DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816

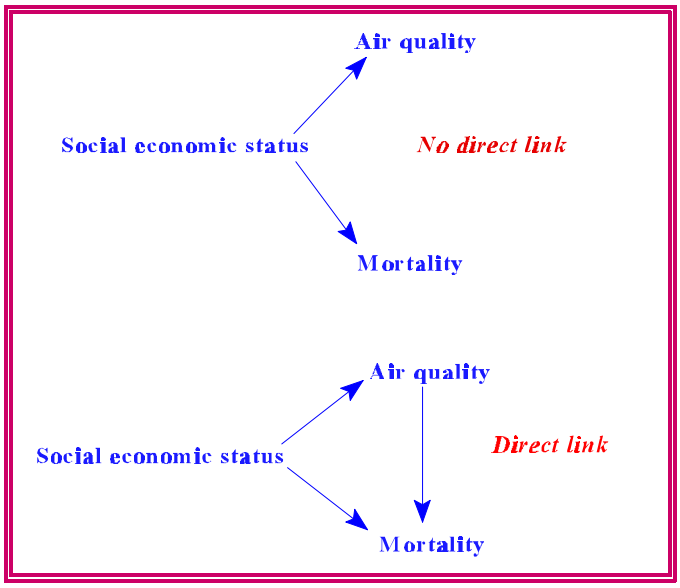
MORE ON MULTIPLE REGRESSION

I. AGENDA:
- A. Illustration of multiple regression
  1. Model building
  2. Coefficient interpretation
  3. Tests
  4. Standardized version
- B. Reading: Agresti and Finlay ***Statistical Methods in the Social Sciences***, 3rd edition, Chapter 11.

II. EXAMPLE OF MULTIPLE REGRESSION:
- A. Model building:
  1. Suppose we want to know if pollution has any impact on mortality (defined as age adjusted mortality per 100,000 population).
     - i. The data were used in an early class.
  2. So, of course, we want to see the effect of air pollution on mortality, but we also know that lots of factors affect the rate.
     - i. A person's standard of living, for example: access to health care, diet, life style or habits, exposure to harmful "agents" including fire arms and the like
  3. We can't measure every factor. Indeed, we won't (for this example) be able to study individuals directly.
     - i. Instead we'll have to look at communities (American cities) and make inferences about the effects of pollution and other factors measured at the aggregate level on individuals.
  4. Since there are a large number of possible explanatory or causal factors, we need some method for deciding which variables to include in the model of pollution and health.
     - i. For present purposes we will rely on intuition or common sense and health research to select variables.
     - ii. But in addition we will use some statistical measures to judge how well adding and subtracting variables improves our model.
- B. Example: out the outset let us see whether or not air pollution is related to communities' mortality rates.
  1. Then we will add other social and economic factors in an attempt to "capture" the effects of life style and life chances on the death rate.
     - i. That is, we might find a relationship between some pollutant or

pollutants and mortality, but we need to consider the possibility that pollution itself is related to social and economic environment.

ii. The model in Figure 1, which is similar to many others we've seen, illustrates the point.



iii. In the first diagram the partial (controlled) relationship between air quality (pollution) and mortality would be zero.

1) The "bi-variate" or two variable relationship that does not take social-economic status into account might not be zero.

iv. The second diagram shows that even if we hold social economic status constant, there will be a relationship between pollution and mortality.

2. In planning our analysis we would think about the direction of the relationships.

i. So, for example, a measure of income or education would no doubt be **negatively** related to mortality. (That is, the greater the education, the less...)

C. Data:

1. The data, used previously, pertain to N = 59 American cities. The variables are:

i. Mortality: Age adjusted mortality per 100,000

ii. Indicator of pollution: SO2-Pot: Sulfur Dioxide pollution potential.

iii. Measures of social-economic envirionment

1) Education: Median education

2) Percentage of that is population non-white

3) Percentage of work force that is white collar

iv. Measures of physical environment:

1) Rain: Annual rainfall (inches)

2) July Temp: Mean July temperature (degrees Farenheit)

2. We won't include all of these variables in the final model.

D. Strategy:

1. Let's start with the bivariate relationship between the two principal variables, mortality (Y) and air pollution (sulfur dioxide, X).
2. We would first examine a plot.
3. We would also look at the individual distributions. Doing so in fact, suggests that sulfur dioxide is skewed toward the low end of the scale. Consequently, we'll transform it later.

E. Results

1. Here are the relevant statistics for the two-variable relationship.

i. We have seen these in a previous class.

```
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
```

ii. The simple regression coefficient indicates that as pollution increases one unit (may be a gram per cubic liter of air; I don't know what the scale is), mortality **increases** .412 deaths per 100,000.

1) That seems like a lot given that the pollution variable may be measured in fairly small quantities.

iii. We see that the coefficient is statistically significant at the .001 level (what does this mean?) and that the explained variation (by X) is

about 18 percent.

1) We can tell the model is significant by examining the F statistic, which has 1 and 59 - 2 = 57 degrees of freedom.

iv. It appears that there is a meaningful relationship. But, of course, it is entirely possible that poor people, who have higher mortality rates than more affluent individuals, also happen to live in more polluted areas and that the relationship above is spurious or weaker than the bivariate result suggests.

2. So let's add a social-factor, median education.

```
Mortality = 1274 + 0.314 SO2 - 31.9 Educat

59 cases used 1 cases contain missing values

Predictor        Coef       StDev          T        P
Constant      1273.99       90.51      14.08    0.000
SO2            0.3138      0.1083       2.90    0.005
Educat        -31.914       8.091      -3.94    0.000

S = 51.02       R-Sq = 35.5%     R-Sq(adj) = 33.2%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2       80205       40102     15.40    0.000
Residual Error    56      145788        2603
Total             58      225993

Source       DF      Seq SS
SO2           1       39698
Educat        1       40507
```

3. The coefficient relating pollution to mortality has changed slightly because it's now a **partial** regression coefficient that shows the impact of sulfur dioxide on mortality after the level of education has been taken into account or controlled.

i. That is, imagine that we divided communities into groups in which the level of education was exactly the same. We then regressed mortality on sulfur dioxide within each group. The partial coefficient can be thought of as a kind of weighted average of those coefficients.

ii. It appears that pollution has an independent effect on mortality.

4. Similarly the partial coefficient relating education to mortality with pollution controlled indicated that there is a net effect; β = -31.9, which means that each year increase in (median) education is associated with a

large decrease in mortality.

i. But note: just because this β is -31 and the other is only .314 we can't say education is a "more important" factor, although it might be.

5. The portion of the variation explained has increase to .355.
   i. In fact, we can and will use the increase in explained variation as a test of whether or not adding a variable increases in a statistically significant way the explanatory power of the model.
   ii. By the same token, note that the standard deviation about the regression line has decreased from 57.17 to 51.02.
      1) This of course makes sense since many of these terms are tied together as we have seen.
   iii. The R^2 always increases as variables are added to a model. The question is: is there a **meaningful** (in a statistical and especially in a substantive sense) increase.

F. Hypothesis tests.

1. We can conduct a **collective** or global of the model of as a whole (see Agresti and Finlay, Statistical Methods, 3rd edition, pages 399 to 401.
   i. It tests the hypothesis that

$$\beta_{mortality,SO^2|education} = \beta_{mortality,education|SO^2} = 0$$

   ii. The alternative hypothesis is that at least one β_i is not zero.
   iii. We use the F test with F_{obs}

$$F_{obs} = \frac{40102}{2603} = 15.40$$

      1) The critical F with **2** and **56** degrees of freedom at the .001 level is about 7.76 (I used Agresti and Finlay's table with 60 degrees of freedom).
      2) Since the observed F exceeds the critical value, we can reject the null hypothesis that both partial regression coefficient equal zero.

2. Tests of individual parameters.
   i. We can test the hypothesis that a particular partial regression coefficient equals 0 or

$$\beta_k = 0$$

$$for\ k = 1,2$$

ii. This is a t test with degrees of freedom N - K - 1 = N - (K + 1) = 59 - 2 - 1 = 56, the error degrees of freedom.

1) Remember K is the number of independent variables. In effect we "lose" a degree of freedom for each parameter in the model.

2) The t statistic is

$$t_{obs} = \frac{\hat{\beta}_i}{\hat{\sigma}_{\beta_i}}$$

3) The critical t a the .005 level with 56 degrees of freedom is about (see Agresti and Finlay, 3rd edition, the row for infinity) 2.576.

4) Both observed t's exceed this value.

3. As Agresti and Finlay note, it is important to estimate the size of a partial regression coefficient, and intervals at the $\alpha = .01$ level can be found by finding the critical t with $\alpha/2 = .005$ and 56 degrees of freedom = 2.576 and

$$\hat{\beta}_i \pm \hat{\sigma}_i \ (2.576)$$

i. If you want other levels of confidence, use the appropriate t with degrees of freedom 59 - 2 - 1 = 59 - 3 = 56.

ii. 99 percent intervals are:

$\hat{\beta}_1$: **.3138 ± (.1083)(2.576)**

lower = **.0348, *upper*** = **.5928**

and

$\hat{\beta}_2$: **−31.914 ± (8.091)(2.576)**

lower = **−57.7564, *upper*** = **−11.0716**

1) Neither includes zero as we would expect from the t tests.

G. Some relationships to look at :

1. The total sum of squares will remain the same as we move from one model to the next so long as we use the same Y. That is, TSS remains the same.

When we add variables the part of this total that is "explained" increases (however slightly) and the unexplained portion decreases.

i. Look at the two tables to make sure that you understand this point.

2. Note by the way that the regression has 2 degrees associated with it; that's because we have two independent variables.

i. And, the error degrees of freedom has decrease by one. It has gone into the regression portion.

ii. Summary:

$$TSS = RegSS + ResSS$$

and

$$df_{total} = df_{regression} + df_{residual}$$

but

$$MSTotal \neq MSRegression + MSResidual$$

1) Sums of squares and degrees of freedom are **additive** but mean squares are not. You can verify this in the table.

iii. We can also partition the regression sum of squares into parts, one for each variable, for example. (Look at the bottom of the table.)

H. Let's add another variable, percent of community that is non-white.

```
Mortality = 1155 + 0.252 SO2 - 24.8 Educat + 3.71 %Nonwht

59 cases used 1 cases contain missing values

Predictor        Coef       StDev          T        P
Constant      1154.99       72.15      16.01    0.000
SO2           0.25182     0.08390       3.00    0.004
Educat        -24.773       6.327      -3.92    0.000
%Nonwht        3.7123      0.5899       6.29    0.000

S = 39.26       R-Sq = 62.5%     R-Sq(adj) = 60.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3      141237       47079     30.55    0.000
Residual Error    55       84755        1541
Total             58      225993
```

1. We see that the explanatory "power' of the model has increased.
   i. R^2 has jumped to .625 and the standard deviation about the regression line (S) has fallen a bit to 39.26.
   ii. Each of the coefficients is significant at the .01 level and two at the .001 level.
   iii. An overall test gives F_{obs} = 30.85, which is "significant" at the .001 level.

I. Lets add still another explanatory variable, percent of the workforce in white collar occupations.
   1. Here are the results.

```
The regression equation is
Mortality = 1171 + 0.256 SO2 - 21.0 Educat + 3.74 %Nonwht - 1.27 %Whtc

59 cases used 1 cases contain missing values

Predictor        Coef       StDev          T        P
Constant      1171.37       73.58      15.92    0.000
SO2           0.25644     0.08387       3.06    0.003
Educat        -20.951       7.224      -2.90    0.005
%Nonwht        3.7416      0.5895       6.35    0.000
%Whtc          -1.270       1.164      -1.09    0.280

S = 39.19       R-Sq = 63.3%     R-Sq(adj) = 60.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         4      143063       35766     23.29    0.000
Residual Error    54       82929        1536
Total             58      225993
```

   i. We see now that percent white collar doesn't add much:
      1) The R^2 increases only slightly.
         a) We could test the increase for statistical significance using the sums of squares but will wait until later to do so.
      2) The coefficient for occupation, moreover, is not significant at even the .05 level.

J. So let's drop it and add an "atmosphere' variable, rainfall.
   1. Here are the results.

```
The regression equation is
Mortality = 1024 + 0.313 SO2 - 16.9 Educat + 3.34 %Nonwht + 1.19 Rain

59 cases used 1 cases contain missing values

Predictor        Coef       StDev          T        P
Constant      1024.32       90.83      11.28    0.000
SO2           0.31270     0.08543       3.66    0.001
Educat        -16.923       7.041      -2.40    0.020
%Nonwht        3.3363      0.5936       5.62    0.000
Rain           1.1872      0.5297       2.24    0.029

S = 37.89       R-Sq = 65.7%     R-Sq(adj) = 63.1%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         4      148450       37113     25.85    0.000
Residual Error    54       77542        1436
Total             58      225993
```

2. Again the multiple R does not increase much but the rainfall coefficient is significant at the .05 level.
   i. It's attained probability is .029.
   ii. S is now down to 37.89.

K. So, let's add one more variable (July temperature) making K = 5 in all.

```
The regression equation is
Mortality = 1229 + 0.284 SO2 - 17.9 Educat + 4.11 %Nonwht + 1.48 Rain
          - 2.85 Julytemp

59 cases used 1 cases contain missing values

Predictor        Coef       StDev          T        P
Constant       1228.5       136.4       9.01    0.000
SO2           0.28444     0.08447       3.37    0.001
Educat        -17.938       6.880      -2.61    0.012
%Nonwht        4.1106      0.6995       5.88    0.000
Rain           1.4849      0.5379       2.76    0.008
Julytemp       -2.852       1.449      -1.97    0.054

S = 36.92       R-Sq = 68.0%     R-Sq(adj) = 65.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         5      153732       30746     22.55    0.000
Residual Error    53       72260        1363
Total             58      225993
```

1. Once again we see an improvement, albeit not a dramatic one.
   i. Each individual coefficient is significant and the standard deviation or standard error of regression is now 36.92.
   ii. To repeat we could compare the R^2 for the previous model (the one without temperature, call it the **reduced** model) with the R^2 for this last one (called it the complete model).
      1) We would use an appropriate F statistic, which could be compared with a critical value to see if the extra variable significantly increased the model's explanatory power.

L. We'll consider this the final model, for now.
   1. Interpretation:
      i. Most of the coefficient have direct and substantively interesting interpretations.
      ii. The exceptions are those relating to rainfall and temperature. I suspect that they are "surrogate" or "indicator" variables for more direct causes of mortality.
         1) Look at rainfall, for example. The estimated β is 1.405, which suggests that as precipitation increases so too does mortality. But it's not likely that there is a **causal** connection.
         2) It's possible, though, that "weaker" people live in areas of greater rainfall, which might account for the apparent relationship. (Why not explore this idea?)

2. What's most important?
   i. We can look at each coefficient to see the impact of the corresponding variable on mortality rates.
   ii. Given the caveat in the previous paragraph we would want to think seriously about each variable.
   iii. Right now it would be hard to say that one is more important than another.
   iv. Even so, let's standardize.

III. STANDARDIZED VARIABLES:

A. As noted in Class 12, we can standardize all of the variables by subtracting means and dividing by standard deviations.
   1. The newest versions of MINITAB makes the process especially easy since they have "standardize" procedures.
   2. But it's easy enough to do with the calculation procedures in both the Student and full versions.

B. Standardized regression coefficients β^*_k are the partial regression coefficients one would get by standardizing Y and X_k for k = 1, 2,...K.
   1. See Agresti and Finlay, Statistical Methods, 3rd edition, page 416.
   2. We can standardize as described above by transforming the variables or by using sample standard deviations.
   3. The general formula is

$$\hat{\beta}_k^* = \hat{\beta}_k \left(\frac{\hat{\sigma}_{X_k}}{\hat{\sigma}_Y} \right)$$

      i. Here $\hat{\beta}_k$ is the estimated partial regression coefficient and the $\hat{\sigma}'s$ are the sample standard deviaitions.
      ii. The standardized regression coefficient measures the effect of X_k in standard deviation units.
      iii. Presumably if $|\hat{\beta}_k^*| > |\hat{\beta}_j^*|$ then X_k has a greater (partial or controlled) affect on Y than X_j does.
   4. Example of calculation:
      i. We'll use the estimated coefficients from the last model, the one that included sulfur dioxide, education, percent non-white, rainfall, and July temperature.

$$\hat{\beta}^{*}_{mortality,so2|X_2...} = .28444\left(\frac{63.552}{62.421}\right)$$

$$= .28959$$

ii. The sample standard deviations for Y and X (sulfur dioxide) are 62.421 and 63.552 respectively.

5. Another example: the standardized coefficient for education is:

$$\hat{\beta}^{*}_{mortality,Educ|X_1...} = -17.938\left(\frac{.85068}{62.421}\right)$$

$$= .24446$$

6. It appears that education and sulfur dioxide emissions have about the same impact on mortality, although in opposite directions.

i. The sample standard deviation for education is .85068.

C. The standardized coefficients for the last model (and test and summary statistics) are :

```
The regression equation is
smort = - 0.0000 + 0.290 sSO2 - 0.244 seduc + 0.593 snonwht + 0.275
srain
           - 0.210 sJuly

59 cases used 1 cases contain missing values

Predictor        Coef       StDev          T        P
Constant     -0.00000     0.07701      -0.00    1.000
sSO2          0.28959     0.08600       3.37    0.001
seduc        -0.24445     0.09376      -2.61    0.012
snonwht        0.5925      0.1008       5.88    0.000
srain         0.27531     0.09972       2.76    0.008
sJuly         -0.2102      0.1068      -1.97    0.054

S = 0.5915      R-Sq = 68.0%     R-Sq(adj) = 65.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         5     39.4547      7.8909     22.55    0.000
Residual Error    53     18.5453      0.3499
Total             58     58.0000
```

1. If we use the size of the standardized coefficients as the yardstick, percent non-white appears to have the greatest impact on mortality, a conclusion which is plausible.
2. The other variables seem to be about equally important.
3. Note the equivalences (R^2 = .68, for instance) and differences (the scales have changed and so too have the sums of squares.
   i. But the t values are all the same, except for the one pertaining to the constant. The constant is now 0--standardization forces it to be 0--so the t is 0 as well.
   ii. Furthermore, F_{obs} = 22.55 in both tables.
   iii. You can make other comparisons yourself.

IV. NEXT TIME:
   A. More on multiple regression.
   B. Dummy variables

Go to Notes page

Go to Statistics page