

**DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816**

MISCELLANEOUS REGRESSION TOPICS

I. AGENDA:

- A. Example of correcting for autocorrelation.
- B. Regression with “ordinary” independent variables.
- C. Box-Jenkins models
- D. Logistic regression
- E. Polynomial regression
- F. Reading
 - 1. Agresti and Finlay, Statistical Methods for Social Sciences, 3rd edition, pages 537 to 538; 543 to 550; Chapter 15.

II. AN EXAMPLE OF SERIAL CORRELATION:

- A. Let's continue with the example of petroleum products import data.
- B. Summary of steps to address and correct for autocorrelation
- C. The steps are:
 - 1. Use OLS to obtain the residuals. Store these in somewhere.
 - i. Use a name such as “Y-residuals”
 - 2. Lag the residuals 1 time period to obtain $\hat{\epsilon}_t - 1$. MINITAB has a procedure:
 - i. Go to **Statistics, Time series**, then **Lag**.
 - ii. Pick the column containing the residuals from the model.
 - iii. Pick another column to store the lag residuals and press OK.
 - iv. You should name this column as well to keep the book keeping simple.
 - 1) Also, look at the data window to see what's going on.
 - 3. Use descriptive statistics to obtain the simple correlation between the residual column and the lagged residuals.
 - i. **This is the estimate of ρ , the autocorrelation parameter.**
 - 4. Lag the dependent and independent variables so they can be transformed.
 - i. Make sure you know where they are stored.
 - ii. Again, name them.
 - 5. Now transform the dependent variable:

$$Y_t^* = Y_t - \hat{\rho}Y_{t-1}$$

- i. That is, create a new variable (with the calculator or mathematical

expressions or **let** command) that is simply Y at time t minus ρ times Y at time t - 1.

1) Example:

- a) Suppose Y at time t is stored in column 10, its lagged version in column 15 and the estimated autocorrelation is .568. Then the let command will store the new Y in column 25:

```
mtb>let c25 = c10 - (.568)*c15
```

2) Label it something such as “New-Y”

6. The first independent variable is transformed by

$$X_t^* = X_t - \hat{\rho}X_{t-1}$$

i. When using MINITAB just follow the procedure describe above for Y.

1) Example:

- a) If X_1 , the counter say, is stored in column 7 and its lag in column 17 (and once again the estimated ρ is .568) then the MINITAB command will be

```
mtb>let c26 = c7 - (.568)*c17
```

ii. Since we are dealing with lagged variables, we will lose one observation (see the figure above).

1) For now we can forget about them or use the following to replace these missing first values.

$$Y_1^* = \sqrt{1 - \hat{\rho}^2}Y_1$$

$$X_{1,1}^* = \sqrt{1 - \hat{\rho}^2}X_{1,1}$$

$$X_{2,1}^* = \sqrt{1 - \hat{\rho}^2}X_{2,1}$$

iii. The Y_1^* and $X_{1,1}^*$ mean the first case for Y and X_1 and so forth.

1) If there are more X's proceed in the same way.

7. Finally, regress the transformed variables (i.e., Y^* on the X^* 's to obtain new estimates of the coefficients, residuals, and so forth.

i. Check the model adequacy with the Durbin-Watson statistic, plots of errors and the usual.

ii. If the model is not satisfactory, treat the Y^* and X^* as “raw” data and go through the process again.

D. Here are the results for the petroleum data.

1. First the “original” equation:

Petroimp = - 0.167 + 0.0961 Counter + 2.47 Dummy - 0.106 Inter

Predictor	Coef	StDev	T	P	VIF
Constant	-0.1670	0.1057	-1.58	0.121	
Counter	0.096077	0.007109	13.51	0.000	7.1
Dummy	2.4732	0.3208	7.71	0.000	18.8
Inter	-0.10569	0.01075	-9.84	0.000	30.6

i. The Durbin-Watson statistic is .53.

ii. The estimate of the autocorrelation coefficient, ρ is .732.

E. Evaluating the Durbin-Watson statistic.

1. Attached to the notes is a table of the Durbin-Watson statistic for the .05 and .01 levels of significance.

2. As noted last time, find (N, the number of time periods; I've sometimes call it T) and the number of parameters K.

i. In the table this is denoted with p.

3. Find the lower and upper bounds.

i. In this instance they are (using $N = 50$, since there isn't a row 48, and $p = 4$): $D_l = 1.42$ and $D_u = 1.67$.

4. Decision: since the observed DW falls below the lower bound, reject H_0 that there is no autocorrelation.

i. In as much as it is positive we proceed to estimate its value and transform the data.

F. Transformation

1. Use the formulas described last time to create “new” variables.

2. After carrying out the transformation the model and test statistics are:

NewY = - 0.153 + 0.120 Newcount + 2.91 Newdummy - 0.130 Newwinter

47 cases used 1 cases contain missing values

Predictor	Coef	StDev	T	P	VIF
Constant	-0.15266	0.08706	-1.75	0.087	
Newcount	0.11969	0.01782	6.72	0.000	6.9
Newdummy	2.9098	0.6808	4.27	0.000	25.2
Newwinter	-0.13019	0.02631	-4.95	0.000	43.7

S = 0.1703 R-Sq = 54.1% R-Sq(adj) = 50.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	1.46723	0.48908	16.87	0.000
Residual Error	43	1.24659	0.02899		
Total	46	2.71382			

Durbin-Watson statistic = 1.85

- i. The estimated DW is 1.85, which falls above the upper bound. Hence, we can assume no serial correlation.
 - 1) Note that we have a smaller N, since I didn't bother estimating the first values of Y and X.
 - ii. The estimated ρ is .035.
 3. Both it and the Durbin-Watson statistic suggest that autocorrelation is not a problem.
 - i. But the plot of residuals suggests that it might be.
- G. Multicollinearity
1. As noted previously in the semester, one almost always encounters multicollinearity when using dummy variables.
 2. This situation can cause parameter estimates to be very unstable, especially since in time series analysis $N = T$ is often not very large.
 3. What to do?
 - i. Try dropping a constant or level dummy variable if it (the level) is not of great theoretical importance or if it obviously changes with time, as it almost always will.

III. LAGGED REGRESSION:

- A. So far we have dealt with a special kind of time series data in which the "independent" variables were all dummy indicators or counters.
- B. But it is possible to use "ordinary" independent variables.

1. One will, however, have to look for and possibly adjust for serial correlation.
 2. An example follows.
- C. Example:
1. Here is a very simple (but real example). The data, which come from a study of crime in Detroit, present several difficult problems, not the least of which is the small sample size.¹ The variables are:
 - i. Y: Number of homicides per 100,000 population
 - ii. X₁: Number of handgun license per 100,000 population.
 - iii. X₂: Average weekly earnings
- D. The time period was 1961 to 1973. Although they are perhaps dated, these data still illustrate the potential and problems of time series regression.
- E. First suppose we simply estimate the model:

$$E(Y_t) = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2}$$

1. Here the X's are number of handgun licenses and average weekly earnings at time t (for t = 1, 2,...,13) and Y is the homicide rate.
2. Simple OLS estimates are:

$$\hat{Y}_t = -34.05 + .0232X_1 + .275X_2$$

(-7.83) (6.39) (10.18)

- i. Note that the observed t's are in parentheses. It is common practice to write either the t value or standard deviation of the coefficient under or next to the estimate.
3. Superficially, the data fit the model very well: $F_{\text{obs}} = 115.21$; $s = 3.661$; $R^2 = .958$.
4. But we know that the errors will probably be a problem. The residuals plotted against the time counter (see Figure 1 at the top of the next page) seem to follow a distinct pattern: negative values tend to be followed by negative numbers and similarly for positive ones. But we have to be careful

¹ J. C. Fisher, collected the data for his paper "Homicide in Detroit: The Role of Firearms," *Criminology*, 14 (1976):387-400. I obtained them via the Internet from "statlib" at Carnegie Mellon University.

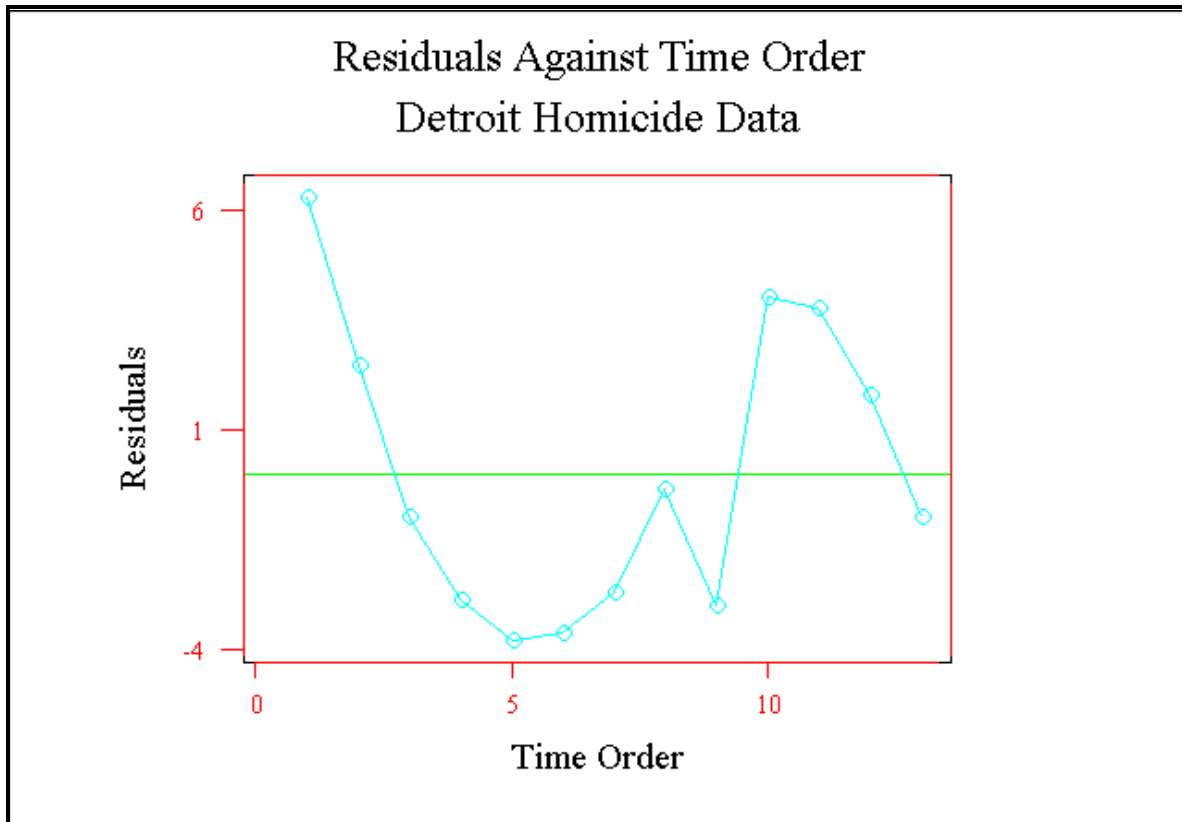


Figure 1: Residuals Suggest Serial Correlation

because the sample is so small, we may find "runs" like these just by chance.

- i. An aside: it's easy to see the pattern of pluses and minuses by applying the **sign** command to the data in a column; it produces a "1" for each positive number, a "-1" for each negative value, and "0" for values equal to zero.

- 1) For example, In MINITAB go to the session window and type, **sign column number column number**.

- a) Example: if the numbers -.2 -.3 .05 .5 .89 0 and -.66 are stored in c1, the instruction **sign c1 c2** will put -1 -1 1 1 1 0 -1 in column c2. (You can see the pattern by printing c2 on your screen.)

5. In addition the Durbin-Watson statistic is .79. At the .05 level, for $N = 13$, the lower bound is 1.08, so we should reject the null hypothesis of no serial correlation (i.e., $H_0: \rho = 0$), concluding instead that $\rho > 0$.
6. **Note: the Durbin-Watson test "works" only under a relatively narrow range of circumstances. Hence, it should be supplemented with other procedures (such as residual plots). Moreover, many analysts argue**

that if DW does not exceed the upper limit, the null hypothesis should be rejected, even if the test result falls in the indeterminate range. That is good advice because correcting for serial correlation when there is none present will probably create fewer problems than assuming it is not a factor.

F. Solutions:

1. As we did in the previous class we can transform the data by estimating ρ and transforming the data. (That is, lag the residuals from the OLS regression and obtain the regression coefficient between the two sets of errors--losing one observation in the process--and creating new X's and Y as in

$$X_1^* = X_1 - \hat{\rho}X_{t-1}$$

2. Then, regressing these variables to obtain new estimates of the coefficients, examine the pattern of errors, and so forth.

IV. ALTERNATIVE TIME SERIES PROCESSES:

- A. The previous example involved a couple of independent variables measured at time t.
- B. It also possible to include **lagged exogenous** or independent variables as in:

$$Y_t + \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{1,t-1} + \epsilon_t$$

1. The term “exogenous” has a particular meaning to many analysts, but I use it loosely as a synonym of “independent.”
2. Here Y is a function of X at both t and t - 1.
 - i. That is, the effects of X last over more the current time period.
3. These models are approached in much the same way the other ones have been: we look for and attempt to correct serial or autocorrelation.

C. Lagged **endogenous** or dependent variables.

1. It is common to find models in which the dependent variable at time t is a function of its value at time t - 1 or even earlier.
2. Example:

$$Y_t + \beta_0 + \beta_1 Y_{t-1} + \epsilon_t$$

- i. I wonder if the budget of the United States government couldn't be modeled partly in this way:

- 1) The budget process used to be called “disjointed incrementalism,” which meant in part that one year’s budget was the same as the previous years plus a small increment.
 - a) The expected value of ε_t would be positive, I suppose.
 3. Models with lagged endogenous (and possible lagged or non-lagged exogenous) variables have to be treated in somewhat different fashion than those previously discussed because the autocorrelation parameter cannot be estimated as before.
- D. ARMA models.
1. So far, I have dealt almost exclusively with first-order positive autocorrelation.
 - i. The error term is a partial function of its value at the immediately preceding time period.
 2. But it is possible that the error term reflects the effects of its values at several earlier times as in:

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \dots + v_t$$

 - i. A model incorporating this sort of error process is called an autoregressive process of order p, AR(p).
 3. A **moving average** error process involves a disturbance that is created by a weighted sum of current and lagged random disturbances:

$$\varepsilon_t = \delta_1 v_t + \delta_2 v_{t-1} + \dots + \delta_q v_{t-q}$$

 - i. This is called a moving average model of order q: MA(q).
 4. It’s possible to have **mixed** models involving both types of error structures such as

$$\varepsilon_t = \rho_1 \varepsilon_{t-1} + v_t + \delta_1 v_{t-1}$$
- E. It’s necessary to used alternative methods to identify and estimate the models, a subject we cannot get to in this semester.
- F. Many social scientist feel that p and q almost never exceed 1 or possibly 2. I have, however, on occasions seen somewhat “deeper” models.
- G. Finally there is still another version called an ARIMA model for **autoregressive integrated moving average**.
1. Again we do not have time to explore them.

V. LOGISTIC REGRESSION.

A. A **binary** dependent variable

1. Consider a variable, Y , that has just two possible values.
 - i. Examples: "voted Democratic or Republican"; "Passed seatbelt law or did not pass seatbelt law"; "Industrialized or non industrialized."
 - ii. The values of these categories can be following our previous practice, be coded 0 and 1.
2. If such a variable is a function of X (or X 's), the mean value of Y , denoted $p(X)$, can be interpreted as a probability, namely, the conditional probability that $Y =$ a specific category (i.e., $Y = 1$ or 0), given that X (or the X 's) have certain values.
3. It is tempting to use such a variable as a dependent or response variable in OLS, and in many circumstances we can obtain reasonable results from doing so. Such an model would have the usual form:

$$Y_i = \beta_0 + \beta_1 X_1 + \dots + \epsilon_i$$

4. Still, there are two major problems with this approach.
 - i. It is possible that the estimated equation will produce estimated or predicted values that are less than 0 or greater than 1.
 - ii. Another problem is that for a particular X , the dependent variable is 1 with some probability, which we have denoted above as $p(X)$, and is 0 with probability $(1 - p(X))$.
5. The variance of Y at this value of X can be shown to be $p(X)(1 - p(X))$.
 - i. Consequently, the variance of Y depends on the value of X . Normally, this variance will not be the same, a situation that violates the assumption of constant variance.
 - ii. Recall the discussion and examples earlier in the course that showed the distribution of Y 's at various values of X . We said that the variance of these distributions is assumed to be the same up and down the line.

B. For these and other reasons social scientists and statisticians do not usually work with "linear probability models" such as the one above. Instead, they frequently use **odds** and **odds ratios** (and the log of odds ratios) as the dependent variable.

1. Suppose π_i is the probability that a case or unit has $Y = 1$, given that X or X 's take specific values. Then $(1 - \pi_i)$ is the corresponding conditional probability that Y is 0.
2. The ratio of these two probabilities can be interpreted as the conditional odds of being Y having value 1 as opposed to 0. The odds ratio can be denoted:

$$\Omega_i = \frac{\pi_i}{(1 - \pi_i)}$$

3. For various reasons it is easier to work with the **logarithm** of this quantity:

$$O_i = \log\left(\frac{\pi_i}{(1 - \pi_i)}\right)$$

4. O_i is considered the dependent variable in a linear model.
5. The approach is called **logistic** regression.
i. We'll discuss it next time.

VI. NEXT TIME:

- A. Additional regression procedures
1. Resistant regression
2. Polynomial regression
B. Models and structural equations

Go to Notes page

Go to Statistics page