DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816

MULTIPLE REGRESSION ASSUMPTIONS AND "DIAGNOSIS" METHODS

I. AGENDA:
- A. Assumptions
- B. Examining the models and looking for improvements
- C. Reading: Agresti and Finlay ***Statistical Methods in the Social Sciences***, 3rd edition:
  1. Chapters 11 and 12 for assumptions.
  2. Pages 534 to 541 for diagnostic techniques

II. ASSUMPTIONS:
- A. The linear model underlying regression analysis is:

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

- B. OLS is used to obtain estimates of the parameters and to test hypotheses.
- C. In order for the estimation and inference procedures to be "valid" certain conditions have to be met.
  1. No "specification" error:
     - i. The relationship between X and Y is linear.
     - ii. No relevant independent variable has been excluded. If one has been omitted, its effects will "show up" in the error term.
     - iii. No irrelevant independent has been included. If so, error variance will be too large.
  2. No measurement error.
     - i. X and Y are accurately measured. Measurement error in Y will "inflate" the error variance if it is random.
  3. Assumptions about the error term, ε_i
     - i. Expected value of the errors is zero:

$$E(\hat{\varepsilon}_i) = 0$$

ii. Constant variation in errors across values of X (**homoscedasticity**).
   1) Example: the variation in errors among in countries with low GNPs (or GDPs) should be the same as in those having higher levels.

iii. The independent variable, X, is not correlated with the error term. This is important! We will discuss this assumption in more detail.

iv. There is no **autocorrelation**; that is, the error at time t = 1 is not related to the error at time t = 2. Again, we will come back to this point.

v. In order to test for significance we also assume the errors are normally distributed:

$$\varepsilon_i \sim N(0,\sigma^2)$$

4. "Colinearity": the correlations among the independent variables are not too large.
   i. If one independent variable is highly correlated with another, they essentially provide the same information (if any) about the dependent variable.
   ii. If two independent variables are perfectly correlated, the estimation procedures break down and we cannot obtain estimators.
   iii. If there are large inter-correlations, say, greater than .6 or .7, then the estimators may be unstable--vary greatly from sample to sample--and may be hard to interpret.

D. Later we will see how to investigate these assumptions using "simulation" procedures.

III. SOME TECHNIQUES AND TOOLS:

A. We can use various techniques to explore how well the observed models "agrees with" the assumptions and to look for improvement in fits.

B. In order to evaluate a model's fit and attempt to assess the reasonableness of assumptions we'll do these things:
   1. Plot Y versus each independent variable to look for
      i. Departures from a linear
      ii. Outliers and "leverage" ("influential") points
      iii. Possible transformation of Y and/or X
   2. Plots of residuals
      i. Stem-and-leaf displays to check for normality and large residuals
      ii. Residuals versus fitted or predicted values of Y
      iii. Residuals versus each X.
   3. Partial regression plots:

i. In effect we will regress Y on all the independent variables except one and obtain "adjusted residuals," $\hat{\varepsilon}_{YX_{[k]}}$. This symbol means the residual--the difference between observed and predicted values of Y--using the model that contains all the X's except X_k. I also use the symbol $Y_{[k]}$ or something similar.

ii. We'll then regress X_k on the other X's and obtain the residual, which is denoted $\hat{\varepsilon}_{X_{[k]}X}$.

iii. We will then plot and adjust the first set of residuals--the ones denoted above as $\hat{\varepsilon}_{YX_{[k]}}$. against the second set, denoted $\hat{\varepsilon}_{X_{[k]}X}$.

iv. The resulting plot is called a partial regression plot.

4. The correlation matrix of the independent variables to look for large inter-correlations among them or multicolinearity.

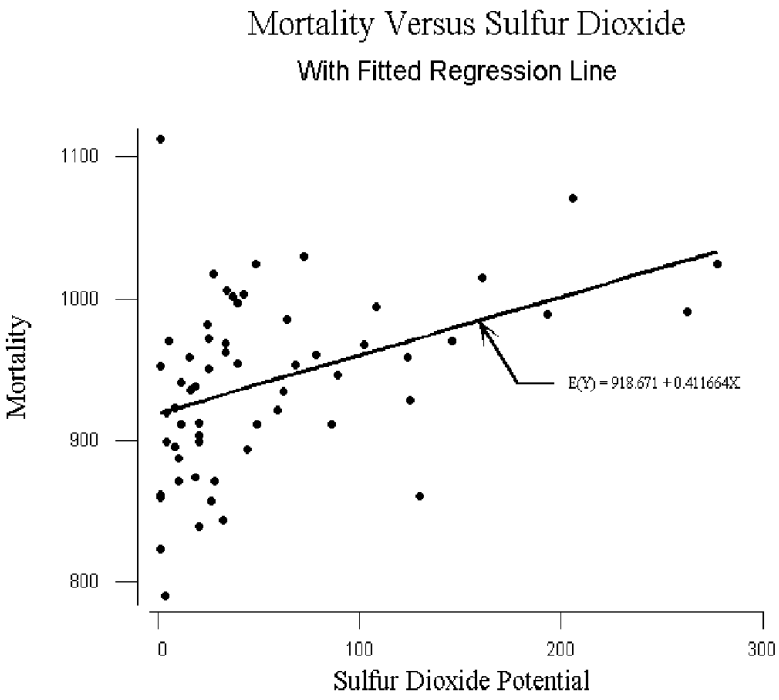
IV. EXAMPLE:

A. Let's use the data pertaining to air quality and mortality that were described in Class 13.

1. The data are for approximately 60 American cities
2. For now we will use just three of the variables:
   i. Y: age-adjusted mortality per 100,000 population.
   ii. X_1: sulfur dioxide potential
   iii. X_2: Median years of education

B. The plots of mortality against X_1 and X_2 provide a great deal of information, as we have seen before.

1. The first plot (next page) displays mortality versus sulfur dioxide with the estimated least square line added.
   i. It suggests that there is strong relationship between mortality and pollution, but that the relationship might be slightly non-linear, a point we'll take up later.
   ii. We also see a couple of points that lie considerably above and below the estimated line. We can explore those later.
   iii. Perhaps because of those points and because most of the data lie below 120, it appears that the variation among the residuals narrows as X increases. Again, we'll keep an eye on it.
2. A plot of mortality on the other X gives similar kinds of information.



C. Now lets obtain the multiple relation of mortality on X_1 and X_2.

1. Here are the results. Since we've been over this material I only summarize them.

```
The regression equation is
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 = .355
```

i. Note that I have changed the R^2 to a proportion.

ii. Note the partial regression coefficient of mortality on education

controlling for sulfur dioxide, $\hat{\beta}_{YX_2|X_1} = -.31.9.$ We'll come across it in a minute.

iii. Similarly note the partial regression coefficient of Y on X_1 controlling for education.

2. The residuals and fitted values for this model were stored so we can plot them in various ways to check assumptions about the error term and the model specification.

D. Residuals:

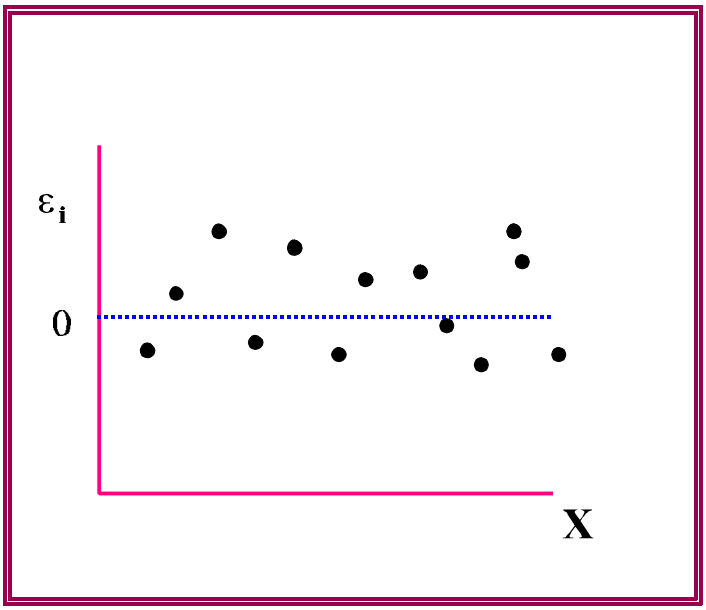
1. A simple stem-and-leaf display indicates that the residuals are normally distributed.

i. There are as we might have guessed from looking at the bivariate plots a couple of "deviant" cases. But they seem more or less symmetrically distributed above and below 0.

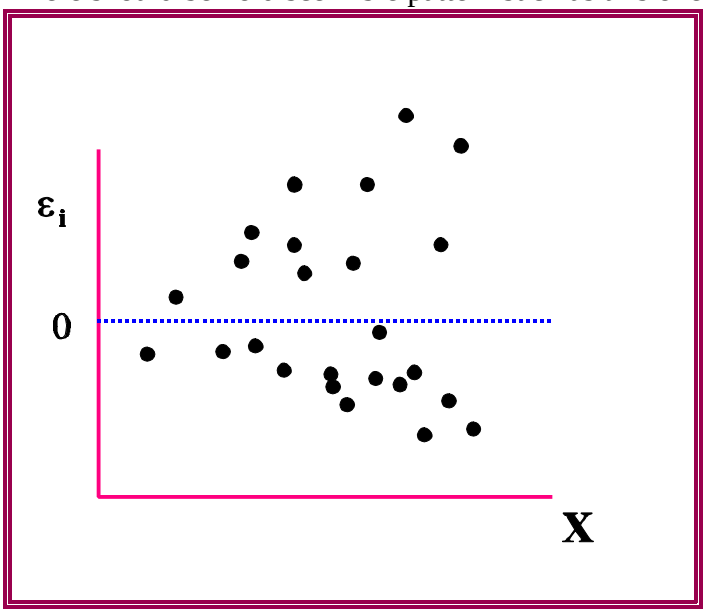
```
Residuals for mortality on sulfur and education


   1   -1 3
   1   -1
   3   -0 99
   5   -0 66
  10   -0 55544
  24   -0 33333222222222
  (6)  -0 111000
  29    0 000001
  23    0 222222333
  14    0 4444555555
   4    0 67
   2    0 8
   1    1
   1    1
   1    1 4
```

2. Plots of residuals: if the model has been correctly specified--errors not correlated with any independent variables, constant variance of errors, and so forth, a plot of residuals against X's or fitted values should look something like this.

i. Also see Agresti and Finlay, ***Statistical Methods***, 3rd edition for other examples.



3. There should be no discernible pattern such as this one



   i. This suggest non-constant error variation.

4. Standardized residuals:

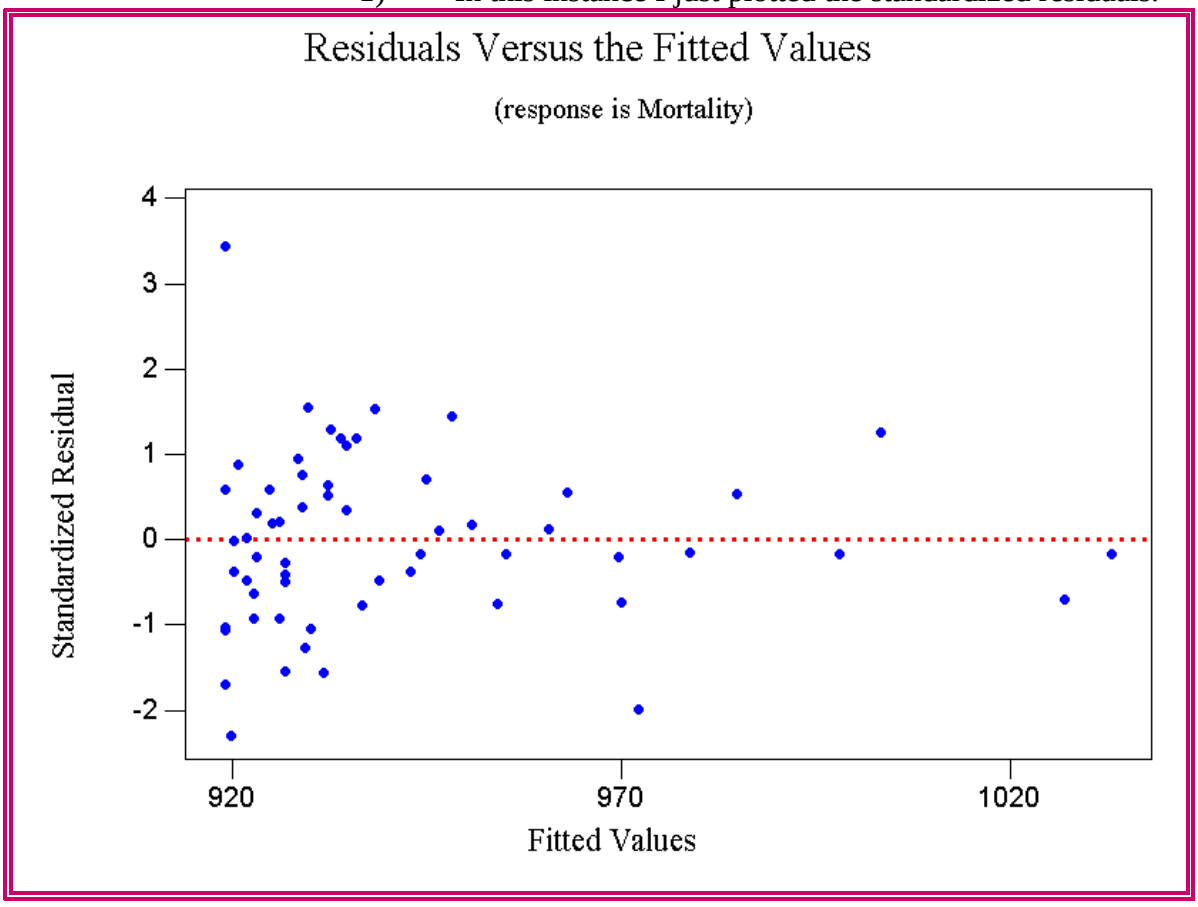
   i. Instead of plotting raw residuals, which are defined as $(\hat{Y}_i - Y_i)$ it is sometimes convenient, if not better, to plot standardized residuals, which are residuals divided by the standard error about

the regression line, s in my notation and S in MINITAB:

$$\frac{(\hat{Y}_i - Y_i)}{\sqrt{MSRes}} = \frac{(\hat{Y}_i - Y_i)}{s}$$

1) Remember that MSRes is the mean square residual and its square root is the standard error (deviation), s, about the regression line.

5. Residuals versus fitted values.
   i. It's also common to plot residuals versus fitted values.
   ii. Most regression programs allow you to store predicted (fitted) values and residuals.
      1) MINITAB labels them "FIT" with a number and "RES" with a number.
   iii. An example follows.
      1) In this instance I just plotted the standardized residuals.



iv. Note that the residuals cluster around 0.

v. Note also that the error variation may not be constant; it seems to shrink as X increases.

1) We can transform one or more of the variable to see if doing so improves the fit, a point we will discuss later.

6. It is also common to plot residuals versus each independent variable.

i. We can demonstrate this in class if time permits.

E. Partial residual plots.

1. Suppose we are interested in the effects of a particular variable, say X_k.
2. One approach to studying its influence on Y is to regress Y on all the independent variables except this one and obtain residuals. The residuals represent "what's left" after the other independent variables have "done their work."

i. For now call these partial or adjusted residuals.

3. A **partial residual** plot is a plot of these residuals against each independent variable.
4. One can also regress the independent variable of interest against the other independent variables and obtain residuals. These are **partial independent variable residuals**.
5. A **partial regression** plot is plot of the partial residuals against the partial independent variable residuals.
6. One can in fact follow this procedure for all the independent variables.

F. Here is the idea stated a bit more formally.

1. Note that I am using slightly different notation than that present previously so as to make clear what is being regressed on whay.

i. Denote residuals obtained when Y is regressed on all X's except X_k as

$$\boldsymbol{Y}_{\mathbf{1,2...(}k\ -\ \mathbf{1)...}K} = \boldsymbol{Y}_{\mathbf{[.}k\mathbf{]}}$$

ii. This is meant to correspond to the $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{YX}_{\mathbf{[}k\mathbf{]}}}$. presented above.

1) The idea is that we are isolating independent variable k to see what its partial effects are.

2. Next, denote the residuals obtained from regressing X_k on all of the other X's as $\boldsymbol{X}_{\mathbf{[.}k\mathbf{]}}$
3. The plot of $Y_{[.k]}$ versus $X_{[.k]}$ is a **partial regression plot** and the regression of $Y_{[.k]}$ on $X_{[.k]}$ has these properties:

i. The OLS estimator of the regression coefficient $\hat{\boldsymbol{\beta}}_{\boldsymbol{Y}_{\mathbf{[}X\mathbf{]}}\boldsymbol{X}_{\mathbf{[.}k\mathbf{]}}}$ is the same as the one obtained from regressing Y on all of the X's; that is, it is the same as $\hat{\boldsymbol{\beta}}_{\boldsymbol{YX}_k|\boldsymbol{X}_1\ldots}$ obtained from the full model estimate.

ii. The constant of this regression will be 0.

iii. Residuals from this partial regression will be the usual residuals obtained from regressing Y on all of the X's.

iv. The influence of individual data values on the beta's are easier to see.

v. Moreover, complicating factors such as leverage values, colinearities, and heteroscedasticity are easier to see.

4. Here's the example from the air quality example.

i. Recall that Y is mortality, X_1 sulfur dioxide, X_2 is median education.

ii. Hence there are K = 2 independent variables.

iii. I first regressed Y on sulfur to let it "explain' all of the variation in mortality that it could.

1) That is, I am interested in isolating the controlled effect of X_2

2) I obtained and stored the residuals, $\hat{\boldsymbol{\varepsilon}}_{YX_{[2]}} = \boldsymbol{Y}_{[2]}$

iv. Next, I regressed X_2 on X_1 and obtained residuals $\boldsymbol{X}_{[.k]}$

1) These residuals are what's left of X_2 after X_1 has explained all of the variation in it that it can.

v. In effect, both Y and X_2 have been adjusted for the effects of sulfur.

vi. If there were more independent variables in the model, I would include them in these regression procedures.

vii. The point is purify Y and education of the effects of sulfur and any other factors and then see if and how they are related.

viii. The partial regression plot is on the next page.

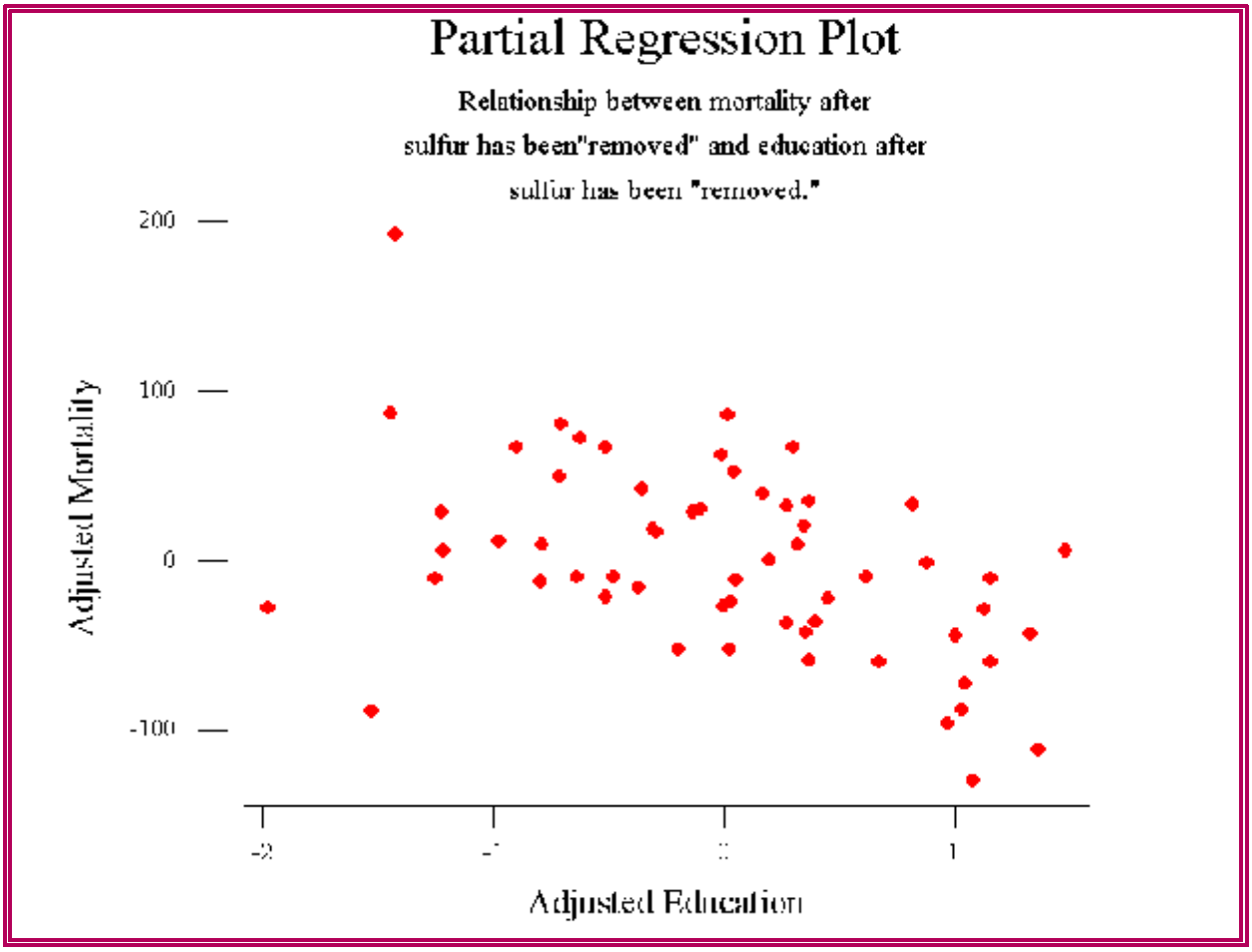
1) We can see that after sulfur dioxide has been taken into account (in both Y and X), there remains a substantial negative correlation.

2) The slope of the line that fits these data is

```
The regression equation is
RESI3 = - 0.00 - 31.9 RESI5

59 cases used 1 cases contain missing values

Predictor        Coef       StDev          T        P
Constant       -0.000       6.584      -0.00    1.000
RESI5         -31.914       8.019      -3.98    0.000

S = 50.57       R-Sq = 21.7%     R-Sq(adj) = 20.4%
```

a) Note that the estimated slope or regression coefficient is -31.94, the value shown previously

ix. Following it is the results of the partial regression analysis



V. NEXT TIME:

A. Still more on multiple regression.

Go to Notes page

Go to Statistics page