Statistics and Quantitative Analysis U4320

Lecture 13*: Explaining *Variation

Prof. Sharyn O'Halloran

Explaining Variation: Adjusted R^2 (cont)

- Definition of Adjusted R^2
  - So we'd like a measure like R^2, but one that takes into account the fact that adding extra variables always increases your explanatory power.
  - The statistic we use for this is call the Adjusted R^2, and its formula is:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k}(1-R^2);$$

n = number of observations,

Includes constant

k = number of independent variables.

  - The Adjusted R^2 can actually fall if the variable you add doesn't explain much of the variance.

Explaining Variation: Adjusted R^2 (cont)

- Back to the Example
  - Comparing Adjusted R^2
    - Model 1: .029
    - Model 2: .031
    - Model 3: .033
    - Model 4: .030
  - Interpretation
    - You can see that the adjusted R^2 rises from equation 1 to equation 2, and from equation 2 to equation 3.
    - But then it falls from equation 3 to 4, when we add in the variables for national parks and the zodiac.

Explaining Variation: Adjusted R^2 (cont)

- Example: Equation 2

Analysis of Variance	DF	f Squares	Mean Square
Regression	2	6.73848	3.36924
Residual	467	182.785	0.3914
F=8.60813		Signif F = 0.0002	

Multiple R	0.18856
R Square	0.03555
Adjusted R Square	0.03142
Standard Error	0.62562

Variable	B	SE B	Beta T	T Sig	T
SKOOL	0	0.01064	0.068897	1.459	0.1453
TUBETIME	-0	0.01444	-0.157772	-3.341	0.0009
(Constant)	2.1	0.15826	13.446	0	

- We calculate:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k}(1-R^2)$$
$$= 1 - \frac{470-1}{470-3}(1-.03555)$$
$$= .0314$$

Explaining Variation: Adjusted R^2 (cont)

- Stepwise Regression
  - One strategy for model building is to add variables only if they increase your adjusted R^2.
  - This technique is called stepwise regression.
  - However, I don't want to emphasize this approach to strongly.
    - Just as people can fixate on R^2 they can fixate on adjusted R^2.
    - If you have a theory that suggests that certain variables are important for your analysis then include them whether or not they increase the adjusted R^2.
    - Negative findings can be important!

Comparing Models: F-Tests

- When to use an F-Test?
  - Say you add a number of variables into a regression model and you want to see if, as a group, they are significant in explaining variation in your dependent variable Y.
  - The F-test tells you whether a group of variables, or even an entire model, is jointly significant.
    - This is in contrast to a t-test, which tells whether an individual coefficient is significantly different from zero.
    - In short, does the specified model explain a significant proportion of the total variation.

Comparing Models: F-Tests (cont.)

- Equations
  - To be precise, say our original equation is:

    Model 1: $Y = b_0 + b_1X_1 + b_2X_2$,

    We add two more variables, so the new equation is:

    Model 2: $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$.
  - We want to test the hypothesis that

    $H_0: \beta_3 = \beta_4 = 0.$

    We want to test the joint hypothesis that X_3 and X_4 together are not significant factors in determining Y.

Comparing Models: F-Tests (cont.)

- Using Adjusted R^2 First
  - There's an easy way to tell if these two variables are *not* significant.
    - First, run the regression without X_3 and X_4 in it, then run the regression with X_3 and X_4.
    - Now look at the adjusted R^2's for the two regressions.
      - If the adjusted R^2 went down, then X_3 and X_4 are not jointly significant.
  - So the adjusted R^2 can serve as a quick test for insignificance.

Comparing Models: F-Tests (cont.)

- Calculating an F-Test

  If the adjusted R^2 goes up, then you need to do a more complicated test, F-Test.

  - Ratio
    - et regression 1 be the model without X_3 and X_4, and let regression 2 include X_3 and X_4.
    - The basic idea of the F statistic, then, is to compute the ratio:

$$\frac{SSE_1 - SSE_2}{SSE_2}$$

Comparing Models: F-Tests (cont.)

- Correction
  - We have to correct for the number of independent we add.
  - So the complete statistic is:

m is the number of additional variables added to the model

$$F = \frac{\frac{SSE_1 - SSE_2}{m}}{\frac{SSE_2}{n-k}};$$

m = number of restrictions;

k = number of independent variables.

  - Remember: k is the total number of independent variables, including the ones that you are testing and the constant.

Comparing Models: F-Tests (cont.)

- Correction (cont.)
  - This equation defines an F-statistic with m and n-k degrees of freedom.
  - We write it like this:

$$F^m_{n-k}$$

  - To get critical values for the F statistic, we use a set of tables, just like for the normal and t-statistics.

Comparing Models: F-Tests (cont.)

- Example
  - Adding Extra Variables: Are a group of variables jointly significant?
    - Are the variables YE OWSTN and MYSI N jointly significant?

$$\text{Model 1: } \text{TRUSTTV} = b_0 + b_1\text{LIKEJPAN} + b_2\text{SKOOL} + b_3\text{TUBETIME}$$

$$\text{Model 2: } \text{TRUSTTV} = b_0 + b_1\text{LIKEJPAN} + b_2\text{SKOOL} + b_3\text{TUBETIME} + b_4\text{MYSIGN} + b_5\text{YELOWSTN}$$

Comparing Models: F-Tests (cont.)

- Adding Extra Variables (cont.)
  - State the null hypothesis

$$H_0 : B_4 = B_5 = 0$$

  - Calculate the F-statistic
    - Our formula for the F-statistic is:

$$F = \frac{\frac{SSE_1 - SSE_2}{m}}{\frac{SSE_2}{n-k}},$$

Comparing Models: F-Tests (cont.)

- What is SSE_1?
  - the sum of squared errors in the first regression.
- What is SSE_2?
  - the sum of squared errors in the second regression

m = 2 N = 4 0 k =

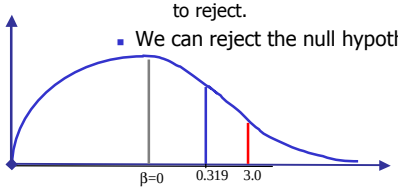
- The formula is:

$$F = \frac{\frac{182.07 - 181.82}{2}}{\frac{181.82}{470-6}}$$

$$= 0.319$$

Comparing Models: F-Tests (cont.)

- Reject or fail to reject the null hypothesis?
  - The critical value at the level
    - F^2_{470-6} from the table, is 3.00.
  - Is the F-statistic F^2_{470-6} ?
    - If yes, then we reject the null hypothesis that the variables are not significantly different from zero otherwise we fail to reject.
  - We can reject the null hypothesis because .319 3.00.



Comparing Models: F-Tests (cont.)

- Testing All Variables: Is the Model Significant?
  - Equation 2: Impact of school and TV watched

	DF	Sum of Squares	an Square	
Regression	2	6.74	3.37	
Residual	467	182.78	0.39	
F Statistic =	8.61		Signif F	2.000E-04

	Multiple R	0.18856
R Square		0.0356
Adjusted R Square		0.0314
Standard Error		0.6256

Dependent Variable: Trust TV					
Variable	B	SE B	Beta	T	Sig T
SKOOL	0.016	0.011	0.069	1.459	0.145
TUBETIME	-0.048	0.014	-0.158	-3.341	0.001
(Constant)	2.128	0.158	13.446	0.000	

Comparing Models: F-Tests (cont.)

- Hypothesis Testing:
  - State Hypothesis

$$H_0 : b_1 = b_2 = 0$$

  - Calculate test statistic
    - Again, we start with our formula:

$$F = \frac{\frac{SSE_1 - SSE_2}{m}}{\frac{SSE_2}{n-k}},$$

Comparing Models: F-Tests (cont.)

- Calculate F-statistic
  - SSE_2 = 1 2.
  - SSE_1 is the sum of squared errors when there are *no* explanatory variables at all.

    If there are no explanatory variables, then SSR must be 0. In this case, SSE=SST.
  - So we can substitute SST for SSE1 in our formula.

    SST = SSR + SSE = . 3 + 1 2. = 1 9. 4

This is the number reported in your printout under the F statistic.

$$F = \frac{\frac{189.54 - 182.78}{2}}{\frac{182.78}{470-3}}$$

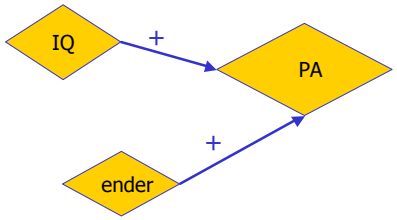
$$= 8.61.$$

Comparing Models: F-Tests (cont.)

- Reject or fail to reject the null hypothesis?
  - The critical value at the level, F^2_{470-} from your table, is 3.00.
  - So this time we can reject the null hypothesis that $\beta_1 = \beta_2 = 0$.
- Interpretation?
  - The model explains a significant amount of the total variation in how much people trust what is said on TV.

Comparing Models: Example

- Study of Seventh rade students in a mid-western school.
  - Path Diagram



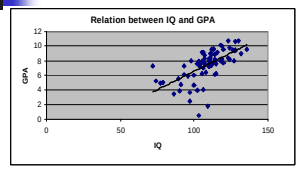
Comparing Models: Example(cont.)

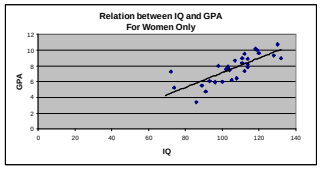
- Variables
  - IQ= student's score on a standard IQ test
  - PA= student's grade point average
  - ender= students gender (1 for male 0 for female)
- Descriptive Statistics

GPA		*IQ*		*Gender*	
Mean	7.45	Mean	108.92	Mean	0.60
Standard Error	0.24	Standard Error	1.49	Standard Error	0.06
Mode	9.17	Mode	111.00	Mode	1.00
Sample Variance	4.41	Sample Variance	173.47	Sample Variance	0.24
Kurtosis	1.10	Kurtosis	0.64	Kurtosis	-1.87
Minimum	0.53	Minimum	72.00	Minimum	0.00
Sum	580.83	Sum	8496.00	Sum	47.00

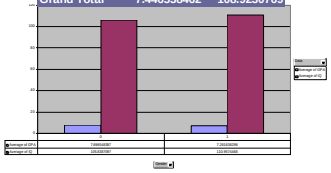
Comparing Models: Example(cont.)

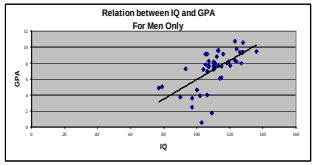
raphs





Gender	Average of GPA	Average of IQ
0	7.696548387	105.8387097
1	7.281638298	110.9574468
Grand Total	7.446538462	108.9230769





Comparing Models: Example(cont.)

- Hypothesis Testing:
  - Hypothesizes concerning coefficients
    - $H_0 : b_1 = 0$

      $H_a : b_1 \neq 0$
    - We want to know if IQ and ender explain a significant amount of the variation in PA.
  - Hypothesizes Concerning Models

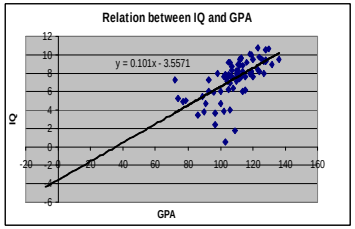
    $H_0 : b_1 = b_2 = 0$

    $H_a : b_1 = b_2 \neq 0$

Comparing Models: Example(cont.)

- Estimation
  - Model I

$$GPA = b_0 + b_1 IQ$$



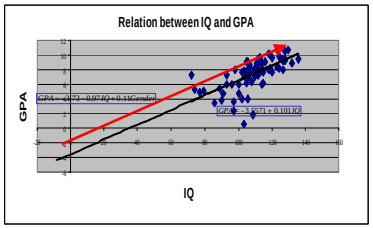
SUMMARY OUTPUT	Dependent Variable: GPA				
Regression Statistics					
Multiple R	0.63				
R Square	0.40				
Adjusted R Square	0.39				
Standard Error	1.63				
Observations	78.00				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	136.32	136.32	51.01	4.7373E-10
Residual	76	203.11	2.67		
Total	77	339.43			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-3.56	1.552	-2.29	0.024658962	
IQ	0.10	0.014	7.14	4.7373E-10	

$$GPA = -3.56 + 0.10 IQ$$

Comparing Models: Example(cont.)

- Model II:

$$GPA = b_0 + b_1 IQ + b_2 Gender$$



SUMMARY OUTPUT	Dependent Variable: GPA				
Regression Statistics					
Multiple R	0.67				
R Square	0.45				
Adjusted R Square	0.44				
Standard Error	1.58				
Observations	78.00				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	153.16	76.58	30.84	0.00
Residual	75	186.27	2.48		
Total	77	339.43			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-3.73	1.50	-2.49	0.01	
IQ	0.11	0.01	7.77	0.00	
Gender	-0.97	0.37	-2.60	0.01	

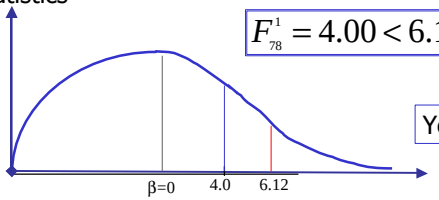
$$GPA = -3.73 - 0.97IQ + 0.11Gender$$

Comparing Models: Example(cont.)

- Is Model I better than Model II?

$$F = \frac{\frac{SSE_1 - SSE_2}{m}}{\frac{SSE_2}{n-k}}, \qquad \frac{\frac{203.11 - 186.27}{1}}{\frac{203.11}{77-3}} = \frac{16.84}{2.75} = 6.124$$

F-test Statistics



$$F_{78}^{1} = 4.00 < 6.124$$

Yes it is.

Comparing Models: Example(cont.)

- Interactive Terms

Regression Statistics	
Multiple R	0.675
R Square	0.456
Adjusted R Square	0.433
Standard Error	1.585
Observations	77.000

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	3	153.526	51.175	20.363	0.000
Residual	73	183.456	2.513		
Total	76	336.982			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-2.196	2.169	-1.013	0.315
IQ	0.093	0.020	4.583	0.000
Gender	-3.899	3.055	-1.276	0.206
IQ*Gender	0.027	0.028	0.974	0.333

The interactive term is not statistically significant. A high or low IQ has the same effect on GPA independent of gender.

Comparing Models: Example(cont.)

- Interpretation
  - Coefficients
    - Both IQ and ender matter.
      - IQ increases PA by .11 points holding ender constant.
      - ender Decreases PA by .9 points holding IQ constant.
  - Models
    - F-statistic shows that the model that includes ender performs significantly better in explaining variation then does the model with only IQ.
    - We are therefore able to reject the null hypothesis that model 1=model 2 at the significance level.

Final Paper

- Clearly state your hypothesis.
  - Use a path diagram to present the causal relation.
  - Use the correlations to help you determine what causes what.
  - State the alternative hypothesis.
- Present descriptive statistics.
  - This includes a correlation matrix and histogram or scatter plot.
- Estimate your model.
  - You can do simple regression, include interactive terms, do path analysis, use dummy variables whatever is appropriate to your hypothesis.
- Present your results.
- Interpret your results.
- Draw out the policy implications of your analysis.
- The paper should begin with a brief which states the basic project and your main findings.