DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816

Regression Methods Continued

I. A SUBSTANTIVE EXAMPLE:

A. Air quality data revisited:

1. Recall that the air quality data showed a possibly curvilinear relationship between mortality and sulfur dioxide.
2. The implicit arrow points down the X axis so I transformed X by taking the square root.
3. Here is the plot:

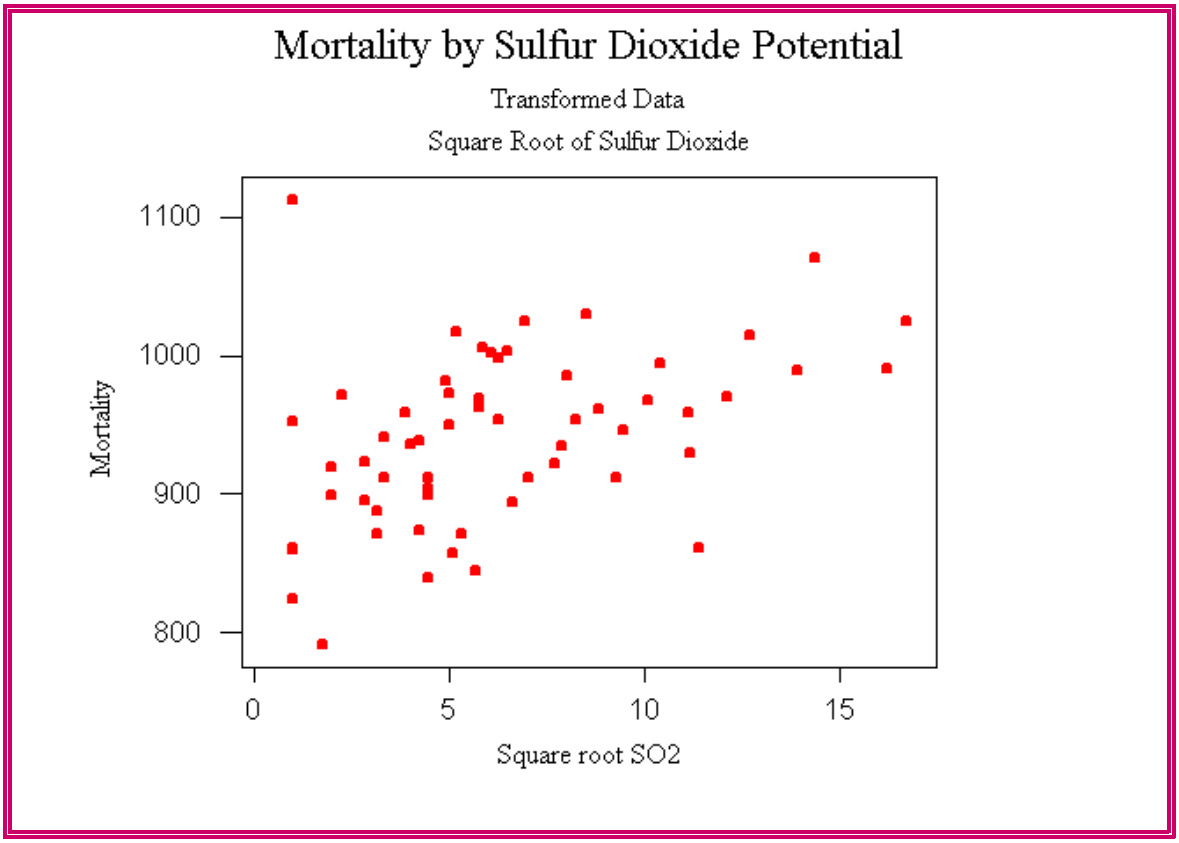

Figure 1

.

4. Now the relationship seems more linear.

B. Example of the Boston crime data.

1. We saw that the crime variable was badly skewed to the lower end of the

scale and its relationship with poverty was perhaps not linear.

2. This skewness can affect regression.
   i. Let's examine a sample (N = 50) cases from the full file.
   ii. One way to see this is plot residuals against fitted or predicted values as we discussed last time.

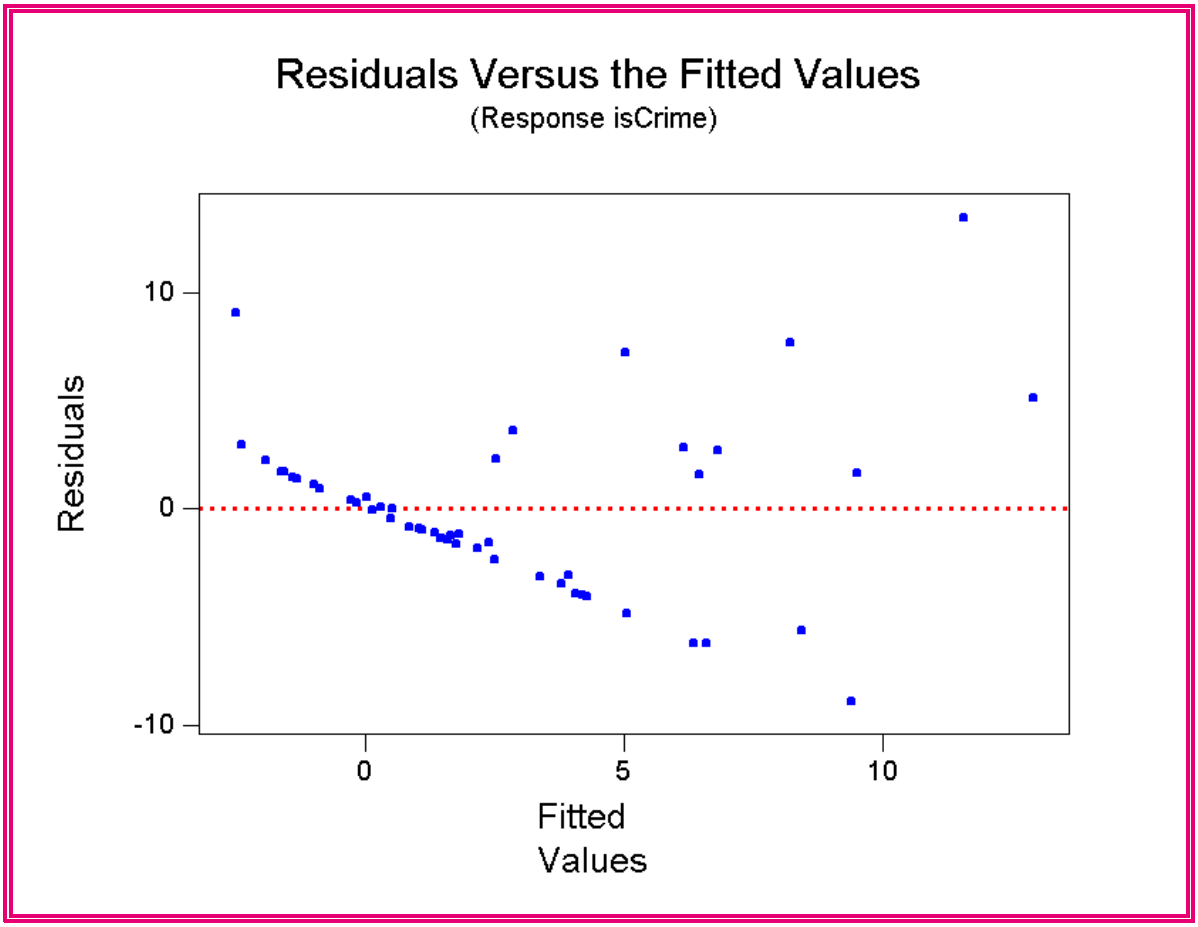

Figure 2

   iii. Something is obviously wrong.

3. The regression analysis shows

```
The regression equation is
samcri = - 4.27 + 0.592 sampoor

Predictor        Coef       StDev          T        P
Constant       -4.271       1.246      -3.43    0.001
sampoor       0.59162     0.09293       6.37    0.000

S = 4.051       R-Sq = 45.8%     R-Sq(adj) = 44.6%
```

i. On the surface these results indicate a good fit.
ii. But we know from the plot of errors that something is wrong.

C. A variable that is skewed toward the lower end of the scale can be transformed by going down the ladder powers.
1. In this example I converted Y to log Y (that is, log to base 10 of per capita crime rate).
2. The regression shows that a linear model is more appropriate for the transformed variable.

```
samlogcr = - 1.65 + 0.114 sampoor

Predictor        Coef       StDev          T        P
Constant      -1.6528      0.2193      -7.54    0.000
sampoor       0.11356     0.01561       7.28    0.000

S = 0.7226      R-Sq = 52.4%     R-Sq(adj) = 51.5%
```

i. Note that the R^2 has increased.
3. The pattern of residuals versus fitted (predicted) values seems more satisfactory.
i. See the figure on the next page.

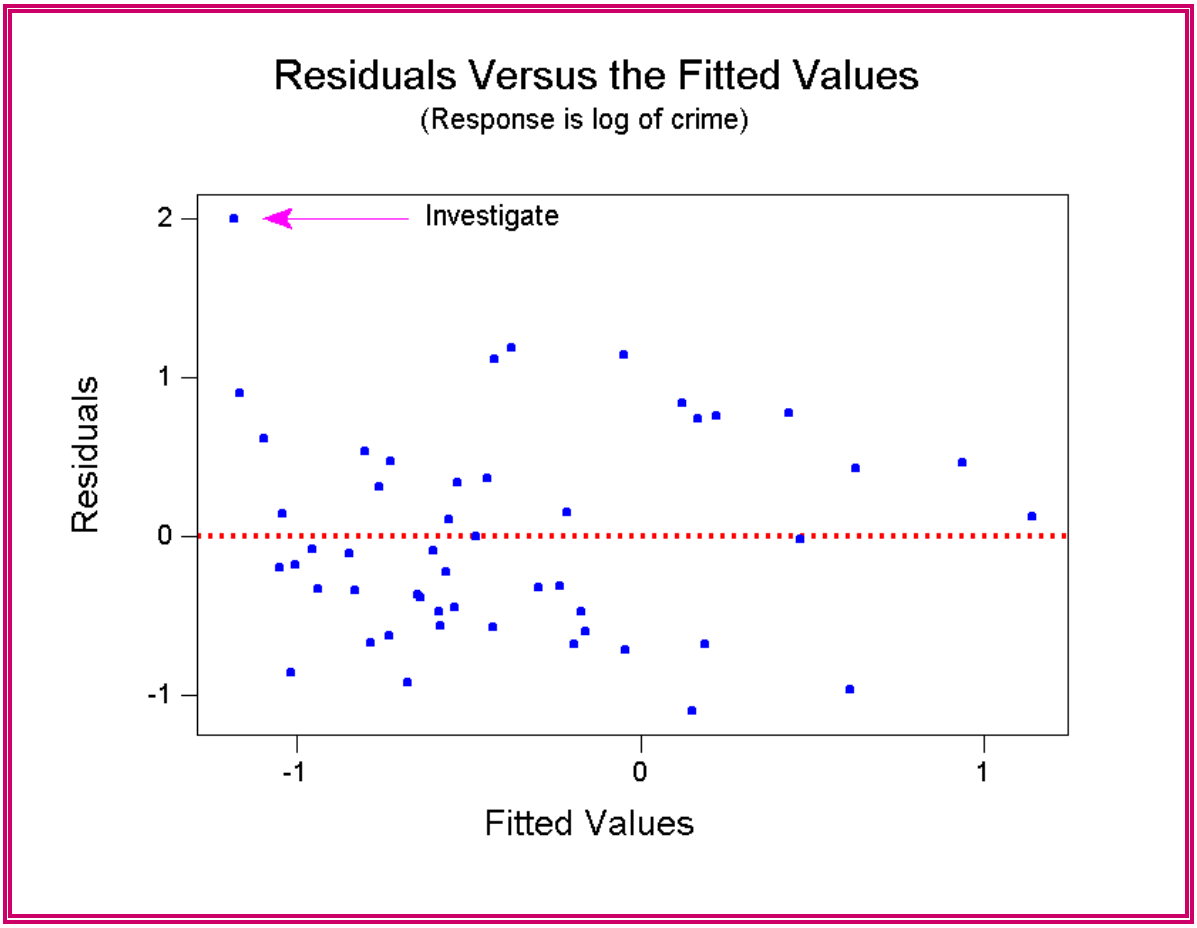

Figure 3

D.

Having analyzed a sample and found a reasonable model we can now apply it to the full data set.

```
The regression equation is
logcrime = - 1.38 + 0.0824 Poor

Predictor        Coef       StDev          T        P
Constant     -1.38146     0.06630     -20.84    0.000
Poor         0.082392    0.004564      18.05    0.000

S = 0.7325      R-Sq = 39.3%     R-Sq(adj) = 39.1%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      174.82      174.82    325.83    0.000
Residual Error   504      270.42        0.54
Total            505      445.24
```

1. These results suggest, as we have found before, that percent classified as poor is closely tied to the (log) crime rate.
   i. Note that R^2 has increased from about .2 for the raw data to about .4 for the transformed variable.

II. NEXT TIME:
   A. Even more on multiple regression.

Go to Notes page

Go to Statistics page