DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816

MULTIPLE REGRESSION

I. AGENDA:
   A. Standardized regression coefficients
   B. Multiple regression model.
   C. Reading: Agresti and Finlay ***Statistical Methods in the Social Sciences***, 3rd edition, Chapter 10 and Chapter 11, 383 to 394.

II. STANDARDIZED VARIABLES AND COEFFICIENTS:

   A. We have noted several times that the magnitude of β_1, or its estimator $\hat{\beta}_1$, reflects the units in which X, the independent variable is measured.

   B. Suppose, for example, we had this equation:

$$\hat{Y}_i = 10 + 1.0X_1 + 1000X_2$$

      1. Furthermore, suppose X_1 is measured in thousands of dollars and X_2 is measured in just dollars.

      2. Note that although $\hat{\beta}_2$ is considerably larger than $\hat{\beta}_1$, both variables have the same impact on Y, since a "one-unit" change in X_2 (when the other variable is held constant, see below) has effectively the same effect as "one-unit" change in X_1. (Changing X_2 one unit of course amounts to a dollar change; but because this dollar change is multiplied by 1000 the effect on Y is 1000, the same as the effect of $\hat{\beta}_1$. Why?

      3. In essence, the regression coefficient β_1 is measured in units that are units of Y divided by the units of X.

   C. For this reason and because many measures in the social and policy sciences are often non-intuitive (e.g., attitude scales) coefficients are frequently difficult to compare.
      1. Hence, it sometimes useful to rescale all of the variables so that they have a "dimensionless" or common or standard unit of measurement.
      2. Standardizing variables, as mentioned in previous classes, provides a means

for doing so.

D. Deviation scores:

1. Transform each X and each Y into deviation scores as follows:

$$y_i = (Y_i - \overline{Y})$$

$$x_i = (X_i - \overline{X})$$

2. The regression of y on x will lead to an equation in which the constant is zero. This is sometimes called regression through the origin.

E. Standardized regression coefficient:

1. The regression coefficient, remember, is measured in units of the original variables.
2. The correlation coefficient can be interpreted as a standardized slope or regression coefficient: it is a slope whose value does not depend on units of measurement. It is, so to speak, scale free.
3. The standardization is achieved by adjusting the unstandardized regression coefficient β_1 by the standard deviations of X and Y. This leads to a coefficient that can be thought of as the value one would get for the slope of the regression of Y on X if the standard deviations of X and Y were equal.
4. How does one make the standard deviations of X and Y equal? By "standardizing" each variable according to the formula"

$$x_i' = (X_i - \overline{X})/\hat{\sigma}_X$$

$$y_i' = (Y_i - \overline{Y})/\hat{\sigma}_Y$$

where $\hat{\sigma}$'s are the sample standard deviations

   i. In other words, one can standardize X and Y to get $\mathbf{x}'$ and $\mathbf{y}'$ and then regress $\mathbf{y}'$ on $\mathbf{x}'$. The resulting equation will yield a slope which is equal to the value of r between X and Y.

5. In more complex situations where there are several variables and equations, the standardized regression coefficients are sometimes called path coefficients.

F. Another way to get the standardized coefficient, a way that might be useful later, is to adjust the estimated slope by the sample standard deviations:

$$r = \left(\frac{\hat{\sigma}_X}{\hat{\sigma}_Y} \right) \hat{\beta}_1$$

G. Here is an example:

1. The data are the percent of the county that voted for H. Ross Perot in 1992 in New Jersey (Y) and the county's population per square mile (X).

County	% for Perot	*Standardized Y*	Density	*Standardized X*
Atlantic	17.6	0.00890	154.4	-0.56493
Bergen	12.9	-0.96683	1359.8	0.55881
Burlington	20.4	0.59018	189.6	-0.53211
Camden	17.6	0.00890	873.0	0.10499
Cape May	20.1	0.52790	143.9	-0.57472
Cumberland	19.0	0.29954	109.0	-0.60725
Essex	9.7	-1.63116	2379.8	1.50971
Glousest	23.1	1.15071	273.6	-0.45380
Hudson	7.9	-2.00484	4571.1	3.55255
Hunterdon	23.6	1.25451	96.7	-0.61872
Mercer	15.5	-0.42707	557.0	-0.18960
Middlesex	15.8	-0.36479	834.5	0.06910
Monmouth	17.1	-0.09490	452.6	-0.28693
Morris	15.5	-0.42707	346.8	-0.38556
Ocean	19.3	0.36182	262.9	-0.46378
Passaic	13.0	-0.94607	945.8	0.17286
Salem	26.0	1.75276	74.6	-0.63932
Somerset	17.4	-0.03262	304.5	-0.42500
Sussex	22.0	0.92235	97.0	-0.61844
Union	11.4	-1.27824	1842.6	1.00890
Warren	23.8	1.29603	98.8	-0.61676

Table: Raw and Standardized Scores

H. Regressing a standardized Y on a standardized X produces standardized regression coefficients. (In some articles they are called <u>beta coefficients</u> or <u>path coefficients</u>, but since this usage can be confusing, I will not use it.)

I. Here are some characteristics of standardized variables and coefficients.

1. Standardized variables have a mean of 0 and standard deviation of 1.0. You can verify this yourself, both by looking at the formula for standardization or, less satisfactorily, calculating the mean and standard deviation of this

sample.

i. Use this property to check your work.

2. In the two variable case the correlation coefficient between Y (e.g., percent for Perot) and X (e.g., population density) equals the standardized regression coefficient.

i. In this example:

$$r_{Perot,density} = \hat{\beta}^{*}_{Perot,density} = .804$$

1) The star (*) indicates a standardized coefficient

3. Standardizing (by subtracting $\bar{X}$) removes the intercept or constant from the equation. Hence, when using standardized variables, β_0 will be zero. In this instance, the predicted vote for Perot is thus:

$$\hat{y} = 0.00 + .804x_1$$

or

$$\hat{y} = .804x_1$$

where **x** and **y** are the standardized variables.

III. MULTIPLE REGRESSION MODEL:

A. There is a single dependent variable, Y, which is believed to be a <u>linear</u> function of **K** independent variables.

1. In the example, K = 4 because there are four independent variables, X_1, X_2, X_3, and X_4.

2. The general model is written as:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + \varepsilon_i$$

3. A model with k = 4 independent variables is:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon_i$$

4. Interpretation:
   i. Systematic part:
      1) The regression parameters, β's, represent the effects of each independent variable on Y when the other variables in the model have been controlled.
      2) Thus, β_1, is the effect of X_1 when the other X's have been controlled.
      3) The reason the word "controlled" appears is that the independent variables themselves are interrelated. Changing the value of, say, X_1, not only changes Y but might also affect X_2 which in turn impacts on Y. To see the "pure" or "uncontaminated" effect of X_1 on Y we need to hold the other X's constant.
   ii. A path diagram may help explain. Consider the models in Figure 1.

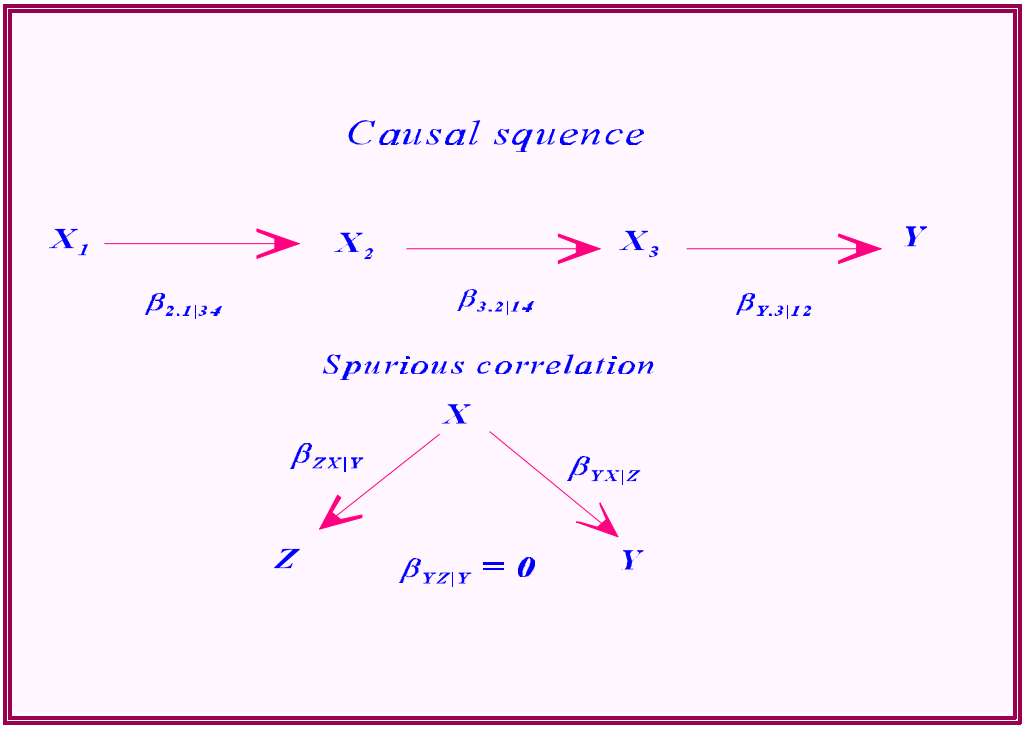

Figure 1: Structural Models

5. Note: that multiple regression coefficients are often written with the dependent variable, Y, an independent variable (X, for example) second, and any variables that are being controlled after the dot. Thus, $\beta_{YZ.X}$ means the regression coefficient between Y and Z, when the X has been (statistically) held constant.
6. In the first part of Figure 1, Y depends on X_1, X_2 and X_3. Changing X_1 will

affect the value of Y, by affecting X_2 constant, which in turn affects X_3. Each of the coefficients is non-zero.

7. Moreover, even if we hold a variable prior to one of the independent variables constant (say, we hold X_1 constant), if we could change the value of the next independent down the line (X_2, say), we will affect Y,
   i. The first model then represents a direct "causal" sequence--sometimes called "developmental sequence."
   ii. It this model holds, then all of its parameters are non-zero.
8. Now look at the second model ("spurious correlation"): the arrows (and lack of arrows) mean that there is a direct (causal?) connection betweeen X and Z and X and Y with X being the causal factor, but no direct connection between Z and Y.
   i. That is, in the second model if we hold X constant, changing Z has no effect on Y.
   ii. This situation means that the **partial regression parameter**, $\beta_{YZ|X}$ is zero. A zero value, in other words, indicates that although there may be a "bi-variate" relationship between Z and Y, the **partial** or **controlled** relationship is nil.

B. Thus, the regression procedure produces partial or controlled coefficients which means, for instance, that Y changes β_1 units for a one-unit change in X_3 when X_1 and X_3 have been held constant.
   1. Note that direct linkages are indicated by arrows; an arrow represents the presence of a non-zero coefficient.

IV. EXAMPLE:

A. Here are some economic and social data:

PROJECTED POPULATION INCREASE

Nation	Birth rate X_1	Death rate X_2	Life expectancy X_3	GNP per capita X_4	Percent projected population increase Y
Bolivia	**42**	**16**	**51**	**510**	**53.2**
Cuba	**17**	**6**	**73**	**1050**	**14.9**
Cyprus	**29**	**9**	**74**	**3720**	**14.3**
Egypt	**37**	**10**	**57**	**700**	**39.3**
Ghana	**47**	**15**	**52**	**320**	**60.1**
Jamaica	**28**	**6**	**70**	**1300**	**21.7**
Nigeria	**48**	**17**	**50**	**760**	**71.6**
South Africa	**35**	**14**	**54**	**2450**	**40.1**
South Korea	**23**	**6**	**66**	**2010**	**21.1**
Turkey	**35**	**10**	**63**	**1230**	**36.9**

B. One can ask some questions:
   1. What explains variation in Y, projected population increase?
   2. What are the "individual" effects of the independent variables?
   3. Are any of them redundant?
   4. How well does a linear model as a whole fit the data?
   5. What policy implications, if any, does the model contain?

C. Here are “estimates” of the coefficients:

$$\hat{\beta}_0 = 27.7$$
$$\hat{\beta}_1 = .728$$
$$\hat{\beta}_2 = 1.46$$
$$\hat{\beta}_3 = -.422$$
$$\hat{\beta}_4 = -.00406$$

i. The estimated model is thus:

$$\hat{Y} = 27.7 + .728X_1 + 1.46X_2 - .422X_3 - .00406X_4$$

1. The first is the constant: it is the value of Y when all X's are zero.
   i. The first regression parameter, $\hat{\beta}_1$, means that Y increases .728 units for a one-unit change in X_1 when X_2, X_3, and X_4 have been held constant.
   ii. The second parameter is interpreted in a similar way: Y changes by 1.46 units for every one-unit change in X_2, assuming that X_1, X_3, and X_4 have been held constant.
   iii. Note that partial regression coefficients are statistical method of physically holding variables constant. In other words, observational analysis limits our ability to manipulate variables so we compensate by making statistical adjustments.

V. RANDOM COMPONENT:
   A. The ε_i in the model once again represents random error--that is, random measurement error in Y (but not X's) and the idiosyncratic factors that affect the dependent variable.
      1. The observed Y scores are thus composed of the effects of the X' plus a random error. The random error is not observed independently; it is estimated from the residuals.
      2. Ideally, these errors really are random: they have an expected value of zero, a constant variance (their variation does not change with changes in X's), they are independent of the X's, and they are serially uncorrelated.
      3. We'll investigate the error component in more detail shortly.

VI. ANOTHER EXAMPLE:
   A. Here's another quick example
      1. We used these data in Class 7. The three variables are
         i. Out-of-wedlock births per 1,000 live births.
         ii. Average monthly Aid to Families With Dependent Children payment.
         iii. Percent of families living below poverty level.
         iv. The data pertain to a sample of 19 states
      2. Data are on the next page
      3. We would expect that, if Murray and other social welfare critics are correct, an increase in welfare spending would be associated with increases in out-of-wedlock births.

Births	AFDC	Poverty
221	110	14.8
204	145	14.9
179	358	7.7
241	227	8.9
231	133	13.2
225	277	8.4
122	271	7.4
138	233	9.8
156	341	7.6
161	379	8.2
176	217	9.1
134	207	6.3
160	185	14.0
92	277	9.8
147	318	7.7
203	107	13.1
133	109	11.1
191	214	9.2
138	366	6.3

B. Let's see whether both AFDC payments **and** the poverty rate can explain variation in the out-of-wedlock birth rate.
   1. This turns out to provide an interesting example of some phenomena we will be dealing with later.

C. The results from MINITAB follow.

```
The regression equation is
Outofwed = 125 - 0.036 AFDC + 5.57 Poverty

Predictor       Coef       Stdev    t-ratio        p
Constant      124.58       87.99       1.42    0.176
AFDC         -0.0356      0.1677      -0.21    0.835
Poverty        5.569       5.370       1.04    0.315

s = 39.63       R-sq = 19.4%     R-sq(adj) = 9.3%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    2        6039        3020      1.92    0.179
Error        16       25133        1571
Total        18       31173

SOURCE       DF      SEQ SS
AFDC          1        4350
Poverty       1        1689
```

D. Why are these results interesting?
   1. There are a relatively small number of cases: N = 19; degrees of freedom for error is 16.
   2. The percent of variation explained by both variables is roughly 19.
   3. But, the F-test for the model is 1.92, which is not "significant."
   4. Neither regression parameter is "significant." In fact, their "attained" probabilities are relatively high.
   5. It turns out that if we serially regressed the independent variables on births rate one at a time, we would discover that the "explained variation" is about the same.
   6. The problem we will see is (partly) "colinearity."

E. In the meantime, interpret the parameters on your own.
   1. Example: when the poverty rate is controlled, AFDC payments seems to have no (linear) relationship with out-of-wedlock births: a one dollar increase in AFDC payments is associated with 4 tenths of a percent **decrease** in out-of-wedlock births and this coefficient is not significant.
      i. But note that a $100 dollar increase in welfare "generosity" is associated with about a 4 percent decrease in births, which seems relatively important.
      ii. We'll investigate the situation further.

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

Go to Notes page

Go to Statistics page