ECMT 461 INTRODUCTION TO ECONOMIC DATA ANALYSIS

Lecture 17

Craig T. Schulman

LINEAR REGRESSION

- Regression R^2 – Measure of Fit
- Regression Interpretation and Hypothesis Testing
- Correlation Analysis
- Multi-Group Correlation Tests

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

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

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

REGRESSION EXAMPLE

- Consider the regression results below, where the dependent variable Y is the Quantity of gasoline sold, measured in gallons, and the explanatory variable X is the Price of gasoline measured in cents per gallon (so Price = 200, is \$2.00)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.792
R Square	0.627
Adjusted R Square	0.619
Standard Error	494.952
Observations	50

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	19733053.11	19733053.11	80.550	0.0000
Residual	48	11758920.92	244977.5191		
Total	49	31491974.02			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	10207.337	420.849	24.254	0.000	9361.164	11053.510
Price	-16.203	1.805	-8.975	0.000	-19.833	-12.573

REGRESSION EXAMPLE

- The "R Square" value of 0.627 says the model explains 62.7% of the variation in Quantity
- The "df" column in the ANOVA table stands for Degrees of Freedom and the Residual df of 48 is the "model" degrees of freedom (number of observations minus 2)
- The "Coefficients" column show the estimated intercept b_0 is 10,207.337, and the estimated slope $b_1 = -16.203$
- The interpretation of the slope parameter b_1 is that, on average, for every 1 cent increase in Price, Quantity sold goes ***down*** by 16.203 gallons

REGRESSION EXAMPLE

- The "t Stat" column shows the calculated t-statistic for the specific hypothesis $H_0: \beta_i = 0$
- For the Price variable, for example

$$t_{calc} = \frac{b_1 - 0}{S_{b_1}} = -\frac{16.203}{1.805} = -8.95$$

- The "P-value" for the Price variable of 0.000 indicates that we would reject this null at any reasonable value of alpha (e.g., 5% or 1%)
- Suppose we wish to test $H_0: \beta_1 = -13$, versus the alternative that it is less than –13, at alpha=5%. Use:

$$t_{calc} = \frac{b_1 - (-13)}{S_{b_1}} = \frac{-16.203 + 13}{1.805} \cong -1.774$$

- Because $t(\alpha, n-2) = 1.667$, we would Reject the Null in favor of the Alternative

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%

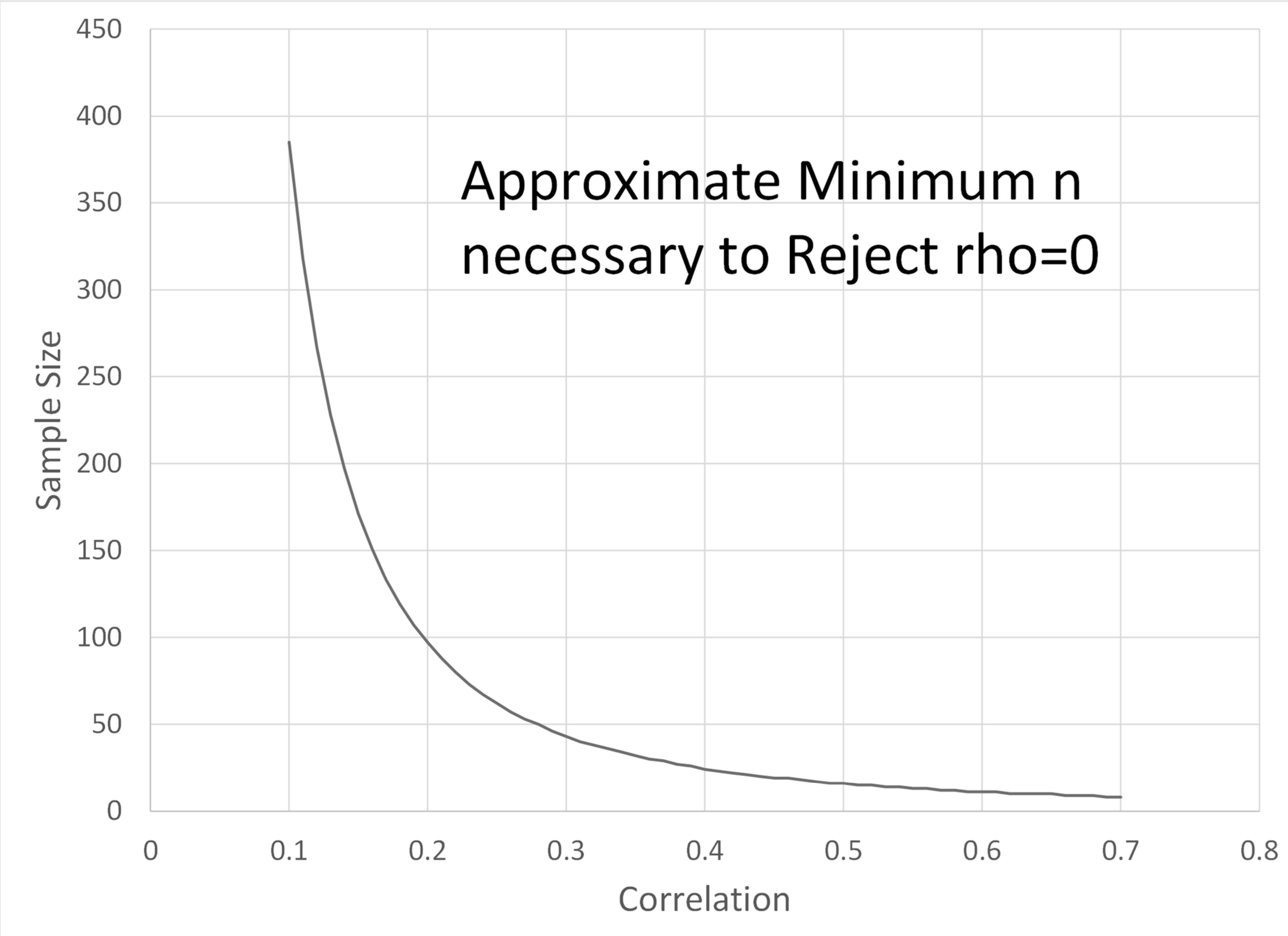
CORRELATION ANALYSIS

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

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

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

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


Approximate Minimum n
necessary to Reject rho=0
Sample Size
Correlation
0
50
100
150
200
250
300
350
400
450
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8

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

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

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

CORRELATION ANALYSIS

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

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

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

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

CORRELATION ANALYSIS

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

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

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

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

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

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)

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

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

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 $\beta_0 \; and \; \beta_1$
- 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)

REGRESSION WITH 2 GROUPS

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

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

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

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

MULTI-GROUP CORRELATION TEST

- To test a hypothesis of equal correlations across multiple groups: H_0: $\rho_1 = \rho_2 = \cdots = \rho_k$ 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 (upper) Chi-Square distribution with $k - 1$ degrees of freedom: $\chi^2(1 - \alpha, k - 1)$

MULTI-GROUP CORRELATION TEST

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