DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816

TWO VARIABLE REGRESSION

I. AGENDA:
- A. Elements of the linear model
- B. Interpretation of regression parameters
- C. Causal inference from non-experimental research
- D. Least squares principle
- E. Reading: Agresti and Finlay ***Statistical Methods in the Social Sciences***, 3rd edition, Chapter 9.

II. GEOMETRY OF LINES:
- A. See the notes from the last class (Class 7)
- B. To understand the linear model let's review some simple math.
- C. The equation of a linear (straight line) relationship between two variables, Y and X, is

$$\boldsymbol{Y = a + bX}$$

- D. Interpretation:
  1. **a** is the intercept, that is the value of Y when X equals zero. If the line is graphed on an Y-X coordinate system (see below), then a is the point where the line crosses the Y axis.
  2. **b**, called the slope, is the amount of change in Y for a one-unit change in X. It's measured in units of the dependent variable, Y, but its numerical value depends on the measurement scale: if X is measured in dollars, then b will equal some particular value, but if the scale is thousands of dollars, b will have a different value.
  3. The figure presented in Class 7 shows a picture of the graph of a linear relationship. Notice that the graph is a straight line.
  4. The linear relationship described by this graph is:

$$\boldsymbol{Y = a + bX = 2 + (2)X}$$

- E. In other words, the intercept of this particular model is 2 and the slope is 2.0.
- F. The numbers a and b are called regression parameters; note that they are constants whereas X and Y are variables. The parameters show you how X affects or at least is connected to Y.

III. INTERPRETING THE REGRESSION MODEL:

A. The equation of a linear (straight line) relationship between two variables, Y and X, is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

B. Interpretation of parameters:

1. β_0 is the **regression constant** or intercept, that is the value of Y when X equals zero. If the line is graphed on an Y-X coordinate system, then β_0 is the point where the line crosses the Y axis.
2. β_1, called the slope or regression parameter, is the amount of change in Y for a one-unit change in X. As noted above, be thoughtful when looking at β_1
   i. Its numerical value depends on the measurement scale: if X is measured in dollars, then it will equal some particular value, but if the scale is thousands of dollars, it will have a different value.
3. Another way of viewing the model: how an individual's (unit's) score on Y is affected by the independent variable, X.
   i. The parameter β_1 is sometimes interpreted as a "causal" mechanism linking X to Y.
   ii. But see the next section.
   iii. A linear model, in brief, is a summary of what we think we know about the dependent variable.

C. Example:

1. Suppose the estimated or observed regression equation turns out to be:

$$\hat{Y}_i = \mathbf{10.1} + \mathbf{.03}X_i$$

   i. Here $\beta_0 = 10.1$.
      a) Sometimes the constant has no “real” or substantive meaning, as when for example we are relating achievement to age. (Age = 0 would be meaningless in most social science studies.)
   ii. The regression constant is $\beta_1 = .03$, which means that as X changes (increases) 1 unit (say, one year), Y increases .03 units of whatever Y is measured on, say an achievement index.
      a) This may or may not be a large change.
      b) You have to ask two questions at least:
         * What is the substantive meaning of a one-unit increase or decrease in X.
         * What is the substantive meaning of a β_1 unit change

in Y.

D. Mean value interpretation of Y:

1. The linear model is sometimes written (see Agresti and Finlay, ***Statistical Methods***, 3rd edition, page 314) as;

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

i. This equation suggests that the average or expected value of Y depends on a corresponding value of X. If β_1 is positive, for example, then the expected value of the dependent variable will increase with increases in X.

ii. This interpretation leads to the next topic.

IV. THE STATISTICAL MODEL:

A. Social and political relationships are seldom "determinate" which means that we have to add "error" to our conceptions of how one thing affects another. Also, we frequently deal with samples, not the total population, so we need to think about estimates versus parameters.

B. Sources of error:

1. Random fluctuations caused by hundreds of idiosyncratic factors, presumably which "cancel" each other out.
2. Random measurement error

C. The systematic part:

1. Suppose we have a quantitative dependent variable, Y, and a quantitative independent variable, X. In a previous example Y is "out-of-wedlock births" and X is the "average monthly AFDC payments." A statistical model describing the relationship is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

2. Interpretation:

i. The systematic part contains:

a) β_0, the intercept or constant which is the value of Y when X = 0

b) β_1, the slope or regression coefficient which shows how much Y changes for a one-unit change in X.

c) Suppose $\beta_1 = 0$? What does that mean?

D. Error part:

1. ε_i represents random error--that is, measurement error in Y (but hopefully not X), random factors causing variation in Y, etc. ε_i symbolizes the part of the variation in Y (e.g., illegitimacy) that is not explained by the model.

2. See Agresti and Finlay, ***Statistical Methods for Social Sciences***, 3rd edition pages 314 to 319.
3. An important goal of the social sciences is to reduce the magnitude of the ε_i's and to ensure that they are really random. Doing so has the effect of increasing the explanatory power of the model compared to the error component.

E. What we need is some method for finding numerical values of β , and β_1 , when the data are scattered about as in the example.
1. Before looking at how parameters are estimated, however, let's interpret regression parameters from another angle.

V. CAUSAL INFERENCE IN NON-EXPERIMENTAL RESEARCH:

A. It is often said that natural science differs from social inquiry because, among other things, investigators working in the former can literally manipulate variables to observe the effects on various phenomena. Hence, a chemist can administer varying amounts of a compound to rats to see what effect it has on, say, the number of lymphocytes.

B. Moreover, so the conventional wisdom continues, the laboratory scientist can hold all relevant factors constant, so that if there is a change in cell counts, the difference can unambiguously be attributed to the compound. The researcher, it is believed, can make a reasonably valid **causal inference**. The inference about causality derives its strength from the experimenter's ability to eliminate alternative explanations for any observed changes.

C. Now compare this situation with that facing the social scientist who wants to know if changes in AFDC payments affect "deviant" or undesirable behavior. It is possible, as we have already demonstrated, to compare areas having differing payment levels. Or, as we just did, we can examine the association between variation in one variable (AFDC payments) and out-of-wedlock births.

D. The problem comes in interpreting the results. Since we are dealing with "observational" data--we have not manipulated anything nor have we control for possible alternative causal factors, it is difficult to interpret our results, especially the regression coefficient, as a "causal" parameter.
1. Why? Suppose, for the moment, our data had confirmed Murray's argument: states with the highest welfare benefits had the highest proportion of out-of-wedlock births. (This is contrary to what we did find, but let's suspend our knowledge for a moment.) But consider this possibility: those states having low AFDC payments also happen to be populated by groups with strong and extended families and consequently illegitimacy violates well established social norms. Suppose, in addition, those places with more generous benefits do not contain as many such groups. There are, in other words, three relationships: one between the dependent variable (births) and AFDC payments; another between births and family structure; and a third between the two independent variables,

AFDC payments and family structure. The question then arises: are the differences in illegitimacy due to a) AFDC payments; b) family structureand social norms; or c) both.

2. Figure 2 suggests alternative models.
3. "Hard scientists" would try to answer the question by manipulating variables. (They would move families **at random** to different states, thus cancelling out the association between welfare payments and family structure.) In a sense they would be comparing apples with apples: the states being compared would be the same in all relevant respects except for AFDC payment level. If their illegitimacy rates differed, they investigators could attribute the differences to the main independent variable.

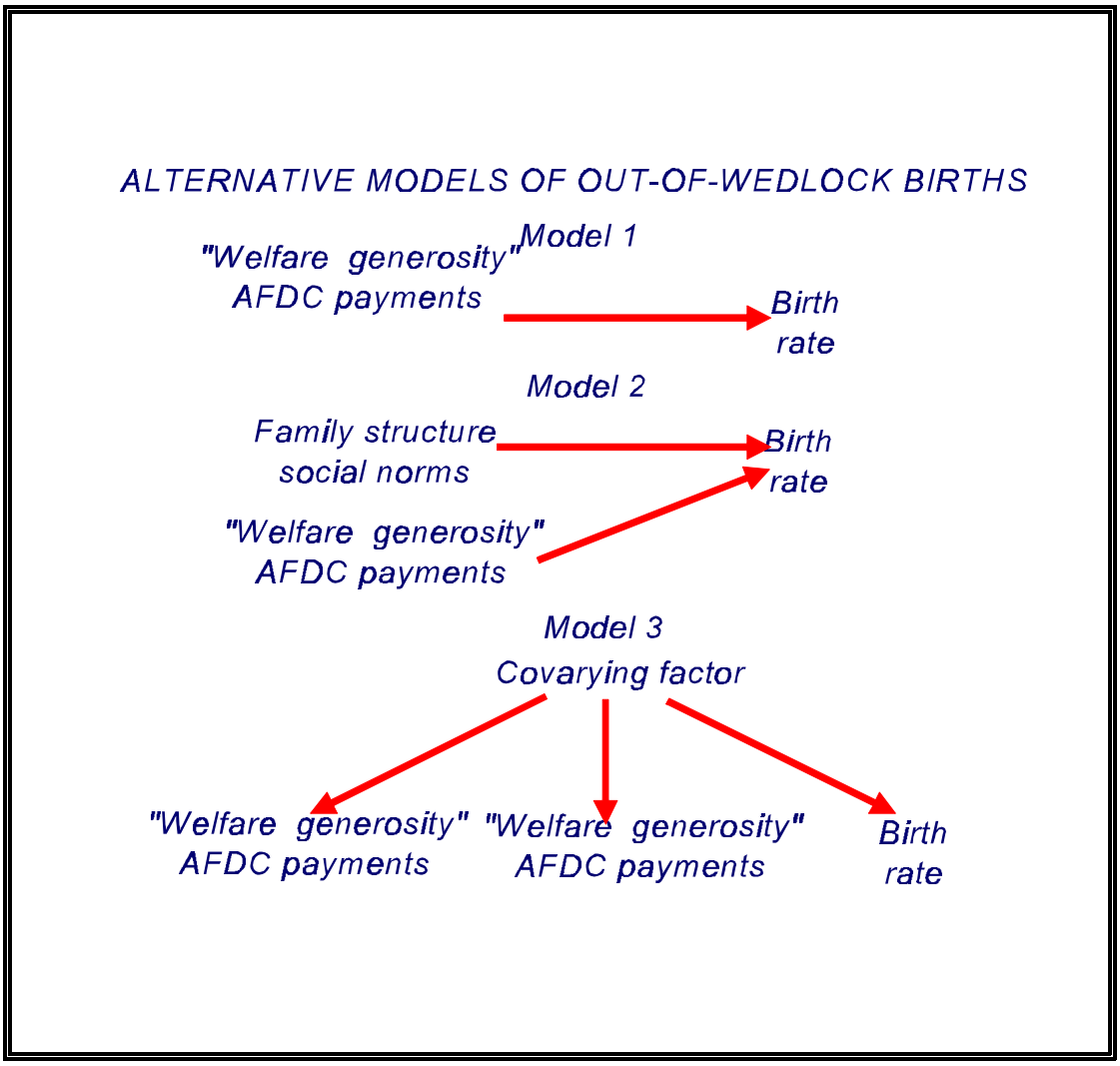

Figure 2: Alternative Causal Models

4. But, of course, in the real world such manipulations are not possible; families cannot be moved around to test hypotheses. (Actually social scientists and policy analysts have attempted to experiment on welfare recipients.)
5. The only solution is to adjust whatever statistical measure of relation between Y and X, β_1 for example, for the effects of other factors.
6. These considerations lead to two conclusions:
   i. We have to be careful about translating statistical relationships, as measured by the betas, into causal assertions of the form "X causes (variation) in Y."
   ii. We need methods to adjust the statistical measures, the β's, to take into account at least some possible confounding influences.

E. This is a matter we will deal with in the remainder of the course.

VI. LEAST SQUARES PRINCIPLE:

A. Suppose we have two estimates of β_0 and β_1; for now it doesn't matter where they came from. As example, suppose the estimates for an equation are 10.1 for β_0 and .03 for β_1. With these numbers we can obtain an <u>estimated</u> model (note the hats):

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X$$

where $\hat{Y}_i$ is the <u>predicted</u> value of Y, and $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated values of the parameters. For example,

$$\hat{Y}_i = 10.1 + .03X$$

1. Here, if the X is 0, the estimated or predicted value of Y is

$$\hat{y}_i = 10.1 + (0) = 10.1$$

2. If X is, say, 250, then the predicted value is

$$\hat{Y}_i = 10.1 + .03(250) = 17.6$$

B. <u>Residuals</u>: A residual is the difference between a predicted value (predicted on the basis of some model) and the corresponding observed value.

1. The formula is:

$$\hat{\varepsilon}_i = (\hat{Y}_i - Y_i)$$

2. Suppose, to continue with the above case, a case had X = 0--in which case we would predict its value on Y to be 10.1 (see above)--but in fact its actual or observed illegitimacy rate is 20. Then the error or residual for this county is 212 - 20 = 9.9.
3. A geometrical interpretation of residuals is shown in Figure 2.

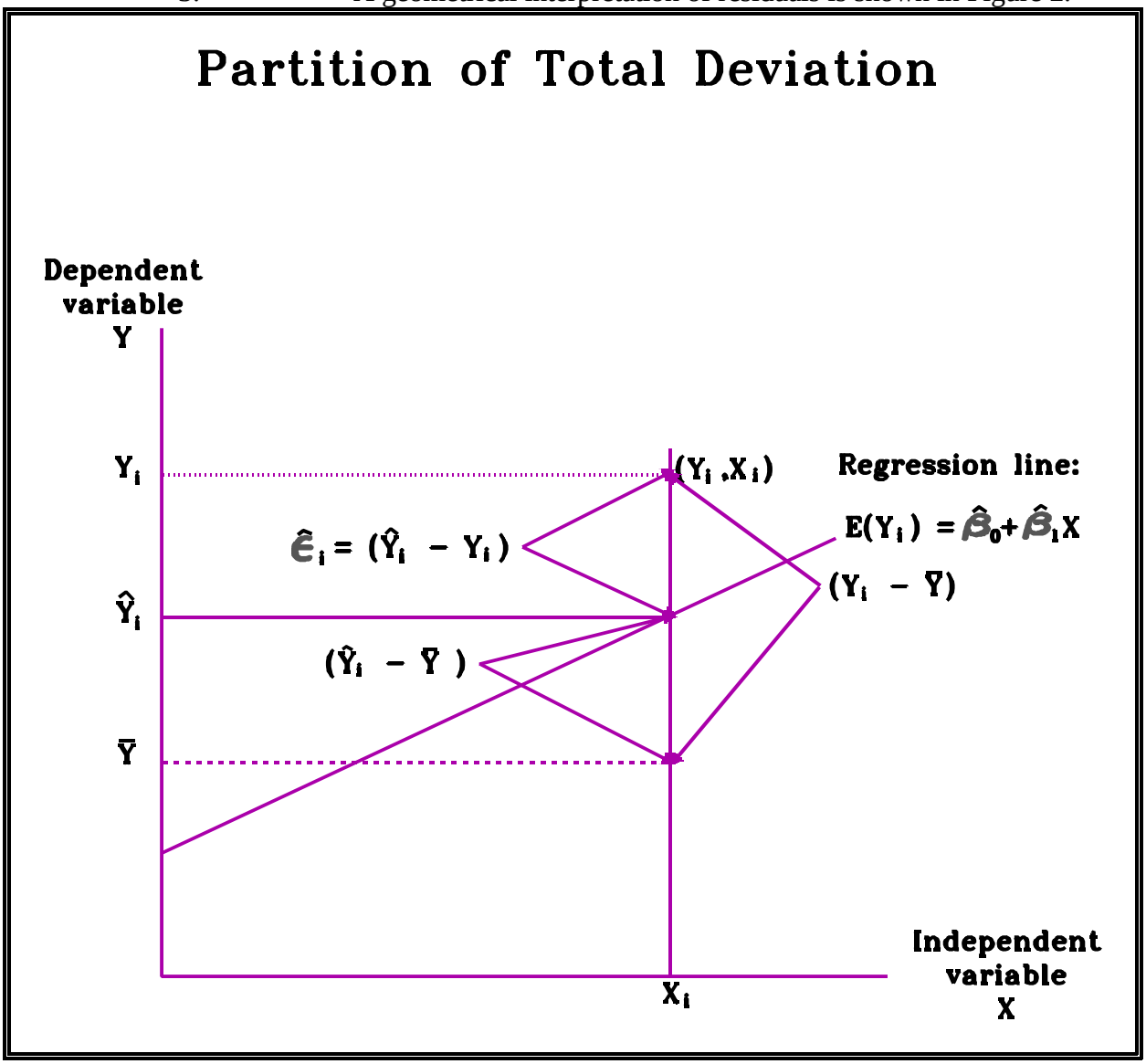

Figure 2: Partition of Deviation

4. Interpretation:

i. ($Y_i - \overline{Y}$) is the difference between the ith unit's score on Y and the grand (overall) mean. This difference when combined with all the other corresponding differences measures the total variation in Y.

ii. When all of these differences are combined by first squaring and then summing them the result is the **total sum of squares** (TSS), an important measure of variation in Y. The formula is:

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

iii. ($\hat{Y}_i - \overline{Y}$) is the difference between the predicted Y and the grand mean. It, in a sense, represents how much we know about Y given our knowledge of X. In other words, if we knew nothing we would "predict" that a typical unit would have a score equal to the grand mean. But with our model of X's impact on Y, we know more than this; in fact we know that as X increases one unit (one dollar in this example) the value of Y will increase .03 units. Thus, a portion of the total variation in Y is "explained" by our knowledge of X which is summarized mathematically in the equation:

$$\hat{Y}_i = 10.1 + .03X.$$

iv. Finally, $\hat{\varepsilon}_i$ represents error in prediction. It is, stated in other words, the difference between what we think Y should be and what it actually is. This error together with all of the others represents the portion of variation in Y that is not accounted for by X.

5. The Least Squares Principle:

i. We pick as estimators of β_0 and β_1 those particular values that minimize the sum of squared residuals for a batch of N observations under study. That is, thinking of β_0 and β_1 as population parameters, we choose estimates of them in such a way that the quantity is a **minimum**.

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

6. Keep S^2 in mind because it comes up again and again.
7. The principle of least squares leads to **computing formulas** used to obtain estimates of the parameters from a set of data. These formulas are describe by Agresti and Finlay and will be discussed later. For now we will rely on MINITAB to compute the numerical estimates.

VII. NEXT TIME:
- A. Examples of MINITAB regression
- B. Measures of fit
- C. Tests of significance

Go to Notes page

Go to Statistics page