUPS

- From a sample of n observations on Y and X, suppose you can identify two sub-groups in the sample – Group 1 and Group 2
- We can identify the two different groups with a single Yes/No identification variable – called a ***Dummy Variable***
- If an observation is part of Group 1, define the variable D=1 and if part of Group 2, D=0, then specify the regression equation

$$Y = \beta_0 + \beta_1 X + \gamma_0 D + \gamma_1 (D \times X) + \varepsilon$$

16

REGRESSION WITH 2 GROUPS

- For Group 1 when D=1, the regression intercept and slope can be written $$Y = (\beta_0 + \gamma_0) + (\beta_1 + \gamma_1)X + \varepsilon$$
- So the intercept for Group 1 is $\beta_0 + \gamma_0$ and the slope is $\beta_1 + \gamma_1$
- For Group 2 when D=0 $$Y = \beta_0 + \beta_1 X + \varepsilon$$
- So the intercept and slope for Group 2 are just β_0 *and* β_1
- The parameter γ_0 measures the difference in the intercept between Group 1 and Group 2 (the difference in the mean of Y after controlling for the effect of X)
- The parameter γ_1 measures the difference in the slope of the regression equation between the two groups (difference in correlation)

17

REGRESSION WITH 2 GROUPS

- Example using Salary Data
- Salary vs. Experience for Females (Group 1) and Males (Group 2)

18

2 GROUP CORRELATION TEST

- An alternative (and more straightforward) test of differences in correlations across two subgroups uses Fisher's z_r transformation
- Suppose you have the sample correlation coefficient for the first group r_1 based on n_1 observations, and similarly for the second group, r_2 based on n_2 observations. First calculate the z_r transformations

$$z_{r_1} = \frac{1}{2}\ln\left[\frac{1+r_1}{1-r_1}\right] \quad z_{r_2} = \frac{1}{2}\ln\left[\frac{1+r_2}{1-r_2}\right]$$

19

2 GROUP CORRELATION TEST

- To test the null hypothesis $H_0\colon \rho_1 = \rho_2$ construct the calculated Z-statistic

$$Z_{calc} = \frac{z_{r_1} - z_{r_2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

- And compare to Z_α for a one-tailed test or $Z_{\alpha/2}$ for a two-tailed test
- Example with Salary data

20

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Lecture 17

Craig T. Schulman

1

MULTI-GROUP TESTS

- A test of equal correlations across 2 or more groups
- Multi-Group Tests using Regression

2

2 GROUP CORRELATION TEST

- Test of differences in correlations across two subgroups using Fisher's z_r transformation

$$z_{r_1} = \frac{1}{2}\ln\left[\frac{1+r_1}{1-r_1}\right] \quad z_{r_2} = \frac{1}{2}\ln\left[\frac{1+r_2}{1-r_2}\right]$$

$$Z_{calc} = \frac{z_{r_1} - z_{r_2}}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

- And compare to Z_α for a one-tailed test or $Z_{\alpha/2}$ for a two-tailed test

3

MULTI-GROUP CORRELATION TEST

- To test a hypothesis of equal correlations across multiple groups: $H_0\colon \rho_1 = \rho_2 = \cdots = \rho_k$ we can use Fisher's z_r transformation to construct a Chi-Square test statistic

$$\chi^2_{calc} = \sum_{j=1}^{k}\left[(n_j - 3)z_{r_j}^2\right] - \frac{\left[\sum_{j=1}^{k}(n_j - 3)z_{r_j}\right]^2}{\sum_{j=1}^{k}(n_j - 3)}$$

- Compared to the Chi-Square distribution with $k - 1$ degrees of freedom

4

MULTI-GROUP CORRELATION TEST

- The test statistic looks intimidating, but it is straightforward to calculate in Excel
- Example

5

MULTI-GROUP TESTS WITH REGRESSION

- When we considered an X-Y regression with two groups, we found that we can identify the two different groups with a single Yes/No identification variable – called a ***Dummy Variable***
- With ***k*** sub-groups in the data, we need $k - 1$ Dummy Variables to uniquely identify the ***k*** sub-groups
- We pick one of the groups to be the *reference group*

6

MULTI-GROUP TESTS WITH REGRESSION

- Suppose we have 4 sub-groups in the data
- Let the fourth sub-group be the reference group
- Define 3 Dummy Variables $D_1, D_2, and\ D_3$ for sub-groups 1, 2, and 3, respectively

Group	D_1	D_2	D_3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

7

MULTI-GROUP TESTS WITH REGRESSION

- Specify the ***Unrestricted*** regression equation

$$Y = \beta_0 + \beta_1 X + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3$$
$$+\gamma_1(D_1 \times X) + \gamma_2(D_2 \times X) + \gamma_3(D_3 \times X) + \varepsilon$$

- And the ***Restricted*** regression equation that forces all the sub-group correlations to be the same but allows the means of the sub-groups to be different

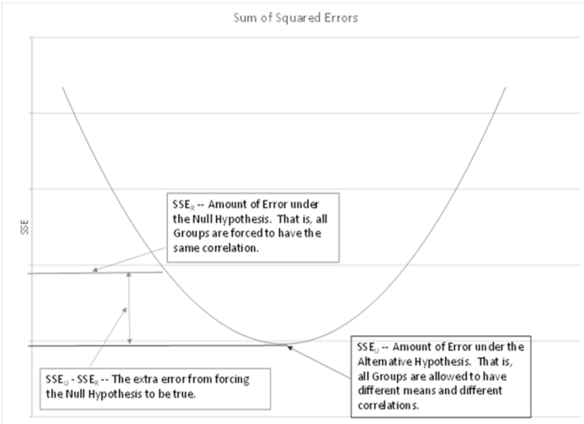
$$Y = \beta_0 + \beta_1 X + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \varepsilon$$

8

MULTI-GROUP TESTS WITH REGRESSION

- In this example, we used sub-group 4 as the reference group
- The "alpha" parameters from the Unrestricted equation measure the difference in the conditional mean of Y for each of the three groups compared to sub-group 4
- The "gamma" parameters from the Unrestricted equation measure the difference in the effect of X on Y for each of the three groups compared to sub-group 4
- Changing the reference group would change the interpretation of the parameters but does not affect the overall results

9



10

MULTI-GROUP TESTS WITH REGRESSION

- The SSE from the Unrestricted equation, $\boldsymbol{SSE_U}$, and the SSE from the Restricted equation, $\boldsymbol{SSE_R}$, allow us to construct a calculated F-statistic

$$F_{calc} = \frac{(SSE_R - SSE_U)/(DF_R - DF_U)}{SSE_U/DR_U} \sim F_{DF_R - DR_U, DF_U}$$

- Where $DF_R \text{ and } DF_U$ are the degrees of freedom in the Restricted and Unrestricted equations, respectively
- Example

11

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Lecture 18

Craig T. Schulman

1

TERM PROJECT SECTION 5

- Two Sample Confidence Intervals and Hypothesis Tests
- You should first conduct pair-wise hypothesis tests of equal variances among sub-groups
- Given the results of your variance tests, next use the Excel Data Analysis Tool to conduct pair-wise means tests with either equal or unequal means as determined by your variance tests
- Use the Analysis Tool results to create tables of your confidence intervals and pair-wise tests
- Example

2

NON-PARAMETRIC TESTS

- Non-Parametric tests are typically used to assess categorical or nominal (discrete) data
- Goodness of Fit Test
- Goodness of Fit for Poisson Distribution

3

CHI-SQUARE GOODNESS OF FIT TEST

- Suppose you have a categorical variable that takes on a fixed number of distinct values, or categories: $C_1, C_2, \dots, C_k$
- Denote the Observed number of outcomes that fall into each category as $O_1, O_2, \dots, O_k$
- The question is whether these observed outcomes are consistent with – ***fit*** – a particular distribution or set of Expected outcomes $E_1, E_2, \dots, E_k$

4

CHI-SQUARE GOODNESS OF FIT TEST

- The assumptions underlying the Chi-Square goodness-of-fit test are as follows:
- Categorical/nominal data representing mutually exclusive categories are used in the analysis
- The data represent a sample of n independent observations
- The expected frequency of each category (or cell) is 5 or greater

5

Category	Observed Frequency Count	Expected Frequency Count
C_1	O_1	E_1
C_2	O_2	E_2
…	…	…
C_K	O_K	E_K
Total Obs.	n	

6

TEST STATISTIC

- Under these assumptions we can formulate a Chi-Square test statistic as follows

$$\chi^2_{calc} = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(K-1)}$$

- Against a one-tailed upper alternative
- Examples

7

NCAA STARTING PITCHERS

- Among all NCAA D-1 pitchers that started at least 5 games, the distribution by class rank is:

Class Rank	Proportion of Starters
Fr	20%
So	22%
Jr	33%
Sr	25%

- Is the distribution of starting pitchers by class rank in the SEC consistent with the overall distribution?

8

OTHER EXAMPLES

- Students doing homework by day of the week – is it evenly distributed?
- Rolls of a six-sided die – is the die ‘fair’?
- Number of television sets per family

9

TEST FOR THE POISSON DISTRIBUTION

- The same basic test procedure can be applied to assess whether a set of observed outcomes are consistent with a particular probability distribution with unknown parameters
- However, we must adjust the degrees of freedom of the Chi-Square statistic for the number of unknown parameters
- For the Poisson distribution, there is one parameter – the mean number of occurrences per observation unit, λ

10

POISSON DISTRIBUTION

- The Poisson probability distribution can be used to model the number of occurrences (or successes) of a certain event in a given continuous interval such as time, spatial area, or length.
  - The number of trucks arriving at a warehouse in a given week.
  - The number of failures in a computer system in a given day.
  - The number of defects in a large roll of sheet metal.
  - The number of customers to arrive at a coffee bar in a given time interval.

11

POISSON ASSUMPTIONS

- Assume that an interval is divided into a large number of equal subintervals so that the probability of the occurrence of an event in any subinterval is small.
- The probability of the occurrence of an event is constant for all subintervals.
- There can be no more than one occurrence in each subinterval.
- Occurrences are independent – an occurrence in one subinterval does not have an effect on the probability of an occurrence in another subinterval.

12

POISSON PDF

- The PDF for the Poisson distribution is shown at the right, where
- P(x) = the probability of x successes over a given time or space unit, given λ.
- λ = the mean (or expected) number of successes per time or space unit, λ>0.
- e = 2.71828… (the base for natural logarithms).

$$\boxed{P(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}\ for\ x = 0, 1, 2, \ldots}$$

13

POISSON CDF

- Recall that the CDF is defined as the probability of X being less than or equal to some specific value, X_0
- We get the CDF by summing up the PDF for all values of $X \leq X_0$

$$P[X \leq X_0] = F(X) = \sum_{X_i=0}^{X_0} P(X_i)$$

- Example

14

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Lecture 19

Craig T. Schulman

1

NON-PARAMETRIC TESTS

- Chi-Square Non-Parametric tests of:
- Homogeneity – Are independent samples of a random variable X homogeneous with respect to some category variable?
- Independence – For a random variable that can be categorized in two dimensions (two category variables), are the two dimensions independent of one another?

2

HOMOGENEITY TEST

- For a random variable X, suppose you have **k** independent samples denoted $X_1, X_2, \dots, X_k$
- Each of these samples can be categorized into one of **m** mutually exclusive categories $C_1, C_2, \dots, C_m$
- Is the distribution of observations among the **m** categories homogeneous (the same) across the **k** samples?
- This can be presented in the form of a contingency table

3

CONTINGENCY TABLE

		Sample				
		X_1	X_2	…	X_k	Row Totals
Categories	C_1	O_{11}	O_{12}	…	O_{1k}	n_{C_1}
	C_2	O_{21}	O_{22}	…	O_{2k}	n_{C_2}
	…	…	…	…	…	…
	C_m	O_{m1}	O_{m2}	…	O_{mk}	n_{C_m}
Column Totals		n_{X_1}	n_{X_2}	…	n_{X_k}	n

- Where
  - O_{ij} is the ***frequency count*** for category i in sample j
  - n_{X_j} is the sample size for sample X_j
  - n_{C_i} is the total frequency count for category i
  - n is the overall sample size

4

HOMOGENEITY TEST

- To construct our test statistic, we need to compare each observed cell frequency count, O_{ij} to the ***expected*** frequency count, E_{ij} under the null hypothesis of homogeneity – that the categories are proportionally similar across all the samples
- As with the Goodness of Fit test, we require each expected cell frequency to be greater than 5

5

HOMOGENEITY TEST

- The expected cell frequency under the null hypothesis is calculated by

$$E_{ij} = \frac{n_{C_i} n_{X_j}}{n}$$

- And the Chi-Square test statistic is then

$$\chi^2_{calc} = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(m-1)(k-1)}$$

- Example

6

INDEPENDENCE TEST

- Consider a random variable that can be categorized in two dimensions – or across two categorical ***factors***
- The Chi-Square test of independence can be used to whether the two factors are independent of one another
- Mathematically, the test is identical to the test of Homogeneity

7

INDEPENDENCE TEST

		Factor A				
		A_1	A_2	…	A_k	Row Totals
Factor B	B_1	O_{11}	O_{12}	…	O_{1k}	n_{B1}
	B_2	O_{21}	O_{22}	…	O_{2k}	n_{B2}
	…	…	…	…	…	…
	B_m	O_{m1}	O_{m2}	…	O_{mk}	n_{Bm}
Column Totals		n_{A1}	n_{A2}	…	n_{Ak}	n

- Compare each observed cell frequency count Oij to its expected frequency under the null E_{ij}
- Example

8

TIME SERIES DATA

- Consider the random variable Y that is observed over some time interval t
- The time interval could be days, weeks, months, years, etc.
- For simplicity, index $t = \{1, 2, \dots, T\}$
- Unlike cross-sectional data, there is a natural ***order*** to time-series data, Y_{t+1} naturally follows Y_t
- This natural ordering can present special estimation issues as well as interpretations related to periodic change and growth

9

ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Lecture 20

Craig T. Schulman

1

LINEAR REGRESSION

- Sample Correlation r_{xy} provides a measure of the degree of association between two variables X and Y
- Linear Regression is a method of "explaining" the observed variation in some variable Y that is believed to be related in some way to the variable X
- Y is denoted the "Dependent" variable
- X is denoted the "Explanatory" variable
- Express Y as a function of X: Y = f(X)

2

BIVARIATE REGRESSION MODEL

- "Bivariate" because there are two variables X and Y
- Assume the relationship between X and Y is linear in the population parameters β_0 *and* β_1 *and the random error term* ε

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Where β_0 *is the intercept and* β_1 *is the slope* of the regression line
- Error ε is the difference between the observed Y and the regression line

3

BIVARIATE REGRESSION MODEL

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- The regression intercept β_0 measures the *conditional* mean of the dependent variable Y: The mean of Y after accounting for the mean effect of X
- The regression slope β_1 measures the *marginal effect* of X on Y: If X changes by 1 unit, the regression predicts that Y will change by β_1 units
- HW 8 Example

4

HYPOTHESIS TESTS

- Given the Normal Distribution assumption we can test hypotheses about the Least Squares parameters as follows:

$$H_0:\ \beta_1 = k$$

$$t_{calc} = \frac{b_1 - k}{S_{b_1}}$$

$$versus\ t_{n-2,\frac{\alpha}{2}}\ or\ t_{n-2,\alpha}$$

5

CORRELATION ANALYSIS

- For the Bivariate Regression model (that includes an intercept)

$$b_1 = \frac{S_Y}{S_X} r_{XY}$$

- So the test of the hypothesis $H_0:\ \beta_1 = 0$ is a direct test of the hypothesis $H_0:\ \rho_{XY} = 0$.

6

ONE-WAY ANOVA

- The Null Hypothesis is that the sub-groups all have the same mean:

$$H_0:\ \mu_1 = \mu_2 = \cdots = \mu_k$$

- Versus the Alternative Hypothesis that at least one of the sub-group means is different.

$$H_A:\ \mu_i \neq \mu_j \ for\ at\ least\ one\ pair, \mu_i, \mu_j$$

7

ONE-WAY ANOVA

- Under the Alternative Hypothesis when we let each sub-group have a different mean, we have the *within group sum of squares*
  - *SSW with Degrees of Freedom (DF):* $n - K$
    - Where n is the total number of observations across all sub-groups and K is the number of sub-groups

$$SSW = \sum_{i=1}^{K} SS_i = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2$$

8

ONE-WAY ANOVA

- Under the Null Hypothesis, when we force all the sub-groups to have the same means, we have the *total sum of squares, SST.*

$$SST = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{\bar{x}}\right)^2$$

- The "extra" error associated with forcing the null hypothesis to be true is the *between group sum of squares, SSG*
  - *SSG with DF:* $K - 1$

$$SSG = \sum_{i=1}^{K} n_i (\bar{x}_i - \bar{\bar{x}})^2$$

9

ONE-WAY ANOVA

- Both SSW and SSG are distributed as Chi-Square. Our test of the Null that all the sub-groups have the same mean will then use an F statistic:

$$F_{calc} = \frac{SSG/(K-1)}{SSW/(n-K)} \sim F_{(K-1),(n-K)}$$

- Excel Examples

10

TWO-WAY ANOVA WITHOUT REPLICATION

- Under the Alternative Hypothesis, where we allow the means to vary over both Groups and Blocks, we have the *error sum of squares*, SSE
  - SSE with DF: $(K-1)(H-1)$
- The "extra" error created when we force the Group means to be the same gives us the *between Groups sum of squares, SSG*
  - SSG with DF: $K-1$
- The "extra" error created when we force the Block means to be the same gives us the *between Blocks sum of squares, SSB*
  - SSB with DF: $\mathrm{H}-1$

11

TWO-WAY ANOVA WITHOUT REPLICATION

- To test the Null Hypothesis that the means are the same across Groups we use:

$$F_{calc} = \frac{SSG/(K-1)}{SSE/(K-1)(H-1)} \; vs. F_{\alpha,(K-1),(K-1)(H-1)}$$

- To test the Null Hypothesis that the means are the same across Blocks we use:

$$F_{calc} = \frac{SSB/(H-1)}{SSE/(K-1)(H-1)} \; vs. F_{\alpha,(H-1),(K-1)(H-1)}$$

12

CORRELATION ANALYSIS

- We can also construct a test statistic based on the sample correlation coefficient r_{XY} to test the hypothesis $H_0\colon\ \rho_{XY} = 0$

$$t_{calc} = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}}$$

$$versus\ t_{n-2,\alpha}\ or\ t_{n-2,\alpha/2}$$

13

2 GROUP CORRELATION TEST

- An alternative (and more straightforward) test of differences in correlations across two subgroups uses Fisher's z_r transformation
- Suppose you have the sample correlation coefficient for the first group r_1 based on n_1 observations, and similarly for the second group, r_2 based on n_2 observations. First calculate the z_r transformations

$$z_{r_1} = \frac{1}{2}\ln\left[\frac{1+r_1}{1-r_1}\right] \quad z_{r_2} = \frac{1}{2}\ln\left[\frac{1+r_2}{1-r_2}\right]$$

14

2 GROUP CORRELATION TEST

- To test the null hypothesis $H_0\colon\ \rho_1 = \rho_2$ construct the calculated Z-statistic

$$Z_{calc} = \frac{z_{r_1} - z_{r_2}}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

- And compare to Z_α for a one-tailed test or $Z_{\alpha/2}$ for a two-tailed test

15

MULTI-GROUP CORRELATION TEST

- To test a hypothesis of equal correlations across multiple groups: $H_0\text{: } \rho_1 = \rho_2 = \cdots = \rho_k$ we can use Fisher's z_r transformation to construct a Chi-Square test statistic

$$\chi^2_{calc} = \sum_{j=1}^{k} \left[(n_j - 3)z_{r_j}^2\right] - \frac{\left[\sum_{j=1}^{k}(n_j - 3)z_{r_j}\right]^2}{\sum_{j=1}^{k}(n_j - 3)}$$

- Compared to the Chi-Square distribution with $k - 1$ degrees of freedom

16

CHI-SQUARE GOODNESS OF FIT TEST

- The assumptions underlying the Chi-Square goodness-of-fit test are as follows:
- Categorical/nominal data representing mutually exclusive categories are used in the analysis
- The data represent a sample of n independent observations
- The expected frequency of each category (or cell) is 5 or greater

17

Category	Observed Frequency Count	Expected Frequency Count
C_1	O_1	E_1
C_2	O_2	E_2
…	…	…
C_K	O_K	E_K
Total Obs.	n	

18

TEST STATISTIC

- Under these assumptions we can formulate a Chi-Square test statistic as follows

$$\chi^2_{calc} = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{(K-1)}$$

- Against a one-tailed upper alternative

19

HOMOGENEITY TEST

- For a random variable X, suppose you have **k** independent samples denoted $X_1, X_2, \ldots, X_k$
- Each of these samples can be categorized into one of **m** mutually exclusive categories $C_1, C_2, \ldots, C_m$
- Is the distribution of observations among the **m** categories homogeneous (the same) across the **k** samples?
- This can be presented in the form of a contingency table

20

CONTINGENCY TABLE

		Sample				
		X_1	X_2	…	X_k	Row Totals
Categories	C_1	O_{11}	O_{12}	…	O_{1k}	n_{C_1}
	C_2	O_{21}	O_{22}	…	O_{2k}	n_{C_2}
	…	…	…	…	…	…
	C_m	O_{m1}	O_{m2}	…	O_{mk}	n_{C_m}
Column Totals		n_{X_1}	n_{X_2}	…	n_{X_k}	n

- Where
  - O_{ij} is the ***frequency count*** for category i in sample j
  - n_{X_j} is the sample size for sample X_j
  - n_{C_i} is the total frequency count for category i
  - n is the overall sample size

21

HOMOGENEITY TEST

- To construct our test statistic, we need to compare each observed cell frequency count, O_{ij} to the ***expected*** frequency count, E_{ij} under the null hypothesis of homogeneity – that the categories are proportionally similar across all the samples
- As with the Goodness of Fit test, we require each expected cell frequency to be greater than 5

22

HOMOGENEITY TEST

- The expected cell frequency under the null hypothesis is calculated by

$$E_{ij} = \frac{n_{C_i} n_{X_j}}{n}$$

- And the Chi-Square test statistic is then

$$\chi^2_{calc} = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(m-1)(k-1)}$$

23

INDEPENDENCE TEST

- Consider a random variable that can be categorized in two dimensions – or across two categorical ***factors***
- The Chi-Square test of independence can be used to whether the two factors are independent of one another
- Mathematically, the test is identical to the test of Homogeneity

24

INDEPENDENCE TEST

		Factor A				
		A_1	A_2	…	A_k	Row Totals
Factor B	B_1	O_{11}	O_{12}	…	O_{1k}	n_{B1}
	B_2	O_{21}	O_{22}	…	O_{2k}	n_{B2}
	…	…	…	…	…	…
	B_m	O_{m1}	O_{m2}	…	O_{mk}	n_{Bm}
Column Totals		n_{A1}	n_{A2}	…	n_{Ak}	n

- Compare each observed cell frequency count Oij to its expected frequency under the null E_{ij}

25