DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816

SIMPLE TWO VARIABLE REGRESSION

I. AGENDA:
- A. Causal inference and non-experimental research
- B. Least squares principle
- C. Regression examples:
  1. "Stories"
  2. Strategies for investigating models
  3. Plots
  4. Interpretation of results
- D. Reading: Agresti and Finlay ***Statistical Methods in the Social Sciences***, 3rd edition, Chapter 9.

II. CAUSAL INFERENCE IN NON-EXPERIMENTAL RESEARCH:
- A. Reprinted from Class 8 notes.
- B. 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.
- C. 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.
- D. Now compare this situation with that facing the social scientist who wants to know if changes in AFDC payments affect "socially 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.
- E. The problem comes in interpreting the results. Since we are dealing with "observational" data--we have not manipulated anything nor have we controlled 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 1 (next page) 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 canceling 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.

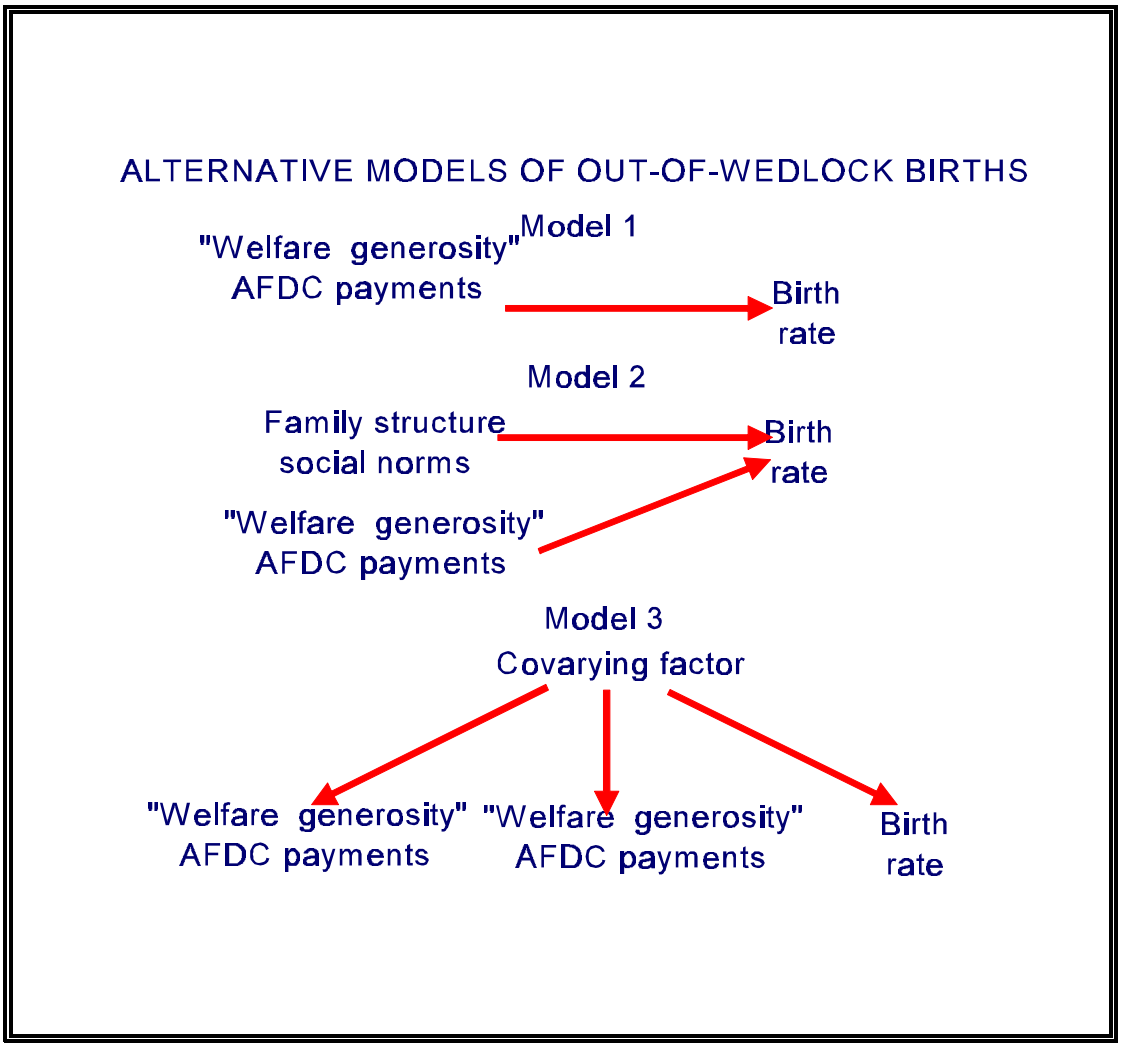

Figure 1: Alternative Causal Models

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.

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

III. LEAST SQUARES PRINCIPLE:

A. Based on the last set of notes.

B. 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 estimated model (note the hats):

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

where $\hat{Y}_i$ is the predicted value of Y, and $\hat{\beta}_0 = \mathbf{10.1}$ and $\hat{\beta}_1 = \mathbf{.03}$ are the estimated values of the parameters. That is,

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

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

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

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

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

C. Residuals: 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 rate is 20. Then the error or residual for this case is 10.1 - 20 = -9.9.

D. A geometrical interpretation.

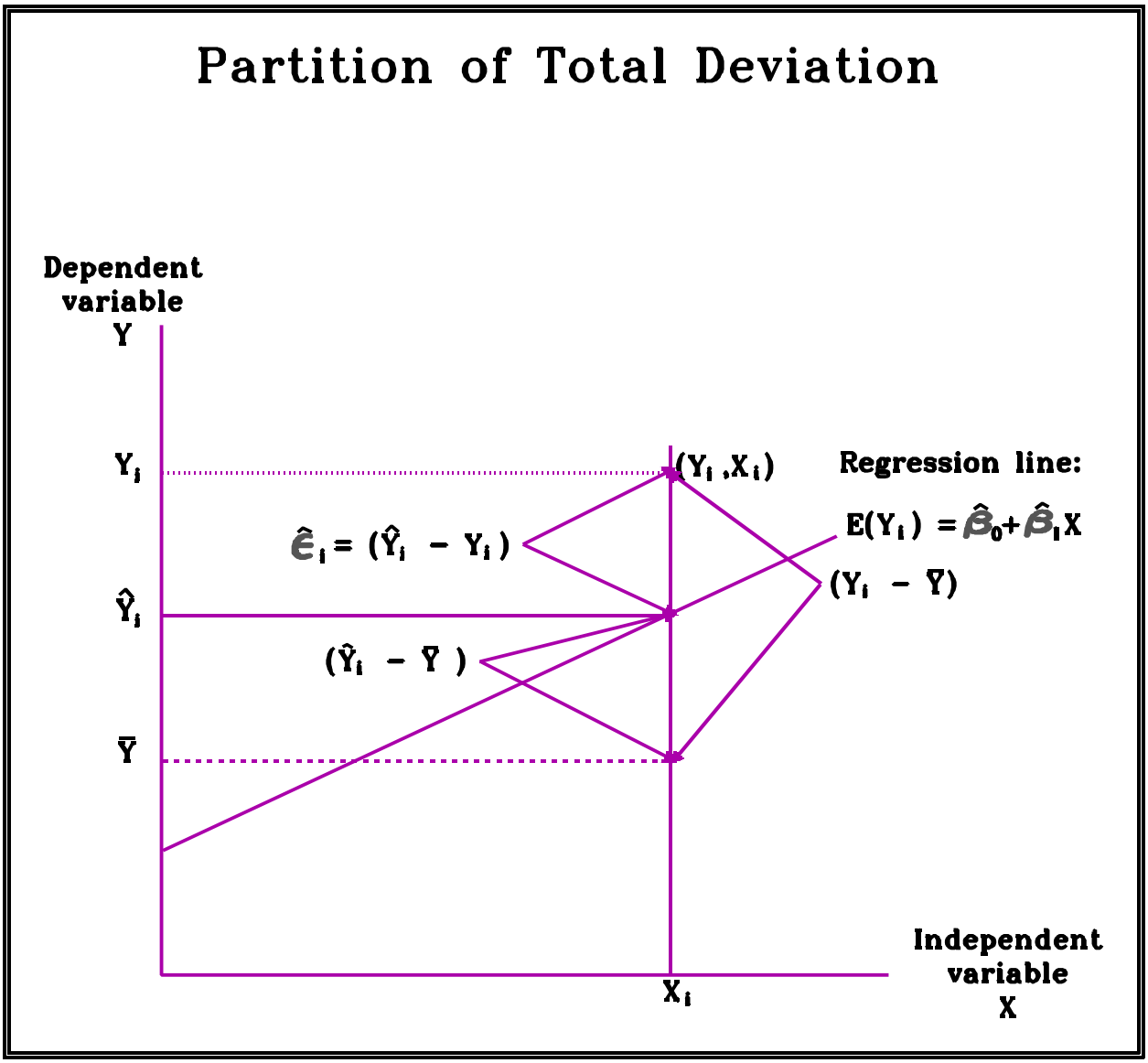

Figure 2: Partition of Deviation

1. There are three kinds of differences in Figure 2.

   i. $(Y_i - \bar{Y})$ 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.

      a) 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:

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

ii. ($\hat{Y}_i$ - $\overline{Y}$) 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:

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

.

a) "RegSS" means "regression sum of squares" or the part of the total sum of squares that can be attributed to the regression model and hence is **explained** (by X).

iii. Finally, $\hat{\varepsilon}_i = (\hat{Y}_i - Y_i)$ 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.

a) Note that we can also sum the estimated errors or sum the observed values of Y minus the predicted values.

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

b) S^2 is often called the **residual** or **error** or "**unexplained**" (by X) sum of squares.

c) This term represents the portion of the total variation in Y that cannot be attributed to X.

E. Note this important relationship.

1. We will refer to it often

$$TSS = RegSS + ResSS$$

$$= Explained\ SS + Unexplained\ SS$$

2. That is, the total variation in Y can be partitioned or divided into two components, an explained and an unexplained part.
3. It is natural to determine what proportion or percent of the total is explained by X. We will, in fact, do so often with the measure R^2.

F. The Least Squares Principle:

1. We pick as estimators of β_0 and β_1 those particular values that minimize the sum of squared residuals (S^2) 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**.
2. 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.

IV. REGRESSION IN PRACTICE - SOME EXAMPLES:

A. The data for these examples comes from the "Data Story Library" at Carneige Mellon University.

1. We can use them to illustrate regression analysis with MINITAB, the interpretation of parameter estimates, possible adjustments that will have to be made, and similar topics.

B. Age and height story:

1. Quoted from the Library:
   i. Description: Mean heights of a group of children in Kalama, an Egyptian village that is the site of a study of nutrition in developing countries. The data were obtained by measuring the heights of all 161 children in the village each month over several years.
   ii. Number of cases: 12
   iii. Variable Names: (c1) Age in months; (c2) Mean height in centimeters for children at this age.
2. Here is a plot.
   i. Notice that the relationship appears to be linear.

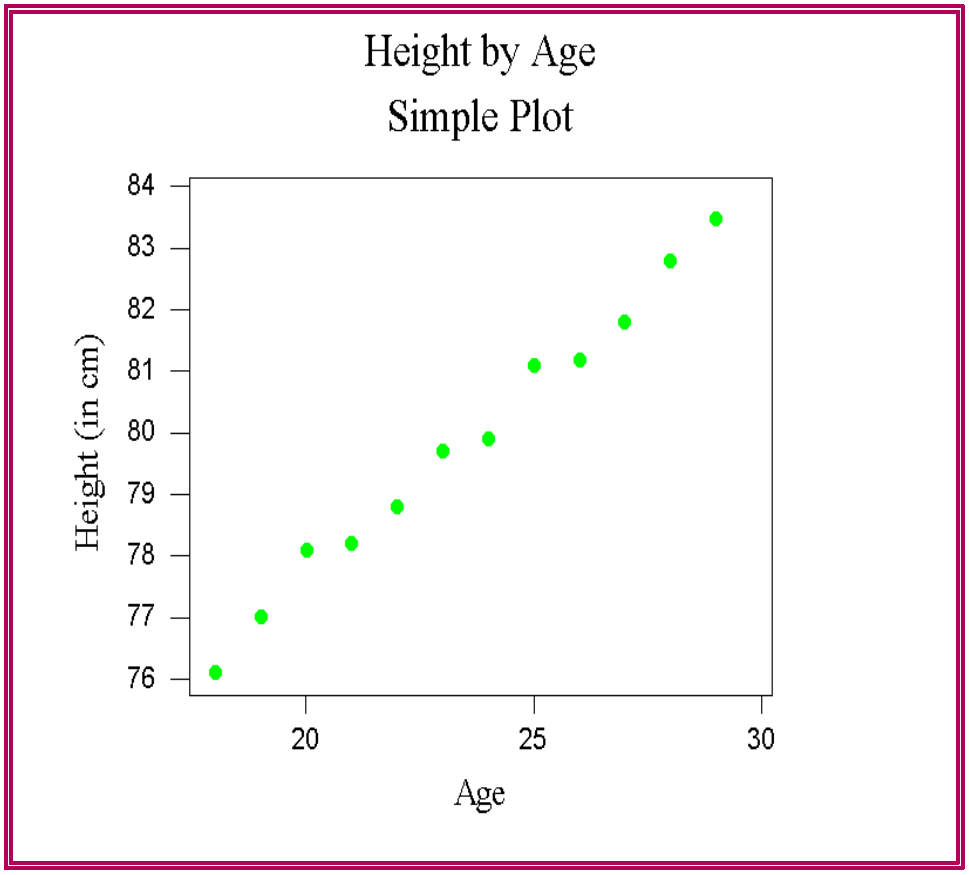

Figure 3: Simple Plot

ii. Now here the results from the regression analysis

```
The regression equation is
C2 = 64.9 + 0.635 C1

Predictor        Coef       StDev          T        P
Constant      64.9283      0.5084     127.71    0.000
C1            0.63497     0.02140      29.66    0.000

S = 0.2560      R-Sq = 98.9%     R-Sq(adj) = 98.8%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      57.655      57.655    879.99    0.000
Residual Error    10       0.655       0.066
Total             11      58.310
```

iii. For now we will ignore most of these results. But notice that
- a) the estimated constant is 64.96 centimeters. (Does this make "substantive" sense?)
- b) the estimated regression coefficient is .635, which means that an extra month translates (on average) into an additional .635 centimeters of height.
  - * How much would a boy be expected to grow after 3 months.
- c) The R^2 means that about 98.9 percent of the variation in height is "explained" by age. We, of course, now know that this refers to statistical explanation. These data do not tell us why children grow to certain heights; only that as they get older, they grow.
- d) The "T" and "p" are used to test the hypothesis that in the population from which the data were drawn, the regression parameters equal zero.
  - * That is, the hypothesis pertaining to the regression coefficient is H_0: $\beta_1 = 0$.
    - 1) The t-value (29.66) and its attained probability (.000) suggest that this hypothesis is not tenable given the data.

C. Alcohol and tobacco story

1. From the Library:

i. Abstract: Data from a British government survey of household spending may be used to examine the relationship between household spending on tobacco products and alcoholic beverages. A plot of spending on alcohol vs. spending on tobacco in the 11 regions of Great Britain shows an overall positive linear relationship with Northern Ireland as an outlier.

ii. These data illustrate the effect of a single influential observation on regression results. In a simple regression of alcohol spending on tobacco spending, tobacco spending does not appear to be a significant predictor of tobacco spending. However, including a dummy variable that takes the value 1 for Northern Ireland and 0 for all other regions results in significant coefficients for both tobacco spending and the dummy variable, and a high R-squared.

iii. Variables:
- a) (c1) Average weekly household spending on alcoholic beverages in pounds
- b) (c2) Average weekly household spending on tobacco products in pounds

2. Here is a plot of all 11 regions:

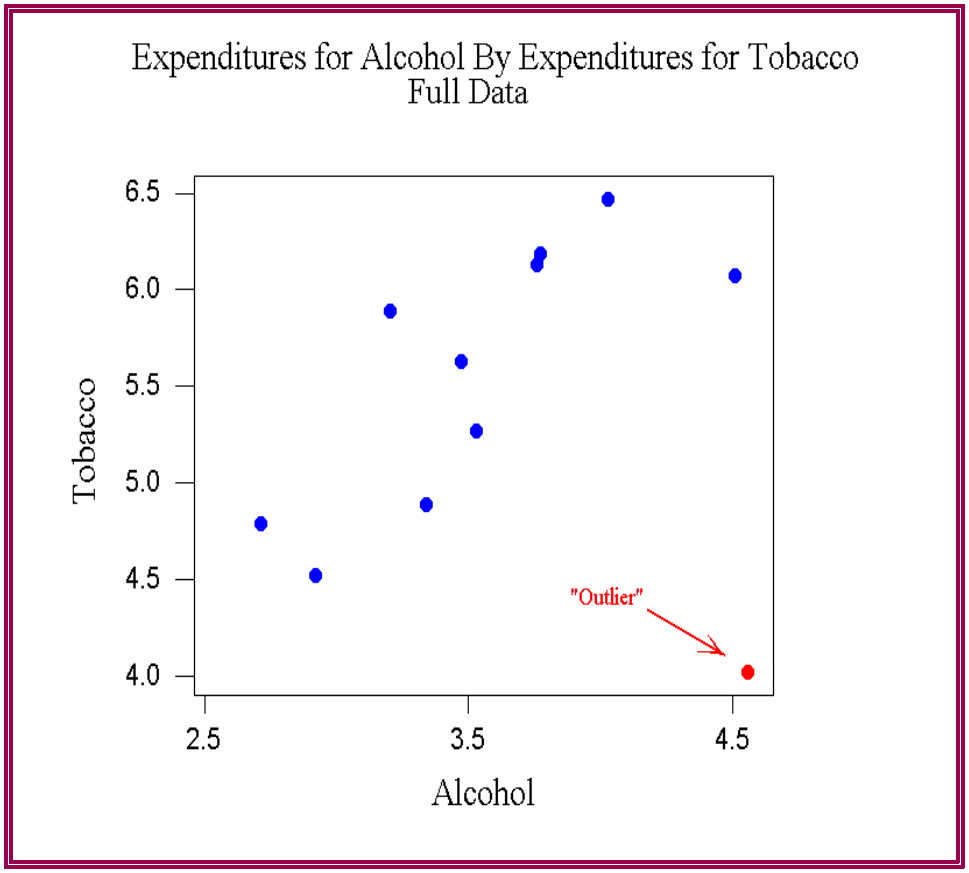

Figure 4: Alcohol and Tobacco Expenditures

a) Except for the point denoted "outlier" most of the observations lie on a straight line.

i. Here are the results of the regression analysis

```
The regression equation is
Alcohol = 4.35 + 0.302 Tobacco

Predictor        Coef       StDev          T        P
Constant        4.351       1.607       2.71    0.024
Tobacco        0.3019      0.4388       0.69    0.509

S = 0.8196      R-Sq = 5.0%      R-Sq(adj) = 0.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      0.3181      0.3181      0.47    0.509
Residual Error     9      6.0461      0.6718
Total             10      6.3643
```

3. The estimated regression coefficient, .302, means that for every one pound increase in spending on tobacco, there is a .302 pound increase in spending for alcohol.
   i. What this tells us is that expenditures for one are positively (but weakly) related to spending for the other.
4. But the R^2 = .05 or 5 percent suggests that a linear model does not fit the data very well.
   i. Moreover, the observed t and attained probability suggest that if these data constituted a sample from some population, there would be little evidence that β_1 is not zero. (Value of zero indicates no linear relationship.)
5. But we have observed that one value lies off the "beaten path" or is an outlier. What happens when it is removed; that is, treated as a missing value.
   i. Here's the regression analysis

```
Reduced Data

The regression equation is
ReducedA = 2.04 + 1.01 ReducedT

10 cases used 1 cases contain missing values

Predictor        Coef       StDev          T        P
Constant        2.041       1.001       2.04    0.076
ReducedT       1.0059      0.2813       3.58    0.007

S = 0.4460      R-Sq = 61.5%     R-Sq(adj) = 56.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      2.5434      2.5434     12.78    0.007
Residual Error     8      1.5915      0.1989
Total              9      4.1348
```

ii. We now see that with the one case removed the estimated coefficient is 1.001, more than 3 times larger.
   a) It's also "statistically significant."
   b) Moreover the R^2 is now .615.

6. Our conclusion is that spending for the two products are closely related: a pound for a pound.

D. Attitudes toward government and public officials story.

1. These data come from a different study, namely the "1991 Race and Politics Survey" and are available at the Data Analysis and Documentation site we went to for assignment 2.
2. The variables are;
   i. "Thermometer 1" President Bush 0-10
      a) These next questions are about some of the political leaders and groups that are in the news these days. I'll read a name and ask you to rate that person or group on a thermometer that runs from zero (0) to ten (10). The higher the number, the warmer or more favorable you feel toward that person or group. The lower the number, the colder or less favorable you feel. If you feel neither warm nor cold toward them, rate that person or group a five.
      b) The first person is President Bush. How would you rate him on a scale from 0 to 10?
   ii. "Thermometer 2" Jesse Jackson 0-10
      a) How about Jesse Jackson? (How would you rate him on a scale from 0 to 10?)

iii. "Thermometer 3" The Federal Government
  a) (How about) The Federal Government? (How would you rate it on a scale from 0 to 10?)
iv. Age of respondent in years
v. Education level
  a) What is the highest grade or year of school you completed?
    1 Eighth grade or lower
    2 Some high school
    3 High school graduate (or GED)
    4 Some college
    5 College graduate
    6 Some graduate work or graduate degree
    8 DK
    9 RF/MD
vi. Race: What race or ethnic group do you consider yourself?
    1 Black, African-American, Negro
    2 Native American, Alaskan native
    3 Latino, Mexican-American, Hispanic,
    4 Filipino
    5 Asian, Pacific Islander
    6 White, Caucasian
    7 Other (SPECIFY) (See Appendix E)
    8 VOLUNTEERED: Jewish

3. We'll have to manipulate these data in order to use the them in the analysis.
  i. In particular we will treat the thermometer questions as (quantitative) dependent variables that we want to "explain" by education and age.
    a) Later we'll add race.
  ii. We need to change the "missing data" codes to the one used in MINITAB, "*."
  iii. We'll use just two categories of ethnicity.
  iv. Note that the data file contains more than 2,000 cases, which presents both opportunities and a few problems.
  v. Note finally, that to be correct technically, the data should be "weighted." But that's a nicety we will skip.

V. NEXT TIME:
  A. Measures of fit
  B. Tests of significance

Go to Notes page

Go to Statistics page