DEPARTMENT OF POLITICAL SCIENCE
AND
INTERNATIONAL RELATIONS
Posc/Uapp 816

LOGISTIC AND NONLINEAR REGRESSION

I. AGENDA:
   A. Some remarks about time series analysis
   B. Logistic regression
   C. Non-linear (polynomial) regression
   D. Reading: Agresti and Finlay ***Statistical Methods in the Social Sciences***, 3rd edition, pages 576 to 585.

II. TIME SERIES:
   A. Adjustment for non-constant variance

III. LOGISTIC REGRESSION:
   A. Substantive problem:
      1. Suppose you wanted to know why certain cities in the United States experienced what have euphemistically been call "urban disorders," that is incidents of widespread looting, arson, and so forth.
         i. An aside: statistics cannot address the "proper" nomenclature for the problems to which it is applied. For example, one person's riot and breakdown in law and order will be another individual's episode of "unconventional" participation.
      2. It might be hard to measure the variable, disorder, any other way than it occurred or did not occur.
      3. If so, we might conceptualize this variable, which can take only two values, namely "present" and "not present," as dependent on or caused by various independent variables such as general level of discontent, unemployment rate, urban decay, and social tension, all of which might be measured as at least categorical variables.
      4. Our problem could be redefined as predicting or "postdicting" the occurrence (or not) of a riot or disorder or massive amount of unconventional participation.
   B. Representation example:
      1. Suppose we wanted to know how well members of Congress represent their constituent.
         i. Another, related question is: what explains their voting behavior? Do they vote for legislation on the basis of party affliation, ideology, or constituency characteristics.
      2. As an example, suppose we want to know why House members voted for

or against a resolution that would have appealed the ban on so-called assault weapons.

IV. MODELING A "BINARY" DEPENDENT VARIABLE:

A. The phenomenon to be explained is really the occurrence or realization of an event that has just two possible states, presence or absence.

B. One way of representing this is to define the dependent variable as Y, which can be either 0 for absence of event or condition, which is sometimes called a "failure" or 1 for (presence of event or condition, which sometimes called a "success."

1. Examples: vote can take on values, say, 0 for "vote for repeal" and 1 for "vote against repeal."

i. Similarly, the variable Y might be defined as 1 = occurrence of civil disorder and 0 = no occurrence.

C. We usually are concerned not with a particular case, but with **probabilities** that Y will be 0 or 1.

1. This, of course, represents a concern with the probability that a city experiences a disorder or a representative votes "yes."

2. We will let π be the probability that Y = 1 and 1 - π stand for the probability that Y = 0.

3. Logistic regression can be thought of as a method for understanding variation in these probabilities.

i. That is, why is the probability that Y = 1 so high for some people or cities and so low for others?

D. The problem could be conceptualized more formally as:

$$\pi(Y = j|X_k) = f(X_k), \quad j = 0,1$$

i. The probability the variable, Y, takes the value j (e.g., 0) given that X's equal certain values is a function of these independent or explanatory factors.

ii. This model could and perhaps should be written more simply as

$$\pi(Y) = f(X)$$

1) This means the probability that a city experiences a disorder or not (or representative votes for repeal) is a function of X.

1. Since in any realistic setting numerous more or less random factors enter the picture, the model should include a random error or disturbance:

$$\pi(Y) = f(X) + \varepsilon$$

i. ε is the usual random error with mean 0 and standard deviation σ_ε

E. Now, doesn't this resemble a typical regression model with an error term?

1. Indeed, if we thought of Y as being a typical variable, albeit one that could only take values of 0 and 1, we might want to consider the equation:

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i$$

2. It turns out, however, that this approach has several drawbacks.
   i. Y, after all, is not a "typical" variable and we are interested in the chances or probability of something happening or not happening.
3. Thus, if we thought of not Y, but the probability that Y equaled 1, then the model would properly be written as:

$$\pi_i = P(Y_i = 1) = \beta_0 + \beta_1 X_1 + \varepsilon_i$$

4. Using probability of Y rather than simple Y has several consequences.

F. The preceding model that treats the probability (that Y equals 1) as linear function of several Xs is called not surprisingly a **linear probability model (LPM).**

1. Such models, which can be estimated and evaluated by standard ordinary least squares (regression) techniques, frequently "work" reasonably well.
   i. Work in the sense that their predictions and estimates are sensible.
2. But they have several drawbacks and are better replaced by modifications.
3. One draw back is that estimated probabilities can sometimes be nonsensical. An estimated model might predict, for instance, that the probability of a "no vote" for some combination of X values is -.12. But probabilities, by definition, must lie between 0 and 1.0.
   i. A predicted probability of 1.2 would be similarly nonsensical.
4. Also, the assumption about the variation of error terms being constant will be violated, since the errors' standard deviation are $\pi(1 - \pi)$.
   i. That is, the variation around the least squares line will change with changes in the X values that give rise to the probabilities.

G. For these reason social scientists and statisticians prefer an alternative formulation.

V. LOGISTIC MODEL:

A. Again let π be the probability that Y = 1.

B. A model for π that addresses the concerns listed above is to use the "log odds," also called the **logit**:

$$\Omega = f(X) = \log\left(\frac{\pi}{1 - \pi}\right) = \beta_0 + \beta_1 X$$

1. This expression can be interpreted directly as the log of the odds of a "success" (e.g., a "yes vote" or the "occurrence of a riot") to the odds of a "failure" (e.g.," voting no" or" non-occurrence of a riot").
2. These log odds are a linear function of X (or X's if there are more than one independent variable).
   i. The terms can be interpreted in the usual fashion. For example, β_0 = is the value of the log odds, when X = 0.
   ii. But like dummy variable analysis there are more straightforward and meaningful interpretations.
   iii. Let's go through some of these.

C. Interpretation:
1. The log odds are a **linear** function of X.
2. The β_1 measures the (additive) change in the **log** odds for a one-unit change in X.
3. If we take the "antilog" of both sides of the model, we see that β_1 is the multiplicative effect on the **plain** odds of a one-unit change in X.
   i. That is, if we apply the antilog (exponential operator) **e** to both sides of the model for the log odds we get the plain odds:

$$O = e^{\log\left(\frac{\pi}{1-\pi}\right)} = e^{\beta_0 + \beta_1 X}$$

   ii. Note the antilog operator, denoted e, of a natural log is $e^{\log(Z)} = Z$.
   iii. So

$$O = \frac{\pi}{1-\pi} = e^{\beta_0 + \beta_1 X}$$

      1) By the way most calculators have a natural log function (usually abbreviated **ln** or **LN**) and antilog key (usually abbreviated e^X).
      2) MINITAB and SPSS, along with all other statistical program packages, have these functions.
4. The last equation can be written equivalently as:

$$O = \frac{\pi}{1-\pi} = e^{\beta_0 + \beta_1 X} = e^{\beta_0} e^{\beta_1 (X)}$$

      1) Note: $e^{a+b} = e^a e^b$
   i. This equation shows that X affects the plain odds multiplicatively

through β_1

5. Defined this way, this model has the property that estimated or predicted **logits** can lie anywhere on the number line from minus to plus infinity, while odds will lie between 0 and infinity..

D. To see better the meaning of the terms consider a very simple example and some data:

$$\log\left(\frac{\pi}{1-\pi}\right) = 1 + 1X$$

1. That is, $\beta_0 = 1$ and β_1 also = 1.
2. Suppose the values of X are -5, -2, 0, 1, 2.

X	$\Omega = 1 + 1X$	$O = e^{1 + 1X}$
-5	-4	.0018
-2	-1	.3679
0	1	2.718
1	2	7.389
2	3	20.09
3	4	54.60

   i. For negative values of X, the log odds can decrease without end, but the odds converge to 0.
   ii. At the other end of the scale, as X becomes increasingly large, the log odds and plain odds increase without limit.
      1) Briefly, the log odds, $-\infty < \Omega < \infty$, whereas $0 \leq O < \infty$

3. We see that X has a **linear** or **additive** effect on the **log** odds: as X increases one unit, the logit increases one unit.
4. For the plain odds, however, $\beta_1 = 1$ (as in this example) indicates that X has a **multiplicative** impact, which depends on the value of X.
   i. The antilog of β_1 is 2.718.
   ii. When X increases by one unit, the odds will multiply by 2.718.
      1) As X increases from 0 to 1 (a one unit change), the odds increase from 2.718 to (2.718 X 2.718) = 7.389; another one unit increase in X changes the odds to 2.718 X 7.389 = 20.09; still another one unit change in X gives odds of 2.718 X 20.09 = 54.60.

E. We can use algebra to find the expression for the probability, π or simple odds of a

"success":

$$\pi = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

1. If we had values for the betas, we could find probabilities that Y = 1, a success, for various values of X.
2. We'll see how to do so shortly.

F. These are of course abstract ideas and artificial data, so let's turn to some real data.

VI. EXAMPLE OF LOGISTIC REGRESSION:

A. At this point we can't show how data are turned into estimated (logistic) regression parameters, except to say the basic idea is roughly the same as least squares:
   1. Estimated values of the coefficients are chosen in such a way as to minimize certain sums or maximize other sums.
      i. Shortly I'll refer to maximum likelihood estimators, which choose estimates that maximize the likelihood of data for a given model with certain parameters.
      ii. Maximum likelihood estimation treats a model and data points as fixed and looks for the values of the betas that maximize the likelihood of that model and data.
      iii. We can't go much further now since more explanation requires some mathematical and statistical understanding that we haven't acquired yet.

B. But we can interpret the parameters with the help of the preceding ideas.

C. I have collected data on California's 52 representatives:
   1. Rural: percent of the representative's district living in rural areas.
      i. They hypothesis: congressmen and women from rural areas will be apt to vote to overturn the ban on assault weapons because the "gun lobby" has greater influence in less populated areas. Moreover, the gun lobby can and does overwhelm party affiliation and political ideology.
   2. ADA score: a measure of "liberalism." The highest score, 100, means most liberal, 0 means most conservative.
      i. This variable, which runs from 0 to 100, allows us to investigate political ideology on representatives' votes.
      ii. It allows us to see how a general political ideology factor explains congressional voting on a "social" issue.
   3. Guns = 1 for no vote on amendment to repeal the ban on assault weapons;

0 for yes.

4. Wages = 1 for yes vote to increase minimum wage, 0 for no vote.
   i. These are just dummy variables that are now treated as dependent or responses.
   ii. In the first case, the reference category is the vote to repeal the ban ("yes") where as in the second the reference category is no.
5. Party affiliation: 1 for Democrat and 0 for Republican.
   i. Another dummy variable that is used as an independent or explanatory factor.
6. We won't use all of these today.

D. Let's start with the amendment to lift the ban on assault weapons.

1. Suppose we hypothesize that votes on this issue are determined by population density, the variable we called "rural."
2. Let π = the probability that Y = 1, that is, a "no" vote.
3. A model for the log odds of a vote to keep the ban is

$$\Omega = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 Rural + \varepsilon$$

4. The corresponding model for the odds will be:

$$O = \left(\frac{\pi}{1-\pi}\right) = e^{\beta_0 + \beta_1 Rural} + \varepsilon$$

5. And a the model for the probability of a vote against the repeal (Y = 1) is

$$\pi = \frac{e^{\beta_0 + \beta_1 Rural}}{1 + e^{\beta_0 + \beta_1 Rural}} + \varepsilon$$

E. We can use MINITAB (or any program that has logistic regression) to estimate the parameters and assess the goodness of fit of the data to the model, just as with simple regression.

1. SPSS has logistic regression as part of an "add on" module.
2. Here are the results of the California representatives data.
   i. I used MINITAB's **binary logistic regression** procedure.

```
Response Information

Variable  Value       Count
Guns      1              26  (Event)
          0              11
          Total          37

   37 cases were used
    9 cases contained missing values

Logistic Regression Table
                                                  Odds        95% CI
Predictor       Coef      StDev         Z      P  Ratio    Lower    Upper
Constant      1.6668     0.5292      3.15  0.002
Rural       -0.08944    0.03747    -2.39  0.017   0.91     0.85     0.98

Log-Likelihood = -18.856
Test that all slopes are zero: G = 7.321, DF = 1, P-Value = 0.007
```

3. We see that 26 representatives voted "no" (that is, voted against the repeal effort) and 11 voted yes.
   i. There are 9 missing cases, representatives who didn't vote or were new to Congress or whatever.
   ii. I also omitted some for whom data were not available on short notice.
   iii. Hence the effective N = 37.
4. The estimated equation is:

$$\hat{\Omega} = \log\left(\frac{\pi}{1-\pi}\right) = \mathbf{1.6668} - \mathbf{.08944}\textit{Rural}$$

   i. MINITAB and SPSS present the results in a form reminiscent of regression output:
      1) The estimated constant and regression coefficient are listed along with their estimated standard errors and an observed significance test statistic and attained probability.
      2) More of these later.

F. The numbers can be interpreted using the ideas discussed above.
1. Suppose rural = 0. The equation would reduce to a predicted logit or log odds = 1.6668.
2. If we take the antilog of this value, we get $e^{1.668} = 5.2952$. This means that for representatives from in totally urban districts (X = 0) the odds of a no vote are 5.3 times those of a yes vote.
   i. In other words, the odds of voting "no" are 5.92952 to 1.
3. What about the regression coefficient, $\hat{\beta}_1 = -.08944$?

i. It gives the linear effect of a one-unit change in X on the log odds. But we would probably be more comfortable interpreting the simple odds so take the antilog:

$$e^{\hat{\beta}_1} = e^{-.08944} = .91444$$

ii. Hence, the odds of a no vote are multiplied (that is, **decreased**) by .91444 for each one percent increase in rural population in the district.

1) Note by the way, that this value appears in the MINITAB results under the "odds ratio" column.

4. If e^{β}_1 is less than 1.0, the odds will be reduced for a unit increase in X; if e^{β}_1 is greater than 1.0, the odds will be increased. If it equals = 1, then X does not affect the odds.

i. Hence we will later be testing the hypothesis that $\beta_1 = 0$ in order to see if e^{β}_1 is close to 1.

5. Example: if the odds of voting no are 5.295, as above, a one percent increase in rural population will decrease them to .91444 X 5.295 = 4.8422.

i. A 10 percent increase in rural population in the district would decrease the odds to $(.91444)^{10}$ X 5.2952 = 2.16496.

1) Remember: .91444 represents the estimated coefficient -.08944 to **e**. That is, $e^{-.08944(10)} = .91444^{10}$

ii. Or we could compare the odds of voting "no" in districts with 0 and 100 percent rural inhabitants:

1) For 0 percent rural (i.e., urban) districts: estimated odds = 5.2952 to 1, as above.

2) For 100 percent rural: estimated odds = .00069 to 1.

iii. By using your calculator or the mathematical expressions (or calculator) functions in MINITAB, you can explore other possibilities.

6. For example, if you look at the estimated odds when X = 5, say, and when X = 6, you can take the ratio of these odds to find the antilog of the estimated regression coefficient.

i. Estimated log odds for X = 5 is

$$\hat{\Omega}_{X=5} = 1.6669 - .08944(5) = 1.2171$$

and

$$\hat{\Omega}_{X=5} = e^{1.2171} = 3.37778$$

ii. When X = 6 we have

$$\hat{\Omega}_{X=6} = 1.6669 - .08944(6) = 1.13016$$

and

$$\hat{\Omega}_{X=6} = e^{1.13016} = 3.09615$$

iii. Finally the ratio of these two odds is 3.09615/3.37778 = .9166, which is about what we got above except for rounding error.

G. Just so we have some more numbers to work with, let's regress the gun vote on ADA scores, the measure of ideology.

1. The results from MINITAB's binary logistic regression are:

```
Logistic Regression Table
                                                    Odds         95% CI
Predictor       Coef      StDev         Z      P    Ratio    Lower    Upper
Constant      -2.945      1.252    -2.35 0.019
ADA96        0.15339    0.07450     2.06 0.040     1.17     1.01     1.35

Log-Likelihood = -5.610
Test that all slopes are zero: G = 33.814, DF = 1, P-Value = 0.000
```

.

i. The estimated equation for the logit is:

$$\hat{\Omega} = -2.945 + .15339ADA$$

ii. And the antilog of the regression parameter, which is the odds ratio, is $e^{.15339} = 1.16578$.

iii. As we've seen, a one unit increase in X is associated with .15339 increase in the log odds and a multiplicative increase in the odds or 1.16578.

iv. To grasp the substantive meaning of the numbers it is sometimes helpful to obtain estimated odds for various meaningful values of X.

1) Suppose, for example, X = 0. Then, the estimated log odds are: -2.945 + .15339(0) = -2.945. Taking the antilog, we see that the odds of a "no vote" to a "yes vote" for extreme conservatives (ADA =) is $e^{-2.945} = .0526$ to 1.

2) That is, there is virtually no chance a strong conservative will vote against the repeal. To see this even more clearly, get the estimated probability of voting "no" from the

formula given above:

$$\hat{\pi} = \frac{e^{-2.945 + .15339(0)}}{1 + e^{-2.945 + .15339(0)}} = \frac{e^{-2.945}}{1 + e^{-2.945}} = \mathbf{.04997}$$

3) NOTE: what we just obtained is the probability of a strong conservative voting "no," which turned out to be about .05. Prior to that we found the **odds** that conservatives vote "no" to the odds that liberals do. That odds ratio was about .05 to 1.

v. Using the same methods we find that the estimated log odds for X = 100, the most liberal members of the California delegation, are: -2.945 + .15339(100) = 12.394.

1) This translates to an estimated odds of $e^{12.394}$ equals about 241,349 to 1. Liberals are overwhelmingly likely to vote against repealing the ban compared to conservatives.

2) So not surprisingly the estimated probability of an extreme liberal (X = ADA = 100) voting "no" is

$$\hat{\pi} = \frac{e^{-2.945 + .15339(100)}}{1 + e^{-2.945 + .15339(100)}} = \frac{e^{12.394}}{1 + e^{12.394}} = \mathbf{1.0000}$$

a) Actually the probability is not 1.0 but .9999...as far as the eye can see.

3) Now what about someone who is middle of the road, ADA = 50? The estimated probability of a no vote is still quite high.

$$\hat{\pi} = \frac{e^{-2.945 + .15339(50)}}{1 + e^{-2.945 + .15339(50)}} = \frac{e^{4.7245}}{1 + e^{4.7245}} = \mathbf{.9912.}$$

vi. It's only with the very conservative members, say ADA = 10, that the probability of a no vote declines. (By the way, what are the estimated log odds, odds, and probability of a no vote for ADA = 10?)

VII. MULTIPLE VARIABLE LOGISTIC REGRESSION:

A. We can approach model building, that is, explaining "variation" in the log odds or odds, in the same way we did with multiple regression.
   1. We can add variables of any of the types discussed under that topic.
   2. Hence, we can add more continuous Xs;
   3. Dummy indicators for categorical variables.
   4. Interaction terms.

B. The regression parameters will be partial regression coefficients that show the effects of a variable on the logit when the other variables in the model have been held constant or controlled.

C. Let's use both percent rural and ADA scores to illustrate the point.
   1. Here are the results:

```
Logistic Regression Table
                                                    Odds        95% CI
Predictor       Coef      StDev         Z     P    Ratio    Lower    Upper
Constant      -2.711      1.562    -1.74 0.083
Rural       -0.01497    0.06533    -0.23 0.819     0.99     0.87     1.12
ADA96        0.14762    0.07474     1.98 0.048     1.16     1.00     1.34

Log-Likelihood = -5.582
Test that all slopes are zero: G = 33.869, DF = 2, P-Value = 0.000
```

D. We need to wait a minute before deciding whether or not using two predictors helps.
   1. But you might anticipate on what has gone before that there really won't be an improvement.
      i. For one thing, there is a relatively strong negative correlation between ADA and percent rural, -.461.
   2. The estimated equation for the log odds is:

$$\hat{\Omega} = -2.711 - .01497Rural + .14762ADA$$

   3. To see what the numbers mean just substitute some meaningful values for X_1 and X_2 such as 0 and 0.
      i. Actually this combination would not make sense in American politics because it's unlikely that an extreme conservative (ADA = 0) would represent a totally urbanized district (rural = 0).
      ii. Anyway, the estimated log odds would be -2.711 and the estimated odds would be $e^{-2.711} = .0665$ to 1.
      iii. What would the log odds, odds and predicted probability be for a representative from a district with a rural population of 30 and an ADA score of 70?

VIII. A NOTE ON ESTIMATION:

A. Logistic regression estimation breaks down if all of the X values that correspond to Y = 1 exceed all of the X values that correspond to Y = 0.

1. Thomas Ryan calls this a **complete separation** of the data.[1]
2. The following table provides a simple example.

X	Y
11	0
14	0
17	0
21	0
33	0
36	0
40	1
43	1
48	1
49	1
55	1
59	1
60	1
64	1
66	1

i. No
te that the X values associated with Y = 0 are all smaller than those associated with Y = 1. In this situation, a program might produce results but will usually flash a warning such as

```
NOTE * Algorithm has not converged after 20 iterations.
       * Convergence has not been reached for the
       * parameter estimates criterion.
       * The results may not be reliable.
       * Try increasing the maximum number of iterations.
```

ii. If this case arises try collecting more cases so that there is some separation in the data.

[1]Thomas P. Ryan, ***Modern Regression Methods*** (Wiley, 1997) page 263.

IX. TESTING AND EVALUATING LOGISTIC REGRESSION:

A. We can test the significance of the estimated parameters using the same as ideas as we employed in regular regression. In particular, we can
   1. Compute statistics roughly comparable to R^2 as a measure of how well the data fit the model.
   2. An overall or global test of the regression parameters.
   3. Tests for individual parameters.
   4. Confidence intervals for estimated parameters.
   5. Confidence intervals for predicted probabilities.

B. Note and warning:
   1. The statistical results for logistic regression usually assume that the sample is relatively large.
   2. If, for example, a statistic such as the estimator of the regression coefficient is said to be normally distributed with a standard deviation of σ_β, the statement applies strictly speaking for estimators based on large N.
      i. How big does N have to be? A rule of thumb: 60 cases.

C. The R^2 analogue.
   1. There really isn't a completely satisfactory version of R^2 available to measure the "explained variation" in Y similar to common multiple R, so we will use a different measure, the **correct classification rate** (CCR).
   2. MINITAB effectively constructs a cross-classification table of predicted and observed results that takes this form.

Observed/Predicted	$\hat{Y} = 0$	$\hat{Y} = 1$
Y = 0	Number correct	Number incorrect
Y = 1	Number incorrect	Number correct

      i. The table cross-classifies predicted Y's by observed Y's.
   3. If the model does a good job, then presumably the total number of correct predictions--the frequencies in the main (shaded) diagonal--should greatly outweigh the incorrect guesses.
      i. For instance, suppose a model led to this pattern of correct and incorrect predictions.

Observed/Predicted	Y = 0	Y = 0
Y = 0	47	4
Y = 1	3	29

ii. Since there a total of 83 observations in the table and 76 of them have been correctly predicted, the CCR is 76/83 X 100 = 91.16%.

4. Some software reports this number or it can be easily calculated from reported data.
5. MINITAB, however, reports "measures of association" for the table.
   i. These measures are bounded between -1.0 and 1.0 and attain maximum values (1.0) when there are no errors.
   ii. So a measure equal to .9 indicates that most of the Y's have been correctly predicted and the model fits reasonably well.
6. The measures for the percent rural and the ADA models are:

```
          Percent Rural
Somers' D                0.51
Goodman-Kruskal Gamma    0.57
Kendall's Tau-a          0.22
               ADA
Somers' D                0.94
Goodman-Kruskal Gamma    0.95
Kendall's Tau-a          0.40
```

   i. The measures association for the independent variable rural are about .5--half way between 0 for no correlation and 1.0 for perfect correlation--so the data fit at best moderately well.
      1) Note I prefer using Somer's measure.
   ii. For the ADA variable, however, the value of the measure is nearly 1, which suggests a quite good fit.
   iii. One would based on these considerations conclude that ADA scores better explain and predict votes on assault weapons than percent rural does. Needless to say, this conclusion undercuts the original hypothesis.

X. NEXT TIME:
   A. More on inference for logistic regression

Go to Notes page
Go to Statistics page