DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816

TWO VARIABLE REGRESSION

I. AGENDA:
- A. Measuring goodness of fit
- B. Inference for regression models
  1. A t test
  2. An F test
  3. Confidence intervals for regression parameters
- C. Reading: Agresti and Finlay ***Statistical Methods in the Social Sciences***, 3rd edition, Chapter 9 as needed.

II. ASSESSING GOODNESS OF FIT:
- A. We will use several tools to decide whether a model fits the data:
  1. Measures of linear association
  2. Tests of significance
  3. Exploratory data analysis techniques
  4. Analysis of residuals and diagnostic plots.
- B. Each of these is by itself insufficient to assess the adequacy of a model but taken together they give one a good indication of how well one variable "explains" another.
- C. A MEASURE OF FIT:
  1. A simple measure, R^2
  2. A quick but sometimes misleading measure of how well an estimated model fits the data is the famous R^2 statistic.
  3. Here is its logic: As noted in class 9, the **total** variation in Y can be partitioned into two additive parts, an "explained" portion and an "error" or "residual" or "unexplained" portion.

$$TSS = RegSS + ResSS$$

***where*:**
TSS is the total sum of squares
RegSS is regression sum of squares
ResSS is residual sum of squares

- D. In symbols this is expressed as:

$$\sum_{i=1}^{N} (Y_i - \overline{Y}_i)^2 = \sum_{i=1}^{N} (\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$$

1. Note that each quantity has been squared.
2. The first term on the right represents what is explained, that is the proportion of the total variation in Y that is accounted for or statistically explained by X, the independent variable.
   i. Therefore, the proportion of total variation that is **explained** (in a statistical sense) is:

$$R^2 = RegSS/TSS = \frac{\sum_{i=1}^{N} (\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}$$

   ii. In MINITAB these quantities (the sums of squares, etc.) are routinely calculated and summarized in an ANOVA table.

E. Example:
1. Let's investigate the possible effects of air pollution on mortality. Here's another example from the Data and Story Library:
   i. "Researchers at General Motors collected data on 60 U.S. Standard Metropolitan Statistical Areas (SMSA's) in a study of whether air pollution contributes to mortality. The dependent variable for analysis is age adjusted mortality (called "Mortality"). The data include variables measuring demographic characteristics of the cities, variables measuring climate characteristics, and variables recording the pollution potential of three different air pollutants.
   ii. The independent variable in this analysis is SO2Pot, the: "sulfur dioxide pollution potential"
2. Here are the results of regressing mortality on sulfur dioxide:
   i. Note incidentally that in this table the total sum of squares equals the regression sum of squares plus the residual or error sum of squares:

225993 = 39698 + 186295

```
Mortality = 919 + 0.412 SO2

59 cases used 1 cases contain missing values

Predictor        Coef       StDev          T        P
Constant      918.671       9.853      93.24    0.000
SO2            0.4117      0.1181       3.49    0.001

S = 57.17       R-Sq = 17.6%     R-Sq(adj) = 16.1%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1       39698       39698     12.15    0.001
Residual Error    57      186295        3268
Total             58      225993
```

3. We can read the R^2 directly from the MINITAB results--it's R^2 = .176--but it can also be calculated from the information provided in the "Analysis of Variance Table:"

$$R^2 = RegSS/TSS = 39698/225993 = .1757$$

   i. Notice that I report the R^2 as a proportion, even though it can be interpreted as "percent of variation in Y explained by X."

4. We see that a linear model doesn't fit very well.
5. There are a couple of conclusions we might draw:
   i. This form of pollution (as measured) has relatively little impact on mortality.
   ii. This form of pollution combined with other factors may or may not contribute to mortality.
   iii. The data need to be adjusted, a point to which we'll return later.

III. TEST OF HYPOTHESES:

A. Since one frequently analyzes sample data, it is important to test propositions about the regression parameters such as β_1 is zero. If we decide that it is zero we will have concluded that there is no linear relationship between Y and X.

1. The null hypothesis of main interest is

$$H_0: \beta_1 = 0$$

B. We consider two tests about the regression parameters, β_0 and β_1.

1. First, assuming a linear model "fits" the data, we test individual coefficients using the t-test.
2. We can also conduct an "Analysis of Variance" or **ANOVA** test of the overall linear model.

C. Since the t-test involves a distribution with which we are already familiar we will start with it. Normally, however, one would begin with the ANOVA procedure.

IV. TESTS FOR INDIVIDUAL PARAMETERS:

A. Recall the steps in a hypothesis test:
1. Define the null and alternative hypotheses, which in turn suggest the tail of the distribution, if appropriate.
2. Determine the appropriate sampling distribution and its **standard error**.
3. Decide on decision rules such as level of significance and critical value(s).
4. Calculate the observed test statistic.
5. Make the decision and interpretation.

B. We'll test a hypothesis about the regression coefficient first.
1. The appropriate test statistic is a t with N - 2 degrees of freedom.
2. The observed t depends on an "effect size"--in this case it's just the estimated value of β_1--and a "standard error" of β_1

C. An important first step in finding the standard error is to find what is called the **Mean Square About Regression**:

$$\hat{}_{Y|X}^{2} = \frac{S^2}{(N - 2)}$$

1. The quantity S^2, you will recall, is the residual sum of squares. You can find it in MINITAB's regression output in the ANOVA table: it is the "Residual Error"

2. The symbol $\hat{}_{Y|X}^{2}$ is called the mean square about regression. Its **square root** is called the <u>standard deviation about the regression line</u> or <u>standard error of estimate</u>. Its formula is simple enough, just take the square root of the mean square of estimate.

$$\hat{}_{Y|X} = \sqrt{S^2/(N - 2)}$$

i. In the air quality and mortality example, we find the residual sum of squares, $S^2 = 186295$, then divide by (N - 2) = 59 - 2 = 57, a number called the degrees of freedom, to get the mean square about regression which is also called the <u>mean square for error</u>:

$$\hat{}_{Y|X}^{2} = \frac{186295}{(59 - 2)} = \frac{186295}{57} = 3268.33$$

ii. Note that the "degrees of freedom" 57, which is 59 minus 2. There is one missing data point, so our effective N is not 60, but rather 59.

iii. The standard error of estimate, denoted **S** in the MINITAB printout is:

$$\hat{}_{Y|X} = \sqrt{3268.333} = 57.1693$$

D. Now we are ready to test hypotheses:

E. **Test of the regression coefficient**: To test a coefficient like β_1, we have to find its standard error. (Recall that tests about the mean required a standard error.) The standard error of the regression coefficient is:

$$\hat{}_{\hat{\beta}_1} = \frac{\hat{}_{Y|X}}{\sqrt{\sum_{i=1}^{N} (X_i - \overline{X})^2}}$$

i. MINITAB labels this quantity the **standard deviation** of the coefficient (its abbreviation is **stdev**) and is found in the fourth column of the previous table that is, $\hat{}_{\hat{\beta}_1} = .1181$

ii. Test of the null hypothesis: H_0: $\beta_1 = 0$

a) Use a t test with N - 2 degrees of freedom

b) The test statistic is:

$$t_{N-2} = \frac{\hat{\beta}_1}{\hat{}_{\hat{\beta}_1}}$$

iii. The t test information--that is, the estimated standard error of β_1 and the estimated value are given in the print out; see the example.

a) The "**T**" in the example is 3.486.

b) Moreover the attained probability is also given. (In this case

it is .001.)

c) The result above suggest that the null hypothesis can be rejected at the .001 level. Or, the estimated regression coefficient is "statistically significant" at the .001 level.

d) Substantively this means that there seems to be a (weak perhaps) relationship between pollution and mortality, although we can by no means assert a cause-and-effect association.

F. <u>If you know ahead of time that the slope (i.e., regression coefficient is going to be, say negative), then use a one-tailed test; otherwise use a two-tailed test.</u> Always us your knowledge of the subject matter in making the test.

1. A table of t values can be found in Agresti and Finlay, ***Statistical Analysis for the Social Sciences***, 3rd edition.

i. In the example we will use a one-tailed test. N - 2 is 59 - 2 = 57, so we have 57 degrees of freedom. The critical value (for a one-tailed test, .05 level) is about 1.645. From the print out we see that the observed t-ratio is 3.49 Since the absolute value of the t_{obs} is greater than the t_{crit} we reject the null hypothesis that β_1 is zero.

G. The test of β_0 follows the same format.

1. Use a t test with N - 2 degrees of freedom and find the standard error of β_0, which is given in the computer results. The corresponding t-ratio is also given.

i. In the example, the degrees of freedom are again 57, so the critical value of the t (<u>two-tailed</u> test, .05 level) is 1.645. The observed t for the constant (β_0)is 93.24, so the t_{obs} is significant at least the .05 level and even beyond the .001 level.

H. Thus, for the population data we reject (at the .05 level) both null hypotheses about the regression parameters.

V. ANALYSIS OF VARIANCE TEST:

A. Our next test involves testing the regression model as a whole.

B. First recall the form of the general model with the error term:

$$Y_i = \beta_0 + \beta_1 X + \quad _i$$

1. This shows the population parameters and an error term.
2. We will essentially obtain two independent estimators of the variance of this error term.

C. Look once again at the analysis of variance table from the example pertaining to mortality and air quality payments presented above:

```
Mortality = 919 + 0.412 SO2

                        Analysis of Variance

Source            DF          SS          MS          F        P
Regression         1       39698       39698      12.15    0.001
Residual Error    57      186295        3268
Total             58      225993
```

D. Example: we can use this information to test an entire model as a whole. That is, when we add variables to "explain" variation in Y, we will want an overall test of, say,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

1. In this particular instance we have two independent variables and we want to test the model as a whole. Once we decide that it is acceptable, we can look at the individual coefficients to see which should be retain or dropped.

E. The table reflects the fact that the total sum of squares (TSS) can be partitioned or divided into two parts:
1. A "regression" or "explained" portion (the "regression" line in the table).
2. An "unexplained" or residual or (as the table indicates) "error" portion.
3. Recall that R^2 measures the proportion of the total sum of squares accounted for by the independent or explanatory or predictor variable.

F. Now we use the two table sum of squares ("regression" and "error" to form two <u>independent</u> estimates of the error variance $_i$.

1. Under the null hypothesis that the betas (regression parameters) are zero the expected value of the two will be the same. But if the null hypothesis does not hold, then one will be larger than the other by an amount that can't reasonably be attributed to chance.

G. **Mean squares:** A mean square is found by dividing a sum of squares by the appropriate degrees of freedom. The general idea is:

$$MS = \frac{Sum\ of\ squares}{degrees\ of\ freedom}$$

1. **Mean squares:** A mean square is an estimator of a variance.
   i. If we have two independent estimators of the same variance, the expected value of their ratios should be 1.0.

ii. If, however, the expected value of one mean square is larger than the expected value of another, then the expected value of the ratio would not be 1.0.

iii. We take advantage of this fact to test the hypothesis that: H_0: $\beta_1 = 0$.

H. The **mean square due error**:

$$MSError = \frac{Error\ sum\ of\ squares}{degrees\ of\ freedom}$$

$$= \frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{N - 2}$$

$$= \frac{S^2}{N - 2}$$

I. Given that certain assumption hold, this is an unbiased estimator of the error variance (the variance of the ε's). More formally:

$$E(MSError) = \quad _i$$

1. Recall that the **E** symbol means expected value. That is, over repeated independent samples from the population, the expected value of the mean squares (for error) will equal the standard deviation of the error variable.

J. The **mean square due to regression**:

$$MSRegression = \frac{Regression\ sum\ ofsquares}{degrees\ of\ freedom}$$

$$= \frac{\sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2}{1}$$

1. Now, <u>if the null hypothesis (that the regression parameter is zero) is true</u> (and certain assumptions hold), this will also be an unbiased estimator of

the error variance.

i. As in the case of the mean square for error, the expected value of this term (assuming $\beta_1 = 0$) is:

$$E(MSRegression) = {}_{i}$$

2. But if the null hypothesis does not hold, then this mean square will have a different and **larger** expected value.

K. To summarize:

1. We obtain (from the ANOVA) table two independent estimates of the error variance.
2. These estimates are called mean squares. There is one associated with residuals or errors and another with regression.
3. If the null hypothesis about β_1 is true, they will be estimating the same quantity. Over all possible sample from this population, the long-run or expected value of the ratio of these estimators should be 1.0.
4. If, on the other hand, the null hypothesis is not true, then the mean squares will not be estimating the same thing. It can be shown that the mean square for regression in this case estimates the variance of the errors plus something else. Hence, it will on average be larger than the mean square for error.
   i. Stated differently, if H_0 does not hold, then the long-run or expected value of the ratio of the mean squares over all possible samples from the population will be greater than 1.0.

L. The long and short of the situation is that we can use the ratio of the mean square for regression to the mean square for error (or residual) as an observed test statistic. This test statistic, like the t, has a distribution.

1. That is, if we took repeated samples from a population with β_1, found the mean squares, calculated their ratios (called them **F ratios**), and plotted these F's, they would have particular distribution called (logically enough) the **F distribution**.

M. The F distribution is really a family of distributions: each member is determined by the degrees of freedom associated with the mean square for regression (in this simple case, 1) and the mean square for error (in this case N - 2).

1. Consequently the table of the F distribution is a bit more complicated than the t or z tables. In particular, there is a table for various levels of significance. In the table for the .05 level the columns represent the degrees of freedom associated with the regression mean square; the rows the degrees of freedom for the mean square for error. (Some tables refer the columns as the numerator degrees of freedom, since the mean square for regression is the numerator of the F ratio) and the columns as the denominator degrees of freedom.

2. You can find an F distribution table in Agresti and Finlay, ***Statistical Methods for the Social Sciences***, 3rd edition, pages 671 to 673.
   i. As noted the first table gives critical values for the .05 level; the second table for the .01 level; and the third for the .001 level. These are the only α's that you can look up.
   ii. The columns are the numerator or regression degrees of freedom.
   iii. The rows are the error degrees of freedom.
3. Example:
   i. In the previous case the regression degrees of freedom is one and the error degrees of freedom are 57 (look in the column labeled "DF").
   ii. At the point .05 level the critical F associated with 1 and 57 degrees of freedom is about 4.00. (It's somewhere between 4.08 and 4.00; look in the first column next to 40 and 60.)
   iii. For the .01 level the critical value is about 7.08.
   iv. And for the .001 level it is about 11.97.

N. Let's use this information and ideas to test the hypothesis that $\beta_1 = 0$.
1. We'll use the F test.
2. If we observe an F greater than or equal to 4.00, we will reject the null hypothesis and conclude that there is a (perhaps weak) linear relationship between mortality and air pollution as measured by sulfur dioxide.
3. The observed F is

$$F_{(1,57)} = \frac{\frac{39698}{1}}{\frac{186295}{57}} = \frac{39698}{3260} = 12.15$$

4. We see that the mean square for regression is 39698/1 = 39698 and the mean square for residual or error is 186295/57 = 3260.
5. The F ratio is thus 12.15.
6. This value exceeds the critical values determined above.
7. Consequently, we would reject the null hypothesis, a decision we reached by an equivalent means with the t test.
   i. In fact, later in the semester we'll see that t^2 equals F.

VI. CONFIDENCE INTERVALS:

A. Confidence intervals can be placed around parameter estimators in the usual way.

B. For (1 - α)% confidence intervals use a t with appropriate degrees of freedom:

$$\hat{\beta}_j \pm t_{(\alpha/2, n-2)} \hat{\sigma}_{\hat{\beta}_j}$$

1. The estimated β_j for j = 0, 1 is bracketed by an appropriate t times the estimated standard deviation (error) of the coefficient.
2. The $t_{(\alpha/2, N-2)}$ is chosen from the t distribution with N - 2 degrees of freedom and $\alpha/2$ level of significance.

C. Example: the estimated regression coefficient of mortality on air pollution was .412, with an estimated standard error of .1181.

1. The 95% intervals are thus

$$.412 - .1181(1.96) \leq \beta_1 \leq .412 + .1181(1.96)$$

$$.181 \leq \beta_1 \leq .643$$

2. 99 percent intervals are

$$.412 - .1181(2.576) \leq \beta_1 \leq .412 + .1181(2.576)$$

$$.108 \leq \beta_1 \leq .716$$

VII. NEXT TIME:

A. More on confidence intervals.

B. Correlation

C. Transformations

Go to Notes page

Go to Statistics page