SOC 8811 ADVANCED STATISTICS
LECTURE NOTES

LOGISTIC REGRESSION

SPRING 2011

Prof. David Knoke
Sociology Department
909 Social Sciences

(612) 624-6816/4300
knoke001@umn.edu

TABLE OF CONTENTS

I. REVIEW OF MULTIPLE REGRESSION

Population Linear Equation: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + \varepsilon_i$

Assumptions: **BLUE Characteristics (Chapter 8, p. 256-7)**

1. The relationship of dependent to independent variables is linear & correctly specified.
2. All variables are measured without error.
3. Error term properties for single equation:
   - Normally distributed
   - Expected value (mean) of errors = 0
   - Errors independently distributed with constant variances (homoscedasticity)
   - Each predictor is uncorrelated with equation's error term
4. In systems of interrelated equations, errors are uncorrelated across equations.

Sample Prediction Equation: $\hat{Y}_i = a + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_K X_{Ki}$

Estimation Method: **Ordinary Least Squares (Chapter 8, pp. 258-59)**

Beta Coefficient Hypothesis: $H_0\!: \beta_K = 0$

b Coefficient Test:

$$t_{N-K-1} = \frac{b_K - \beta_K}{s_{b_K}}$$

Confidence Interval for b: $b_K \pm s_{b_K} t_{\alpha/2}$

Coefficient of Determination:

$$R^2 = \frac{SS_{REGRESSION}}{SS_{TOTAL}} = \frac{SS_{REGRESSION}}{SS_{REGRESSION} + SS_{ERROR}}$$

R-Square Hypothesis: $H_0\!: \rho^2 = 0$

R-Square Test:

$$F_{K,\,N-K-1} = \frac{SS_{REGRESSION} / K}{SS_{ERROR} / (N-K-1)}$$

R^2 Difference for 2 Eqns:

$$F_{(K_2-K_1),(N-K_2-1)} = \frac{(R_2^2 - R_1^2)/(K_2 - K_1)}{(1 - R_2^2)/(N - K_2 - 1)}$$

EXAMPLE: LINEAR REGRESSION WITH INTERACTION

To illustrate OLS multivariate linear regression with Stata using the 2008 General Social Survey (2008 GSS), I estimate two equations where sexfreq ("About how often did you have sex during last 12 months?") is the dependent variable. It's an seven-category ordered measure from (0) "Not at all" to (6) "More than 3 times a week." I recoded those values into annual frequencies (sexfreq2); see below. Three independent variables are age, gender, and their interaction. I recoded sex into a 1-0 dummy variable, female. Then I computed a variable for the interaction of female and age by multiplying those two variables (femage). Thus, in femage, every man's value = 0, while each woman's value equals her age in years.

The Stata instructions to create the variables used this sequence of dialog boxes and this set of command lines:

```
recode sexfreq (0=0)(1=1.5)(2=12)(3=30)(4=52)(5=130)(6=208),
        generate(sexfreq2) label(sexfreq times per year)
recode sex (2=1)(1=0), generate(female)
generate femage=female*age
codebook sexfreq2 female age femage
```

The variable descriptive statistics:

```
summarize sexfreq2 female age femage
```

Variable	Obs	Mean	Std. Dev.	Min	Max
sexfreq2	1686	51.64858	61.68792	0	208
female	2023	.54078	.49845	0	1
age	2013	47.7084	17.35084	18	89
femage	2013	25.88276	27.31426	0	89

The first OLS regression equation I estimated has only the additive "main effects" of age and female.

regress sexfreq2 age female

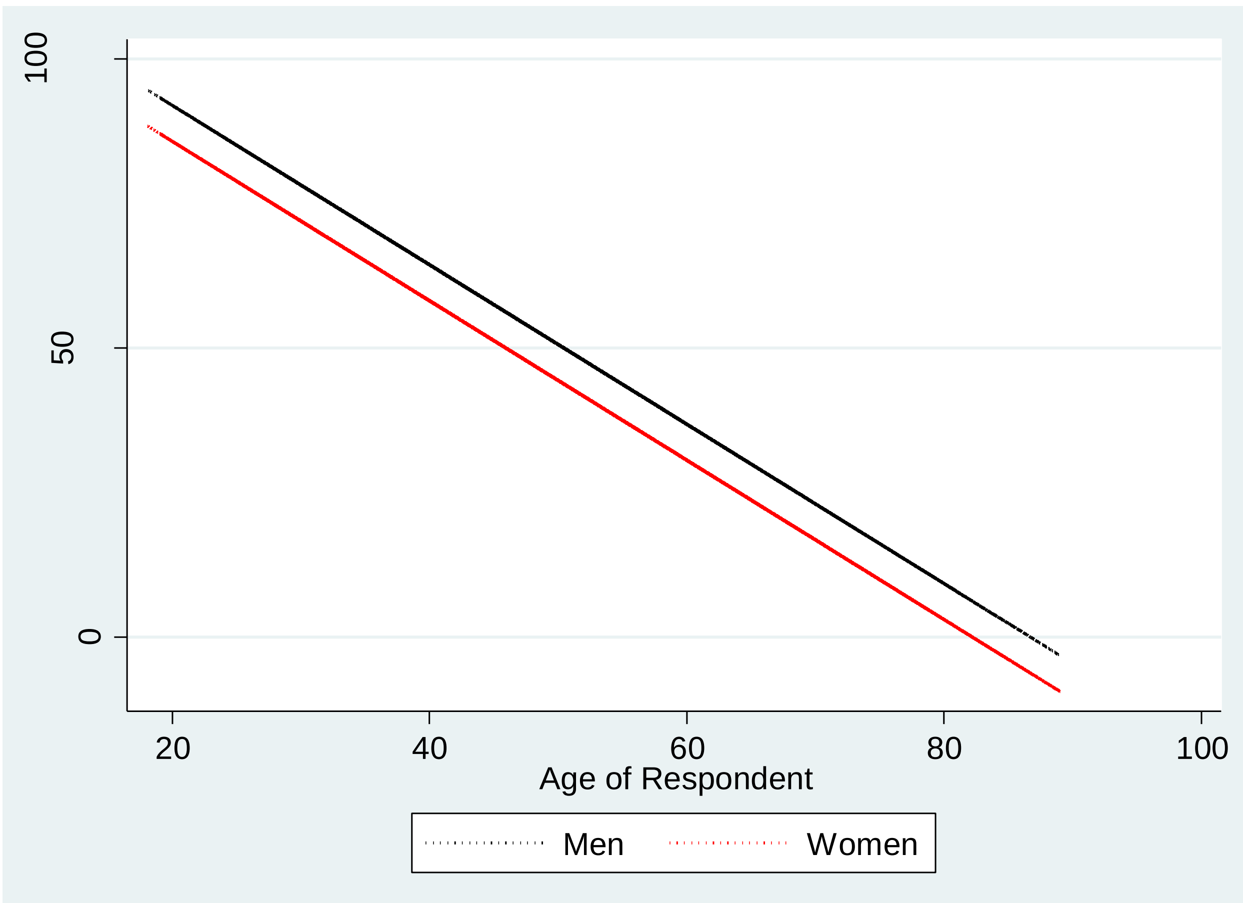
```
      Source |       SS       df       MS              Number of obs =    1680
-------------+------------------------------           F(  2,  1677) =  148.92
       Model |  964467.741     2  482233.871           Prob > F      =  0.0000
    Residual |  5430331.13  1677  3238.12232           R-squared     =  0.1508
-------------+------------------------------           Adj R-squared =  0.1498
       Total |  6394798.87  1679  3808.69498           Root MSE      =  56.905

------------------------------------------------------------------------------
    sexfreq2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -1.379259    .080797   -17.07   0.000    -1.537733   -1.220786
      female |  -6.224432   2.786511    -2.23   0.026    -11.68984   -.7590262
       _cons |   119.5691   4.272797    27.98   0.000     111.1885    127.9497
------------------------------------------------------------------------------
```

The significant negative effect of age indicates that in the population sexual activity declines with age. The negative sign for female means that, relative to men (the omitted reference category), women have less sexual activity than men at all ages (about 6.22 times fewer). Both coefficients are statistically significant at p < .05 or less, so we can infer that these effects probably occur in the population with only a small chance of Type I error (false rejection error).

If the predicted scores for each gender are graphed over the respondents' age range, the two lines are parallel.

```
predict sexmale if female==0
predict sexfemale if female==1
twoway (line sexmale age if female==0, lcolor(black) lpattern(dot)
msymbol(diamond) msize(small)) (line sexfemale age if female==1,
lcolor(red) lpattern(dot) msymbol(circle) msize(small)),
ytitle(Predicted Sexfreq2) xtitle(Age of Respondent) legend(order(1
"Men" 2 "Women"))
```

In this example, women's line is -6.22 units below the men:



The second OLS regression equation includes the femage interaction term:

regress sexfreq2 age female femage

```
      Source |       SS       df       MS              Number of obs =    1680
-------------+------------------------------           F(  3,  1676) =  100.40
       Model |  974157.617     3  324719.206           Prob > F      =  0.0000
    Residual |  5420641.26  1676  3234.27283           R-squared     =  0.1523
-------------+------------------------------           Adj R-squared =  0.1508
       Total |  6394798.87  1679  3808.69498           Root MSE      =  56.871

------------------------------------------------------------------------------
    sexfreq2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -1.214335   .1248968    -9.72   0.000    -1.459305   -.9693646
      female |   6.997664   8.130674     0.86   0.390    -8.949681    22.94501
      femage |  -.2833732   .1637148    -1.73   0.084    -.6044802    .0377338
       _cons |    111.916   6.146892    18.21   0.000     99.85963    123.9724
------------------------------------------------------------------------------
```

The female coefficient changed to a positive sign (+7.00, rounded) after including its interaction with age (-0.28). Now the two predicted lines are no longer parallel:

$$Male: \quad \hat{Y}_i = 111.92 - 1.21\, X_{AGE} + 7.00(0) - 0.28(0)\, X_{AGE}$$

$$\hat{Y}_i = 111.92 - 1.21\, X_{AGE}$$

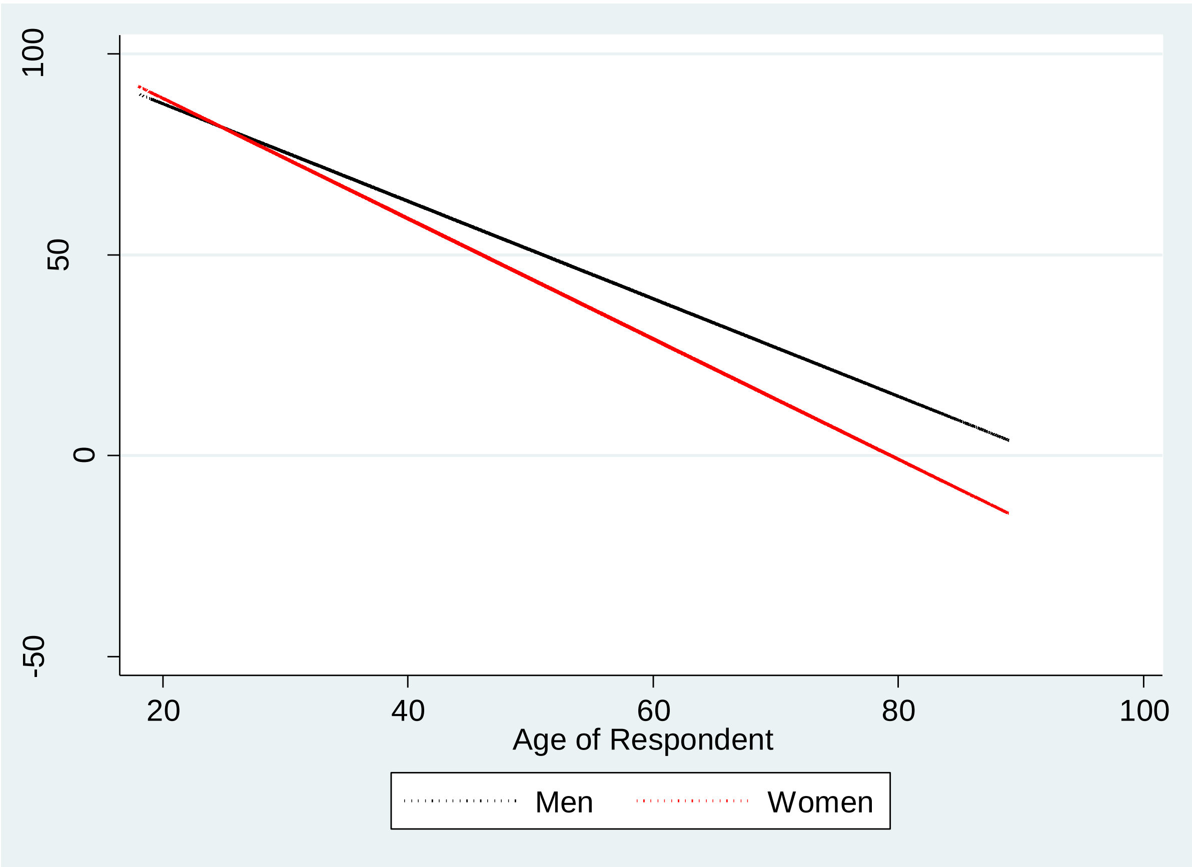
$$Female: \quad \hat{Y}_i = 111.92 - 1.21\, X_{AGE} + 7.00(1) - 0.28(1)\, X_{AGE}$$

$$\hat{Y}_i = 118.92 - 1.50\, X_{AGE}$$

To graph the two lines, recompute the predicted scores for each gender then run the twoway program as before:

```
predict sex_male if female==0
predict sex_female if female==1
twoway (line sex_male age if female==0, lcolor(black) lpattern(dot)
msymbol(diamond) msize(small)) (line sex_female age if female==1,
lcolor(red) lpattern(dot) msymbol(circle) msize(small)),
ytitle(Predicted Sexfreq2) xtitle(Age of Respondent) legend(order(1
"Men" 2 "Women"))
```

The lines are no longer parallel. Although teenage women are slightly more sexually active than men, the rate falls off with age, so the gender gap grows increasingly wider.



The two OLS regression equations still produce straight line relations of sexual activity and age for both genders. We could try other transforms, such as adding a squared age component, to see whether a nonlinear relation occurs.

II. LOGISTIC REGRESSION

Multiple regression assumes a normally distributed continuous dependent variable, and is generally robust for multi-category ordered variables. If the dependent variable is a 1-0 dichotomy, applying the OLS estimation method results in the linear probability model. The estimated regression coefficients predict the expected proportion of cases in the "1" category.

The extended example below analyzes visart: "How many times did you visit an art museum during the last year?" I recode visart into visartd a dichotomy with no visits (0) and 1 or more visits (1); to ensure that any very frequent visitors were included in the latter category, I used a maximum value of 500 for the upper range:

recode visart (0=0)(1/500=1), generate(visartd) label(binary visit art museum)

The frequency distribution of visartd:

table visartd

```
missing .: 519/2023
----------------------------------------------------------
RECODE of visart (how often r visited art museum last year)
          |       Freq.
----------+-----------
        0 |       1,032
        1 |         472
----------------------------------------------------------
```

Now estimate the OLS regression of visartd on educ (5 Rs had missing values):

regress visartd educ

```
Source      |      SS        df      MS           Number of obs =    1501
------------+-------------------------           F(  1,  1499) =  275.87
Model       |   50.23675      1   50.2367        Prob > F      =  0.0000
Residual    |  272.96777   1499    .18209        R-squared     =  0.1554
------------+-------------------------           Adj R-squared =  0.1549
Total       |  323.20453   1500    .21546        Root MSE      =  .42673
-------------------------------------------------------------------------
visartd     |   Coef.    Std. Err.   t     P>|t|   [95% Conf. Interval]
---------+---------------------------------------------------------------
educ        |  .060888   .003665    16.61  0.000    .053698    .0680797
_cons       | -.503488   .050423    -9.99  0.000   -.602389   -.4045757
-------------------------------------------------------------------------
```

If predicted values are computed, the linear probability model may generate expected scores that are less than 0.0 or greater than 1.00, which are nonsensical probabilities.

EX: For educ = 6 years: $\hat{p}_i = -0.503 + .061(6) = -0.137$

Logistic regression is preferable to the linear probability model because it does not require the OLS-BLUE assumption of normally distributed error terms in multiple regression.* Logistic regression does not generate impossible predicted scores because they are bounded between 0 and 1.

The logit transformation of p is defined as a natural logarithm of ratio of two probabilities:

$$\mathbf{L_i} = \ln\left(\frac{\mathbf{p_i}}{1-\mathbf{p_i}}\right) = \log_{\mathbf{e}}\left(\frac{\mathbf{p}_1}{\mathbf{p}_0}\right)$$

Logit is also called log-odds. Because $p_1 + p_0 = 1.00$, so $p_1 = 1 - p_0$

*** From page 298 in *SSDA*:** $e_i = Y_i - \hat{Y}_i$

$e_i = Y_i - (a + bX_i)$

But, Y has only two values (1 or 0); so at every X-value, an error term can have only two scores: $e_i = 1 - (a + bX_i)$

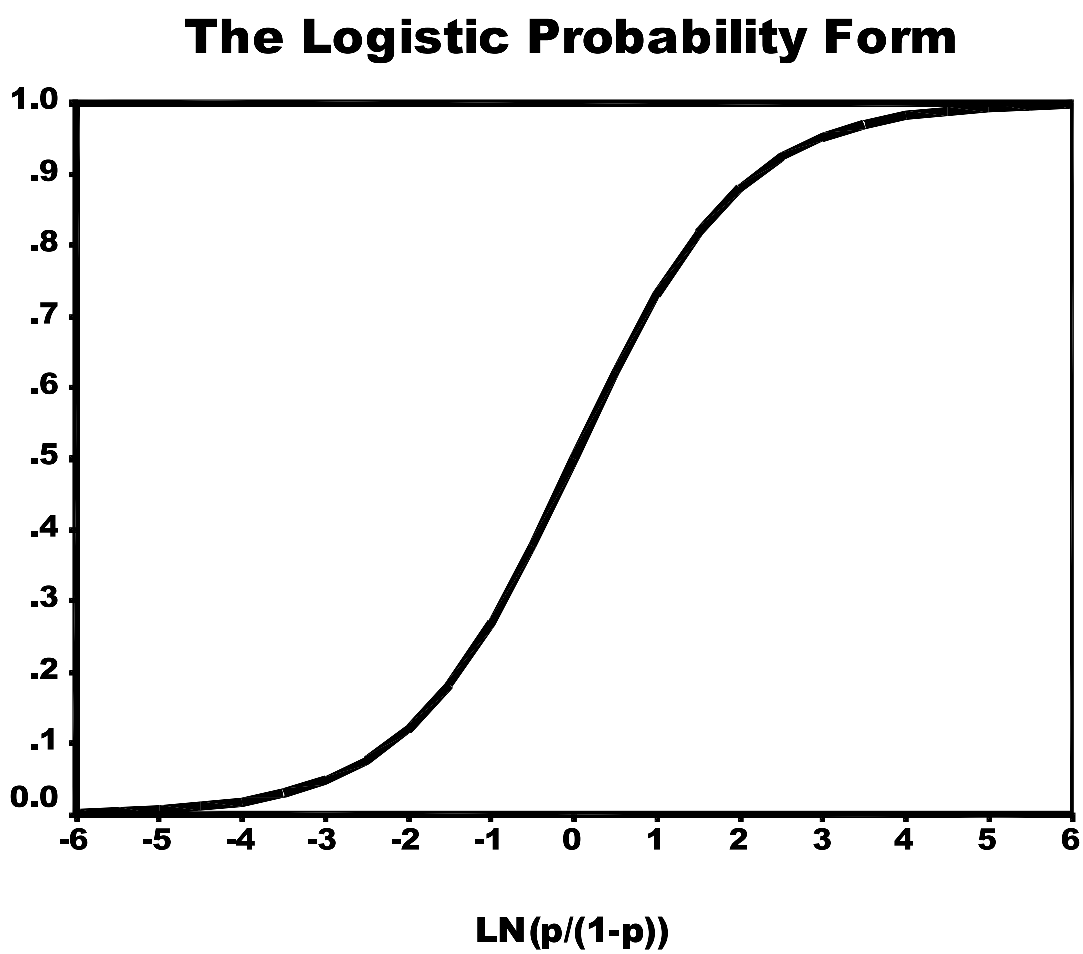
$e_i = 0 - (a + bX_i)$

Hence, the error terms cannot be normally distributed, violating a key assumption in OLS regression. Although unbiased, the linear probability model's coefficients are not efficient (i.e., do not have the smallest possible sampling variances).

The diagram below shows that logit values are symmetrical around p = 0, and roughly linear within the range of (.15 < p < .85). However, as probability approaches either extreme of its range, the logit grows very large or small but never reaches its asymptotic limits of 0 or 1.

EXERCISE: Use your calculator to convert these probabilities into logits:

p = .05	p = .25	p = .50	p = .75	p = .95
L = ____	L = ____	L = ____	L = ____	L = ____



(SOURCE: *SSDA,* 4th Ed., Fig. 9.4, p. 300)

ABOUT LOGARITHMIC TRANSFORMATIONS

Recall from school that using logarithms is a convenient way to do multiplication and division by simply adding or subtracting the logs of numbers. Contrariwise, taking the anti-log of a logarithm is called exponentiation, which restores the original Arabic numerical value with which you began. Logs and exponents are VERY important for understanding logistic regression and event history analysis.

EX: For base of 10: 100 X 1,000 = 10^2 X 10^3 = 10^{2+3} = 10^5 = 100,000

First key: $\log_{10} 100 + \log_{10} 1{,}000 = 2 + 3 = 5$

Second key: antilog$_{10}$ 5 = exp^5 = 10^5 =100,000

Statistics more commonly uses the natural logarithm, which has Napier's constant (e = 2.71828...) as the base. Two symbols for the natural log are "ln X" and "$\log_e$ X". Two notations for exponentiation where Napier's constant is the base are "exp" or just "*e*".

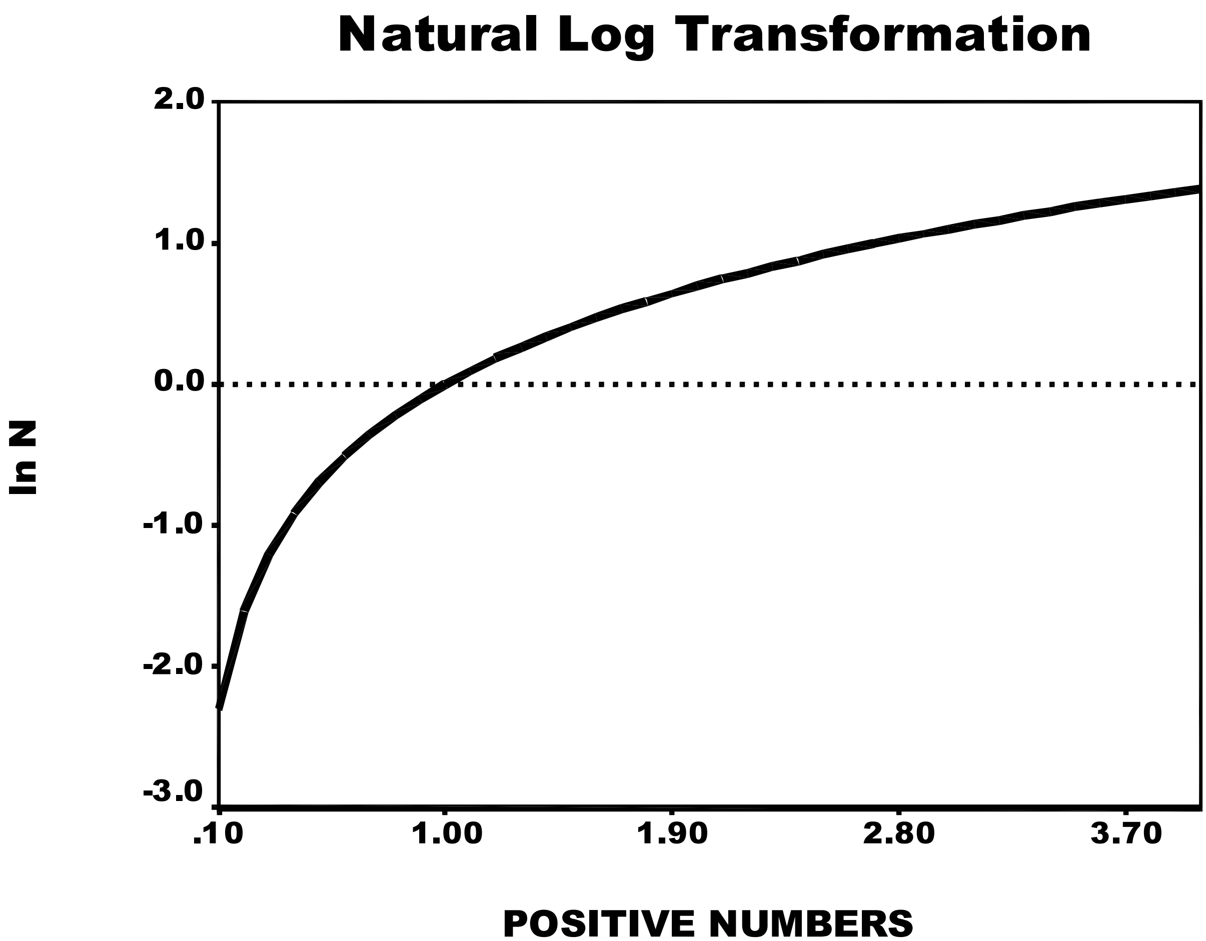
Some important features of next figure showing N transformed into ln N:

1. Only positive values of N strictly greater than 0 can be changed into logarithms; ln 0 is undefined.

2. Values of N between zero and 1 have negative logarithms; as N approaches 0, ln N approaches negative infinity at an accelerating rate. Despite the figure, the curve never touches the Y-axis (where X = 0).

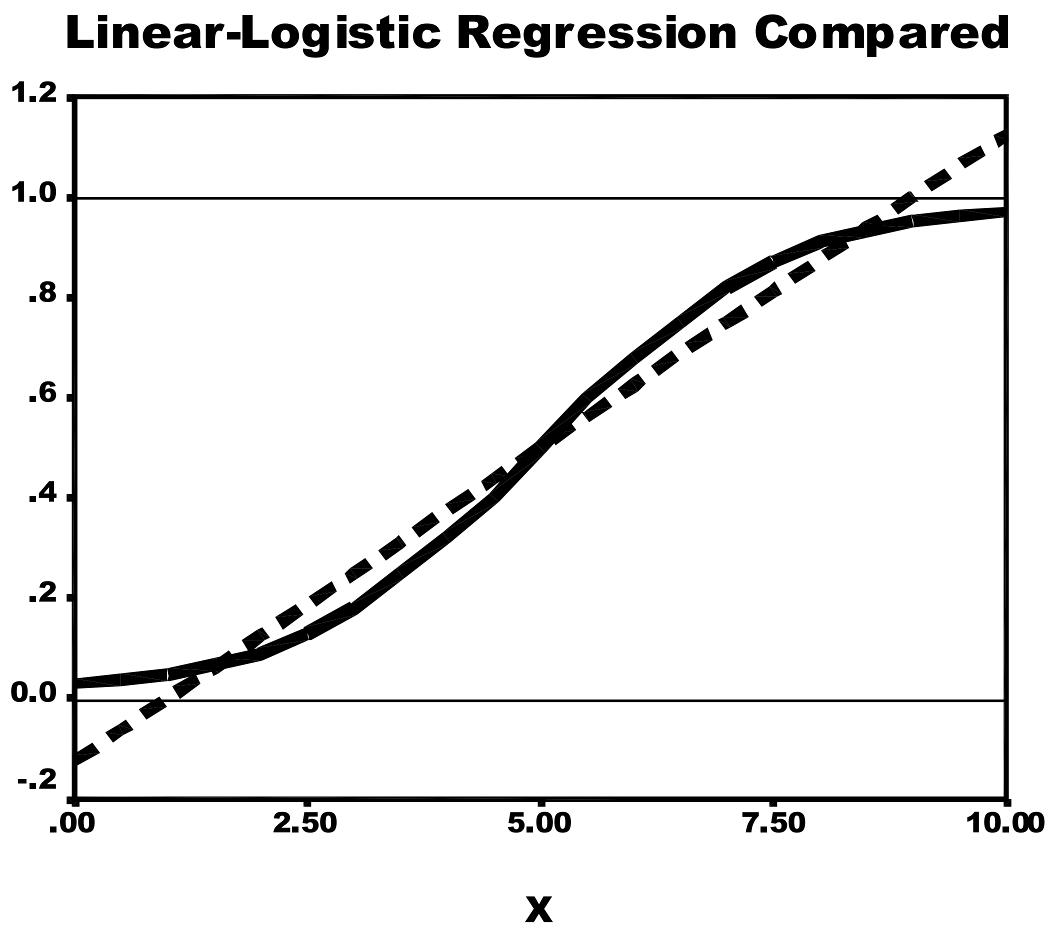
3. ln 1 = 0. Because logging and exponentiating are reverse operations, taking the log of an exponentiated term cancels that operation: ln (e^X) = X. Recall that any number taken to the "0th" power is 1; for example, 2^2 = 4, 2^1 = 2, and 2^0 = 1. Therefore, e^0 = 1. By substitution, ln e^0 = ln 1. But ln e^0 = 0. Therefore, ln 1 = 0. Use your calculator to verify these facts.

4. The ln N for values of N greater than 1 are positive and approach positive infinity at a decreasing rate. That is, unit changes of ln N are smaller as N increases.


Natural Log Transformation
2.0
1.0
0.0
-1.0
-2.0
-3.0
ln N
.10
1.00
1.90
2.80
3.70
POSITIVE NUMBERS

The diagram below illustrates schematically how the linear and the logistic regression lines differ for the same data. The lines very close in the middle-range of the probability scale (from p = .25 to .75), but depart widely at the extreme upper- and lower-ranges. Observe how the logistic regression predictions stay within the 0-1 probability bounds, but the linear probability model may predict probabilities that are negative or larger than 1.0!

Although the dichotomous dependent variable Y takes only two observed values (0 and 1), the <u>expected values</u> $(\hat{Y}_i)$ **calculated from either regression equation fall across the full range between 0 and 1.**



ODDS & LOGITS: OBAMA JOB RATINGS

As an exercise, convert these percentages in Presidential job-approval polls, all conducted during October, 2010, into odds and logits (log-odds):

POLL*	% Approve	% Disapprove	Odds	Logits
Newsweek	54	40		
ABC/Washington Post	50	45		
AP-GfK	49	50		
McClatchy-Marist RV	48	43		
NBC/Wall Street Journal RV	47	49		
Bloomberg LV	47	48		
Battleground LV	46	51		
CNN/ORC	46	51		
Pew	46	45		
CNN/ORC	45	52		
Ipsos/Reuters	45	51		
NBC/Wall Street Journal RV	45	50		
USA Today/Gallup	45	49		
CBS/New York Times	45	47		
CBS	44	45		
Ipsos/Reuters	43	53		
FOX/OD RV	43	47		
FOX/OD RV	41	50		

Variations on “"Do you approve or disapprove of the way President Bush is handling his job as president?” Most *N*s are 800-1,000 respondents.
Approve +Disapprove do not sum to 100% due to omitted “don’t know”, “mixed feelings”, “not familiar”, “note sure”, etc. responses.

* RV = Registered voters   LV = Likely voters

SOURCE: Polling Report.com
<http://www.pollingreport.com/obama_job.htm>

NATURAL LOGARITHMS

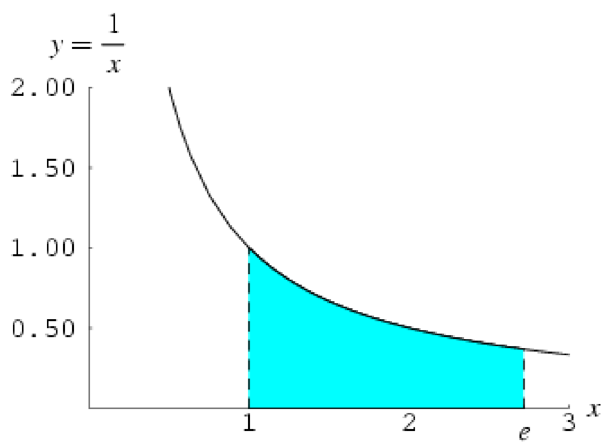
The natural logarithm (ln x), invented by John Napier (1550-1617), is the logarithm with base e, where e = 2.718281828….

The ln x function is referred to as *natural* because, unlike other logarithms, it can be defined using a simple integral or Taylor series. In mathematics, expressions with an unknown variable as a function of the exponent e occur much more often than exponents of 10 (the "natural" properties of the exponential function provide a better description of growth and decay).

This function can be defined

$$\ln x \equiv \int_1^x \frac{dt}{t} \qquad (2)$$

for $x > 0$.



The integral of ln x, from x = 1 to x = e is the shaded area under the hyperbola y=1/x, which has area = 1 (unit area). ln x is very useful for calculus and statistics because its derivative is:

$$\frac{d}{d\,x}\ln x = \frac{1}{x}$$

In contrast, the derivatives of logarithms using other bases (b), such as base = 10, are more complicated:

$$\frac{d}{d\,x}\ln_b x = \frac{1}{x \ln b}$$

SOURCES: Wikipedia.org; mathworld.wolfram.com

DICHOTOMOUS LOGISTIC REGRESSION

As in OLS regression, a prediction equation for logistic regression specifies the expected logit as a linear additive function of one or more independent variables:

$$\hat{\mathbf{L}}_{\mathbf{i}} = \mathbf{a} + \mathbf{b}_1\mathbf{X}_{1\mathbf{i}} + \mathbf{b}_2\mathbf{X}_{2\mathbf{i}} + \ldots + \mathbf{b}_{\mathbf{K}}\mathbf{X}_{\mathbf{Ki}}$$

Anti-logging (exponentiating) both sides of the equation, yields this result (the two forms of notation on the right-hand sides are equivalent):

$$\exp \hat{\mathbf{L}}_{\mathbf{i}} = \exp(\mathbf{a} + \mathbf{b}_1\mathbf{X}_{1\mathbf{i}} + \mathbf{b}_2\mathbf{X}_{2\mathbf{i}} + \ldots + \mathbf{b}_{\mathbf{K}}\mathbf{X}_{\mathbf{Ki}})$$

$$\exp\left(\ln\left(\frac{\mathbf{p}_{\mathbf{i}}}{1-\mathbf{p}_{\mathbf{i}}}\right)\right) = \mathbf{e}^{\mathbf{a}+\mathbf{b}_1\mathbf{X}_{1\mathbf{i}}+\mathbf{b}_2\mathbf{X}_{2\mathbf{i}}+\ldots+\mathbf{b}_{\mathbf{K}}\mathbf{X}_{\mathbf{Ki}}}$$

Note that by exponentiating a logarithm, these two inverse operations essentially cancel one another, yielding the odds on the left-hand side of the equation:

$$\frac{\mathbf{p}_{\mathbf{i}}}{1-\mathbf{p}_{\mathbf{i}}} = \mathbf{e}^{\mathbf{a}+\mathbf{b}_1\mathbf{X}_1+\mathbf{b}_2\mathbf{X}_2+\ldots+\mathbf{b}_{\mathbf{K}}\mathbf{X}_{\mathbf{K}}}$$

or, equivalently, where Pr(Y=1) means “probability that Y equals 1”:

$$\frac{\Pr(\mathbf{Y}=1)}{\Pr(\mathbf{Y}=0)} = \mathbf{e}^{\mathbf{a}}\mathbf{e}^{\mathbf{b}_1\mathbf{X}_1}\ldots\mathbf{e}^{\mathbf{b}_2\mathbf{X}_2}\mathbf{e}^{\mathbf{b}_{\mathbf{K}}\mathbf{X}_{\mathbf{K}}}$$

Again the two right-hand side expressions are equal; recall that $e^a e^b = e^{a+b}$.

Thus, logistic regression coefficients tell how the dependent variable's expected <u>log odds</u> change with unit differences in the predictors. More below on their interpretation.

Estimation Method

OLS techniques, which estimate the unknown population regression coefficients by minimizing the sum of squared errors, do not suffice for the logistic model.

Instead, statisticians use <u>maximum likelihood estimation</u> (MLE) methods. Because no general closed-form solution exists, computer programs for MLE use an iterative procedure to generate parameter estimates:

1. Start with an initial set of estimates (e.g., use OLS).

2. Successively revise the parameter estimates, finding new values that maximize the joint probability density function (likelihood function) of observing the dependent variable values that were actually sampled.

MLE involves maximizing a <u>log-likelihood function</u>, whose core is this negative expression:

$$-\sum_{\mathbf{i}=1}^{\mathbf{N}}(\mathbf{Y}_{\mathbf{i}}-\boldsymbol{\mu})^2$$

> **where <u>Y</u> is the dependent variable and μ is the central tendency of the parameter distribution. NOTE: the negative sign produces a parabola that opens downward: the 2nd derivative thus identifies the location of the parameter's <u>maximum</u> value.**

We return below to the log-likelihood function when assessing how well an equation "fits" the data.

3. Stop iterating when "peak" values are obtained (i.e., the local maximum), as indicated by calculus (the point at which the equation's first derivative equals zero). That is, quit when no further increase in the MLE occurs.

For details on MLE principles and procedures, see Scott R. Eliason. 1993. *Maximum Likelihood Estimation: Logic & Practice*. Newbury Park, CA: Sage Publications.

Now re-estimate the art museum-education relationship with Stata's binary logistic regression. In Stata, two logistic commands must be submitted to produce both sets of output:

logistic visartd educ, coef

```
Logistic regression                          Number of obs   =       1501
                                             LR chi2(1)      =     264.19
                                             Prob > chi2     =     0.0000
Log likelihood = -801.67728                  Pseudo R2       =     0.1415
------------------------------------------------------------------------
 visartd |  Coef.     Std. Err.    z    P>|z|   [95% Conf. Interval]
---------+--------------------------------------------------------------
    educ |   .3552712  .0247871  14.33  0.00     .306689   .403853
   _cons | -5.726947   .3586061 -15.97  0.00   -6.429802 -5.024092
------------------------------------------------------------------------
```

logistic visartd educ

```
Logistic regression                          Number of obs   =       1501
                                             LR chi2(1)      =     264.19
                                             Prob > chi2     =     0.0000
Log likelihood = -801.67728                  Pseudo R2       =     0.1415
------------------------------------------------------------------------
visartd | Odds Ratio  Std. Err.    z    P>|z|   [95% Conf. Interval]
+-----------------------------------------------------------------------
educ    | 1.426567   .0353605   14.33  0.000   1.358919  1.497584
------------------------------------------------------------------------
```

Using Stata's logit command produces the same coefficient output, but also this Iteration history, some of which is used below in measuring equation fit (page 31):

logit visartd educ

```
Iteration 0:   log likelihood = -933.77247
Iteration 1:   log likelihood = -805.81088
Iteration 2:   log likelihood = -801.68997
Iteration 3:   log likelihood = -801.67728
Iteration 4:   log likelihood = -801.67728
```

INTERPRETING LOGISTIC COEFFICIENTS

Effects are nonlinear in the independent and dependent variables, but are linear in their logs.

The estimate b = 0.355 means that, for each year of educ, the expected logit of attending an art museum increases by 0.355. This number is not at all enlightening. Let's calculate some expected logits. For a person with 6 years of education:

$$\hat{L}_6 = -5.727 + 0.355(6) = -3.597$$

The negative sign shows the expected value favors that person having NOT visited an art museum. Why is this interpretation correct? (HINT: What two categories form the ratio of probabilities whose ln is taken?)
Compare this logistic regression prediction to the linear probability model's prediction of –0.137 (see page 8).

For someone with 20 years of schooling:

$$\hat{L}_{20} = -5.727 + 0.355(20) = +1.373$$

The positive sign shows that such persons were more likely to visit than not to visit. Again, why?

These two calculations reveal little beyond showing that the log-odds of visiting an art museum for the more educated person are expected to be much higher than log-odds for the less-educated. The meaning of these numerical magnitudes cannot be grasped. What we require is a measuring rod with which we are more familiar. So, translate each log-odds into an <u>expected probability</u>, using this generic formula:

$$\hat{p}_i = \frac{e^Z}{1 + e^Z}$$

where Z is a short notation standing for the entire right-hand side of the logistic regression equation: $(a + b_1X_1 + ... + b_kX_k)$.

Here's a step-by-step derivation, starting from the definition of the logit (page 8) and the logistic regression equation (page 15):

$$\mathbf{L_i} = \ln\left(\frac{\mathbf{p_i}}{1-\mathbf{p_i}}\right) = \mathbf{Z} \quad \textbf{where } \mathbf{Z} = \mathbf{a} + \Sigma \mathbf{b_i X_i}$$

Exponentiate both sides:

$$\exp\ \ln\left(\frac{\mathbf{p_i}}{1-\mathbf{p_i}}\right) = \exp\ \mathbf{Z}$$

$$\left(\frac{\mathbf{p_i}}{1-\mathbf{p_i}}\right) = \exp\ \mathbf{Z}$$

$$\mathbf{p_i} = (\exp\ \mathbf{Z})(1-\mathbf{p_i})$$

$$\mathbf{p_i} = \exp\ \mathbf{Z} - \mathbf{p_i} \exp\ \mathbf{Z}$$

$$\mathbf{p_i} + \mathbf{p_i} \exp\ \mathbf{Z} = \exp\ \mathbf{Z}$$

$$\mathbf{p_i}(1 + \exp\ \mathbf{Z}) = \exp\ \mathbf{Z}$$

$$\mathbf{p_i} = \frac{\exp\ \mathbf{Z}}{1 + \exp\ \mathbf{Z}}$$

Now replace Z in the last equation with the full expression to show the expected probability of a category 1 response:

$$\mathbf{p_{Y=1}} = \frac{\mathbf{e}^{\mathbf{a}+\mathbf{b_1 X_1}+\mathbf{b_2 X_2}+\ldots+\mathbf{b_K X_K}}}{1+\mathbf{e}^{\mathbf{a}+\mathbf{b_1 X_1}+\mathbf{b_2 X_2}+\ldots+\mathbf{b_K X_K}}}$$

Why will all three terms <u>always</u> have positive values? Therefore, the expected probability that Y = 1 can never be 0 or lower. Because the denominator must always be larger than the numerator (again, why?), the expected probability can never be 1 or greater. Thus, the probabilities predicted by a logistic regression equation are confined inside the range between 0 and 1.

Next, consider two expected probabilities, obtained by substituting our previous example education into the equation:

For a person with educ = 6 years of schooling, the predicted probability is:

$$p_{Y6=1} = \frac{e^{-5.727+0.355(6)}}{1+e^{-5.727+0.355(6)}}$$

$$p_{Y6=1} = \frac{e^{-3.597}}{1+e^{-3.597}} = \frac{0.0274}{1+0.0274} = 0.027$$

For someone with educ = 20, the expected probability that Y = 1 is:

$$p_{Y20=1} = \frac{e^{-5.727+0.355(20)}}{1+e^{-5.727+0.355(20)}}$$

$$p_{Y20=1} = \frac{e^{1.327}}{1+e^{1.327}} = \frac{3.770}{1+3.770} = 0.79$$

Where are these two probability values located on the logistic probability curve in the figure on page 9?

EXPONENTIATING LOGISTIC REGRESSION B's

A multiplicative transformation of the logistic regression coefficients is also insightful. In the column headed "Odds Ratio" Stata displays the exponentiated values of the logit coefficient. (Verify on your calculator that the transformation of B = 0.355 into Odds Ratio = 1.310 by using the e^B key on your hand calculator; that is, the inverse of the "LN" key.) An exponentiated value reveals the percentage change in the expected odds of the dependent variable for a one-unit change in the independent variable. This transformation involves exponentiating both sides of the basic logistic regression equation. For example, the expected logit equation is:

$$\hat{L}_i = -5.727 + 0.355\,X_i$$

Exponentiating both sides (see page 15) gives:

$$\frac{p_1}{p_0} = e^{-5.727+0.355X_i}$$

Finally, plug in educ = 6 years to obtain the expected odds for X:

$$\frac{p_1}{p_0} = e^{-5.727+0.355(6)} = e^{-3.597} = 0.0274$$

On page 20 we saw that the expected probability of visiting an art museum by someone with 6 years of schooling is .027 (more, precisely it's 0.0266693). This probability exactly corresponds to the expected odds of (p_1 / (1 - p_1)) = p_1 / p_0 = (.0266693 / .9733307) = 0.0274, which is exactly what the exponentiation above shows.

If you exponentiate the estimated B for educ (found in the Coef. column of the Stata output on page 17), you obtain the value for educ in the Odds Ratio column:

[exp(0.3552712) = 1.4265675]

which shows that the expected odds of a museum visit increases by (1.427 - 1)(100%) = 42.7% per year of education.

The generic transformation formula is: (exp(B) - 1)(100%)
In Stata's output notation: ((Odds Ratio) - 1)(100%)

Equivalently, to apply the exponentiated coefficient to calculate the changing odds from one year to the next, simply multiply the preceding year's odds by 1.427 to obtain the next year's odds. To illustrate, what is the expected odds of visiting an art museum by a person with 7 years of schooling? Simply (0.0274)(1.427) = 0.039. For a someone with 8 years: (0.039)(1.427) = 0.056, etc. Each successive year of education increases the expected odds value by 42.7% across the entire EDUC range from 0 to 20!

Because odds have no upper limit, expected values can increase indefinitely (unlike a probability, which is bounded from 0 to 1). Note that, as we continue multiplying successive values by any constant amount (e.g., by Exp(B)), the cumulative increases compound exponentially. That curve is similar to what happens to your bank balance when the interest rate remains constant.

The process works in reverse. If a logistic regression b-coefficient has a negative sign (e.g., decreasing the log-odds of attendance), then exponentiating this coefficient will produce a transformed value less than 1. Suppose that $e^{-0.355}$ = 0.701. Thus, a year of schooling would reduce the odds of visiting by (0.701 -1)(100%) = -29.9%. For every additional year of educ, the odds would decrease by another –29.9%.

If a predictor has no impact on the dependent variable, its B = 0 and thus e^{0} = 1.00. Then (exp(0)-1)(100%) = (1-1)(100%) = 0%. Hence, the expected odds of visiting an art museum change by 0.0% as the predictor changes.

In a multiple logistic regression equation, these exponentiated values facilitate comparisons of the net effects of several predictors that are measured using different scales.

LINEAR vs NONLINEAR FORMS of LOGISTIC REGRESSION EQUATIONS

The two forms of the logistic regression parameter express an independent variable's effect on one of two measures of the dependent dichotomy, the logit or the odds. These measures are functions of one another other, via natural logarithm or exponentiation transformation. This section uses the museum visit logistic regression equation to illustrate the basic equivalence of logistic regression parameters in their additive and multiplicative forms.

Here is the additive equation for the expected logit (natural log of the odds):

$$\hat{L}_i = -5.727 + 0.355\, X_i$$

Exponentiate both sides, changing it into a multiplicative equation for the expected odds. The right-most expression is based on the calculation rule for multiplying powers of the same base (e.g., $2^{2+3} = 2^2\, 2^3 = 32$):

$$\frac{p_1}{p_0} = e^{-5.727+0.355X_1} = e^{-5.727}\, e^{0.355X_1}$$

Now compute the expected logits and odds for some education levels. The table below uses the full equations only to compute the initial values (for persons with 0 years of educ). For each succeeding year, I changed the immediately preceding logit or odds by either the respective additive or multiplicative increment.

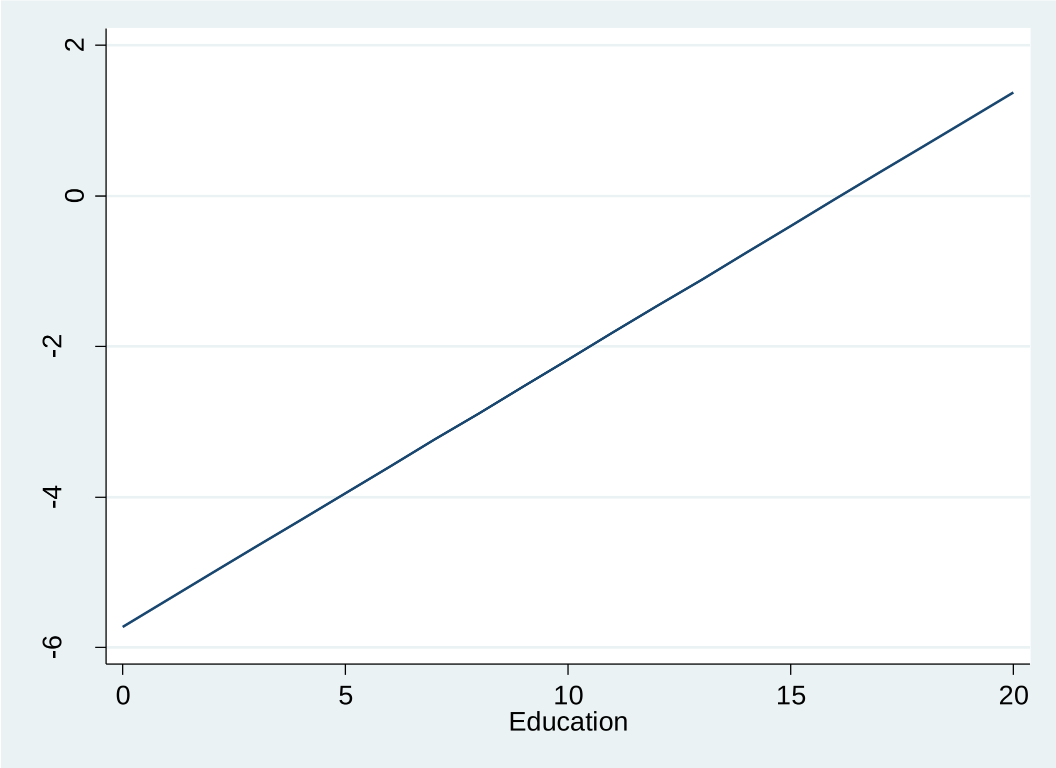
This table shows that (1) the logit changes by a constant amount (linearity in the logs); (2) the odds change by a constant proportion (multiplicity in the odds; recall that $e^{0.270}$ = 1.427.)

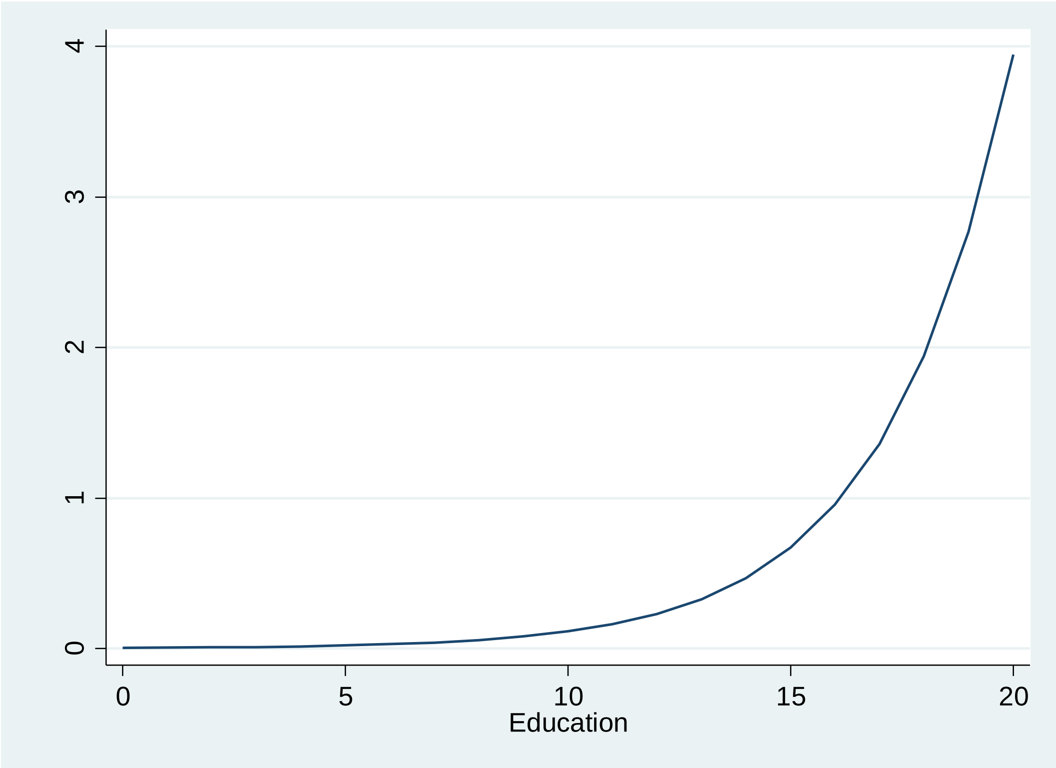
educ	$\hat{L}_i = -5.727 + 0.355 X_i$	$\frac{p_1}{p_0} = e^{-5.727} e^{0.355 X_1}$
0	**-5.727 + 0.355(0) = -5.727**	**$e^{-5.727}$ = 0.00326 and $e^{0.355(0)}$ = 1.0: (0.00326)(1.0) = 0.00326**
1	**+0.355 = -5.372**	*** (1.427) = 0.00465**
2	**+0.355 = -5.017**	*** (1.427) = 0.00663**
3	**+0.355 = -4.662**	*** (1.427) = 0.00947**
16	**+0.355 = -0.047**	*** (1.427) = 0.95502**
20	**+0.355 = +1.373**	*** (1.427) = 3.95101**

Use your calculator to show that the paired values in each row are equivalent (i.e., take their antilogs and natural logs, respectively), within rounding error.

Graphs of these values on the next page reveal that the logit equation forms a straight line while the odds equation forms a nonlinear curve, reflecting their parameters' respective additive and multiplicative relationships with educ. Exponentiating the first graph produces the second plot; logging the second figure yields the first diagram.

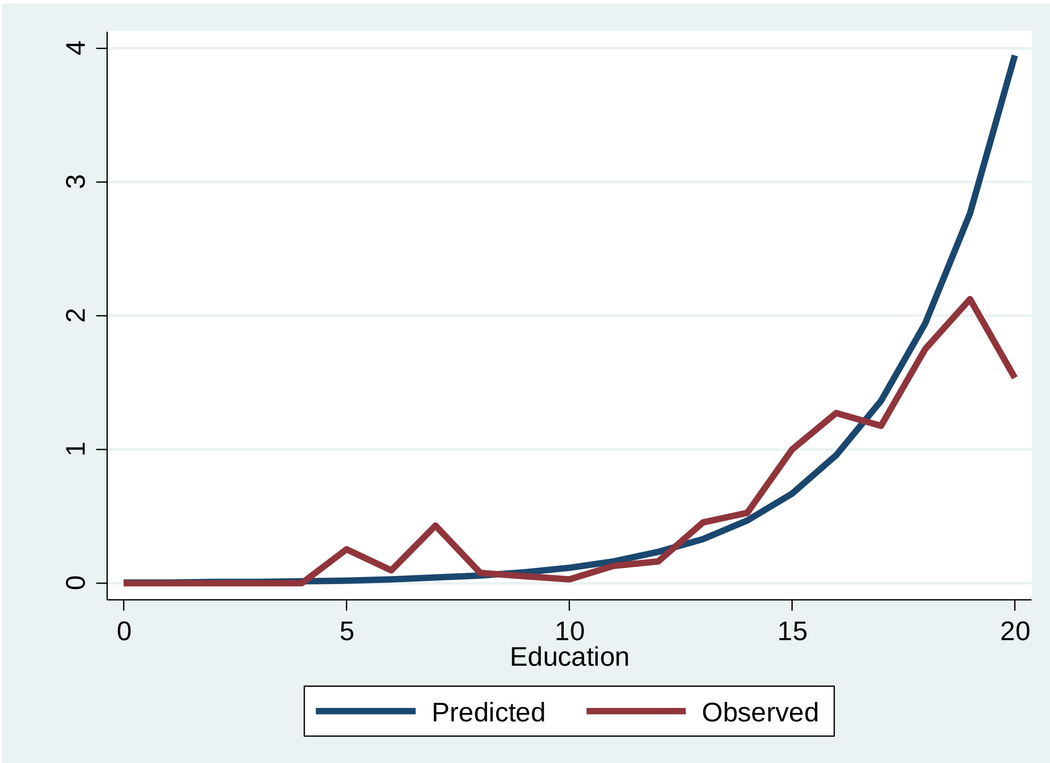
```
generate logit_visartd = -5.727 + 0.355*educ
twoway (line logit_visartd educ), ytitle(Value of LOGIT)
        xtitle(Education)

generate odds_visartd = exp(-5.727 + 0.355*educ)
twoway (line odds_visartd educ), ytitle(Value of ODDS)
        xtitle(Education)
```


2
0
-2
-4
-6
0
5
10
15
20
Education


4
3
2
1
0
0
5
10
15
20
Education

This final graph plots both the predicted and observed odds of visiting an art museum for each year of educ. I created a small dataset with 21 years of educ (from 0 to 20); prodds, the predicted odds from the logistic regression; and obsodds, the observed odds calculated as the ratio of frequencies in the two categories of visartd (visit divided by novisit).

twoway (line prodds educ) (line obsodds educ), ytitle(Value of ODDS) xtitle(Education) legend(order(1 "Predicted" 2 "Observed"))



How closely do you think the predicted values fit the observed data?

THE CHI-SQUARE TEST STATISTIC

To assess the overall equation fit and, for some programs such as SPSS, the significance of individual logistic regression coefficients requires familiarity with the "family" of chi-square distributions χ^2. As explained in *SSDA* (Section 3.11, pp. 102-104), a specific chi-square distribution is constructed from a normally distributed population by drawing a random sample of N cases, changing their observed scores into squared standardized (Z score) values, then summing them:

$$\chi^2_N = \sum_{i=1}^{N} Z_i^2 = \sum_{i=1}^{N} \frac{(Y_i - \mu_Y)^2}{\sigma_Y^2}$$

Although Z scores range from negative to positive values, the squaring eliminates any zero values from the chi-square distribution. In effect, the probability density function for a chi-square resembles a sum of N "folded over" normal distributions.

Different chi-square distributions result from choosing differing Ns. Associated with each chi-square distribution is its degrees of freedom (df), symbolized by the Greek letter ν (nu). The df for a given chi-square equals N, the number of observations in the sample used to construct that distribution. Note my practice of subscripting the χ^2 with its df.

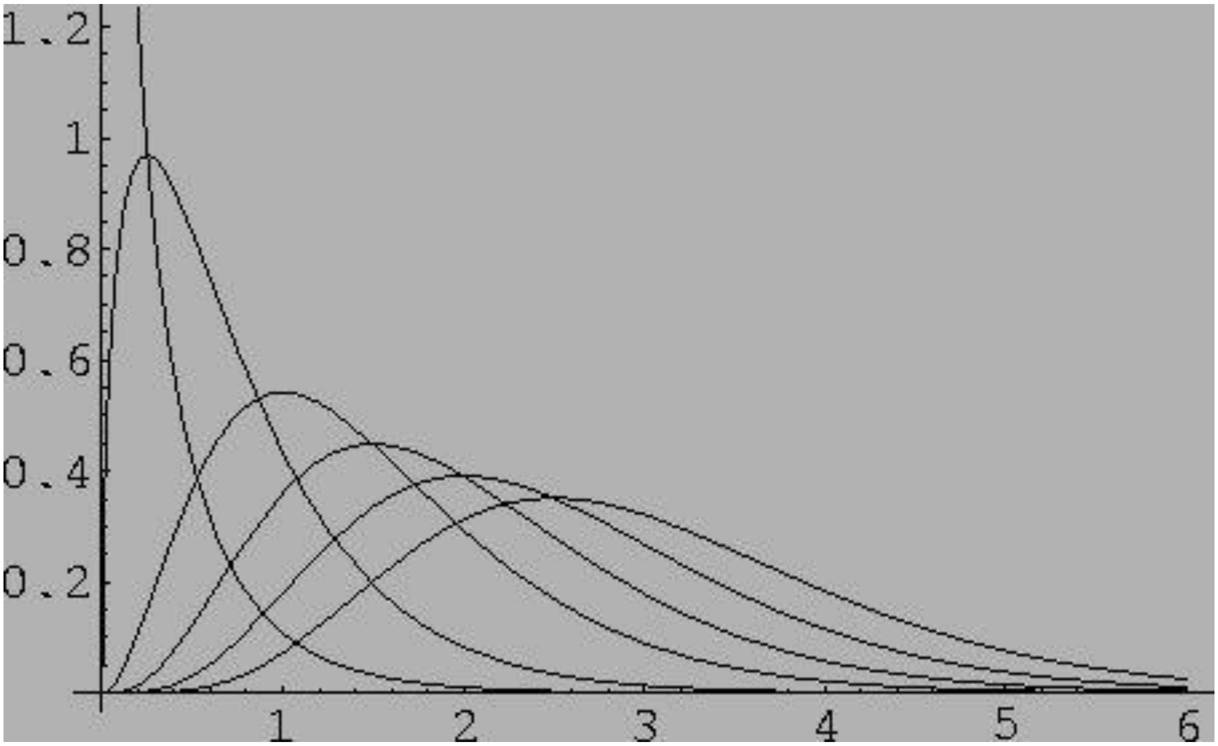
The figure on page 28 shows several chi-square distributions with different dfs. Note that as df increases, the distribution becomes more symmetrical and bell-shaped. The expected value ($E(\chi^2)$) for a given chi-square distribution is ν, its df. The variance of the distribution is 2ν. Appendix B in *SSDA#4* (pp. 457-8) displays critical values of chi-square for several distributions at conventional α-levels.

If you're stranded on a desert island without your solar powered computer, approximate critical values for large df (*N* > 100) as:

$$\chi^2_\alpha = \nu \left(1 - \frac{2}{9\nu} + Z_\alpha \sqrt{\frac{2}{9\nu}}\right)^3$$

where Z_α is the score that puts the entire alpha region into the right tail of the standardized normal distribution.

Note that, like the F-ratio, the χ^2 region of rejection falls entirely into the right-hand tail of the distribution, since the squaring eliminates the negative values of Z. On *SSDA* page 103 we show that $F_{1,\nu} = t^2_\nu$, which means that, for a large N sample, $F = Z^2$. Similarly, because F is the ratio of two chi-squares (p. 116), $\chi^2 = Z^2$. Thus, at α = .05, the critical values for $\chi^2_1 = 3.8414$ and $F_{1,\infty} = 3.8414$, while the critical values for $Z^2_{\alpha/2} = (\pm 1.96)^2 = 3.8414$ and for $t^2_{\infty;\alpha/2} = (\pm 1.96)^2 = 3.8414$. These test statistics are all cousins in the normal distribution family.



Testing the Logistic Regression Estimate of β:

As in OLS linear regression, so in logistic regression with sample data a typical null hypothesis is that the parameter (β) is zero in the population. A two-tailed research hypothesis is:

$$\mathbf{H}_0 : \boldsymbol{\beta}_{\mathbf{K}} = 0$$
$$\mathbf{H}_1 : \boldsymbol{\beta}_{\mathbf{K}} \neq 0$$

To help you decide whether to reject the null hypothesis, H_0, Stata's logit output displays the t-test statistic and reports the probability of making a Type I error (false rejection error) if your reject the null hypothesis in favor of the two-tailed alternative, H_1. It also calculates the 95% confidence interval around the sample point estimate. Here's the t-test statistic for the educ coefficient:

$$t = \frac{b_k - \beta_k}{s_{b_k}} = \frac{0.355 - 0}{0.025} \cong 14.33$$

which can easily be rejected at p < .001. (Stata's output truncates the p-level; it's not equal to zero but your risk of making a false rejection decision is surely much less than one chance in a thousand). Although the Stata labels the column a z-test, it's identical to a t-test statistic for large samples. The convention when reporting results is to call them t-tests.

For some reason, many logistic regression programs (such as SPSS) don't report t-test statistics but instead calculate "Wald statistics." Wald is distributed as a chi-square variable with one degree of freedom:

$$Wald = \left(\frac{b_k - \beta_k}{s_{b_k}} \right)^2$$

Note that this formula is simply the square of the usual t-test above, consistent with the observation on page 28 that $\chi^2 = Z^2 = t^2$ for large N.

Regressing visartd on educ would yield a Wald test statistic of $(0.3552712/0.0247871)^2 = 205.4322$; its square root is 14.33, which is the t-test reported on both outputs on page 17. The probability of a false rejection error is exceedingly small ($p < .001$), hence the 2008 GSS sample statistic b_k is very unlikely to come from a sampling distribution with a population parameter of $\beta_k = 0$. Thus, your decision to reject H_0 runs only a teensy-tiny risk of making a Type I (false-rejection) error. What substantive conclusion do you draw about the art museum-education relationship from the logistic regression analysis of the 2008 GSS data?

MEASURING EQUATION FIT

Several descriptive and inferential statistics are available to assess how well a logistic regression equation (model) fits the data. Stata produces some of these measures, either as part of the logistic output or via additional commands.

1. LOG-LIKELIHOOD RATIO

As noted above, ML estimation of the logistic regression parameters maximizes the equation's log-likelihood (LL) function. Its numerical value is always <u>negative</u>, because the function to be maximized is an inverted parabola in hyper-space whose largest value lies below 0 on the vertical (dependent variable) axis. Computer packages differ in reporting either this negative log-likelihood value, or minus twice the value (-2LL), which has distributional properties enabling application of chi-square distributions.

The log-likelihood value is not used in isolation, but always in comparison to an alternative equation specification. A pair of multivariate equations are said to be <u>nested equations</u> if all the parameters included in the first equation also appear in the second equation (i.e., the first is "nested inside" the second). The difference in -2LL's for a pair of nested equations tests whether the additional parameters specified in the second equation improve its fit to the data over the first equation's fit. We seek to reject the null hypothesis that adding predictors to the second equation does not reduce the size of the –2LL relative to the difference in degrees of freedom:

$$H_0: (-2LL_1) - (-2LL_2) = 0$$
$$H_1: (-2LL_1) - (-2LL_2) > 0$$

The <u>log-likelihood ratio</u> for comparing two nested equations is:

$$G^2 = -2\ln\left(\frac{L_1}{L_2}\right) = \left(-2\ln L_1\right) - \left(-2\ln L_2\right)$$

where equation 1 is nested inside equation 2. The G^2 test statistic is distributed as a chi-square value with degrees of freedom equal to the difference in the two equations' <u>df</u>s $df_{G^2} = df_2 - df_1$. Determine the appropriate critical value to reject a null hypothesis at your chosen α-level

(region of rejection). With a large sample size and nested equations differing by one df: c.v. χ^2 = 3.84 for α = .05; c.v. χ^2 = 6.63 for α = .01; and c.v. χ^2 = 10.83 for α = .001.

Here are relevant portions of outputs on page 17, where visartd was regressed on educ:

```
Iteration 0:   log likelihood = -933.77247
Iteration 1:   log likelihood = -805.81088
Iteration 2:   log likelihood = -801.68997
Iteration 3:   log likelihood = -801.67728
Iteration 4:   log likelihood = -801.67728
```

```
Logistic regression                           Number of obs   =       1501
                                              LR chi2(1)      =     264.19
                                              Prob > chi2     =     0.0000
Log likelihood = -801.67728                   Pseudo R2       =     0.1415
```

The iteration history reports the initial value of the log likelihood for a "constant only" equation that has no independent variables: -933.77. The LL at the final iteration step (= -801.68) is for the equation with all predictors included. The difference between those LL values is ((-933.77) - (-801.68)) = -132.09). Multiply this difference by -2 to obtain the -2LL test-statistic for a pair of nested equations – G^2 = (-2)(-132.09) = 264.18 – which appears on the Stata output as "LR chi2(1)." Often called the "model chi-square," this G^2 has one degree of freedom because the intercept-only model has 1 df (for the constant) and the second model has 2 df (the constant plus the K = 1 predictor, educ).

For this bivariate visit-education logistic regression, what is your decision about the null hypothesis? If you set α = .001, what critical value of chi-square is required to reject the null H_0 of no improvement in fit?

We must reject the null hypothesis with a very low probability of Type I (false rejection) error. Conclusion: adding the single predictor to the equation probably improved the model's fit in the population data.

2. PERCENT OF CASES CORRECTLY CLASSIFIED

OLS regression equations can be used to predict the score of every case, which can then be compared to the observed value to see how accurate is the prediction. Similarly, logistic regression equation can be used to decide

that, if the expected probability is < .50, then predicted score is 0; if the expected probability ≥ .50, then predicted value is 1. The percentages of correctly predicted cases are then calculated and displayed in a two-by-two classification table. If the equation "completely explains" the variation of the dependent variable, all cases would fall on the main diagonal, and the overall percentage correct would be 100%. That is, all cases predicted to equal 0 would be observed 0s, and all predicted 1s would be observed 1s. After running a logistic regression, command Stata to produce the classification table:

estat classification

```
Logistic model for visartd
              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       116            72  |        188
     -     |       355           958  |       1313
-----------+--------------------------+-----------
   Total   |       471          1030  |       1501
                                                  .
Classified + if predicted Pr(D) >= .5
True D defined as visartd != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   24.63%
Specificity                     Pr( -|~D)   93.01%
Positive predictive value       Pr( D| +)   61.70%
Negative predictive value       Pr(~D| -)   72.96%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    6.99%
False - rate for true D         Pr( -| D)   75.37%
False + rate for classified +   Pr(~D| +)   38.30%
False - rate for classified -   Pr( D| -)   27.04%
--------------------------------------------------
Correctly classified                        71.55%
--------------------------------------------------
```

At first glance, the example classification table above seems to indicate a high level of correct predictions (the main diagonal has 1074 of the 1501 case = 71.55% correctly classified). However, we could correctly "predict" 1030/1501 = 68.62% of the cases just by assuming that no one visited an art museum last year. Hence, the bivariate logistic regression equation produces just a small increment over guessing the most frequent response (no visit) for every case.

The two sets of four conditional probabilities help to pinpoint where the equation does a good and poor job of classifying cases. In this example, the low Sensitivity value (116/471 = 24.63%) indicates that the equation correctly classified only one-fourth of the respondents who really visited a museum. Apparently educ alone doesn't identify very accurately who goes to view pictures at an exhibition! To improve classification accuracy, we should consider including additional independent variables in the logistic regression equation predicting art museum visitation. Any suggestions?

3. GENERALIZED "COEFFICIENTS OF DETERMINATION"

In OLS linear regression, the ratio of the between sum of squares to total sum of squares is called the coefficient of determination (R^2). It ranges between 0.00 and 1.00 and can be interpreted as the proportion of the dependent variable's variance "explained" by the linear combination of the independent variables. Further, the sample statistic R^2 is used to test the null hypothesis that the population parameter $\rho^2 > 0$. Because logistic regression uses iterative MLE methods to estimate the equation parameters, instead of variance-minimizing OLS methods, it does not produce a comparable statistic to indicate model fit to the data. Instead, many statisticians have proposed goodness-of-fit measures for logistic regression.* Unfortunately, all lack known sampling distributions and thus can't be tested statistically. Furthermore, many of these generalized "coefficients of determination" produce differing values for the same data.

Stata's basic logistic regression output reports a "Pseudo R2". For the visartd-educ equation above, pseudo-R^2 = 0.1415. Knoke et al. (2002:313) provide another formula:

$$pseudo - R^2 = \frac{G^2}{N + G^2}$$

Applied to the example data, pseudo-R^2 = (264.18)/ (1501+264.18) = 0.1497, which is close to Stata's value.

---

* Liao, J.G. and Dan McGee. 2003. "Adjusted Coefficients of Determination for Logistic Regression." *American Statistician* 57:161-165.

Additional fit statistics are available by running fitstat, a post-estimation command that produces scads of fit statistics for many Stata single-equation regression commands, including: regress, logistic, logit, mlogit, poisson, and probit. Written by J. Scott Long and Jeremy Freese, fitstat can be found on the Web and installed in your Stata program by using this command:

```
findit fitstat
```

After installation, run a logistic regression followed by the command:

```
fitstat
```

```
Measures of Fit for logistic of visartd
Log-Lik Intercept Only: -933.772  Log-Lik Full Model:     -801.677
D(1499):                1603.355  LR(1):                   264.190
                                  Prob > LR:                 0.000
McFadden's R2:             0.141  McFadden's Adj R2:         0.139
Maximum Likelihood R2:     0.161  Cragg & Uhler's R2:        0.227
McKelvey and Zavoina's R2: 0.257  Efron's R2:                0.174
Variance of y*:            4.430  Variance of error:         3.290
Count R2:                  0.716  Adj Count R2:              0.093
AIC:                       1.071  AIC*n:                  1607.355
BIC:                   -9360.162  BIC':                   -256.876
```

Long and Freese (2006:104-113) discuss fitstat methods and formulas.* UCLA Academic Technology Services summarizes eight "commonly encountered pseudo R-squareds" on its FAQ Webpage. Neither source makes recommendations, although Long and Freese call the Count R^2 a "seemingly appealing measure." It is the proportion of correct predictions. Adjusted Count R^2 "is the proportion of correct guesses beyond the number that would be correctly guessed by choosing the largest marginal."**

All measures are descriptive statistics that provide a rough approximation for judging a model's predictive efficacy. No test statistic is available to test the null hypothesis that a generalized $\rho^2 = 0$ in the population.

---

* Long, J. Scott, & Freese, Jeremy (2006). *Regression Models for Categorical Dependent Variables Using Stata (Second Edition)*. College Station, TX: Stata Press.

** UCLA Academic Tech Services. "What are pseudo R-squareds?" <http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm>

---

4. HOSMER-LEMESHOW STATISTIC

The Hosmer-Lemeshow (HL) statistic compares the observed sample scores (y) to the probabilities (π) predicted by the logistic regression equation. First, the HL program sorts the N cases from the lowest to highest predicted probability, then divides them into G groups (quantiles) of approximately equal size (if N/G is not an integer, the G groups may differ slightly in size). Typically, analysts choose G=10, resulting in 10 deciles. Next, within each group, compute the mean predicted probabilities and mean observed scores (i.e., the proportion of cases = 1). Finally, the program calculates HL as a chi-square test statistic with G-2 degrees of freedom:

$$\chi^2_{G-2} = \sum_{g=1}^{G} \frac{(n_g \bar{y}_g - n_g \bar{\pi}_g)^2}{n_g \bar{\pi}_g (1 - \bar{\pi}_g)}$$

If the probability of the HL statistic is $p \leq .05$, we reject the null hypothesis of no difference between observed and predicted values of the dependent variable. If $p > .05$, then we fail to reject the null hypothesis of no difference, implying that the model's parameter estimates fit the data at an acceptable level. Although the model may not explain a large proportion the dependent variable's variation in the population, it's probably more than none.

To perform the HL test, after running a logistic regression equation, use Stata command:

estat gof, group(10) table

```
Logistic model for visartd, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
  (There are only 8 distinct quantiles because of ties)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.1021 |    11 |  10.0 |   149 | 150.0 |   160 |
  |     4 | 0.1879 |    72 |  96.4 |   464 | 439.6 |   536 |
  |     5 | 0.2482 |    35 |  27.8 |    77 |  84.2 |   112 |
  |     6 | 0.3201 |    66 |  61.5 |   126 | 130.5 |   192 |
  |     7 | 0.4018 |    35 |  28.1 |    35 |  41.9 |    70 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     8 | 0.4893 |   136 | 118.9 |   107 | 124.1 |   243 |
  |     9 | 0.6610 |    76 |  79.5 |    49 |  45.5 |   125 |
  |    10 | 0.7987 |    40 |  48.7 |    23 |  14.3 |    63 |
  +--------------------------------------------------------+
       number of observations =      1501
             number of groups =         8
      Hosmer-Lemeshow chi2(6) =        25.60
                  Prob > chi2 =         0.0003
```

In this example, the program could create only G=8 quantiles because the predicted probabilities had many ties (especially in groups #4 and #8), An equation with multiple independent variables would be less likely to encounter this problem. The HL output indicates that the visartd-educ equation does not fit well.

*** Hosmer, D.W., Jr. & S. Lemeshow. 2000. *Applied Logistic Regression*. 2d Ed. NY: Wiley.**

ANALYZING PREDICTED PROBABILITIES

To examine and display the range of predicted probabilities, first run the equation. Then use Stata predict to compute the probability of an art museum visit, and store the results in a new variable (predlogit). (Because Stata will calculate predicted probabilities of all cases in the 2008 GSS, you must include only cases with no missing values.) Next, summarize the predicted values and create a Stata dotplot histogram:

logistic visartd educ
predict predlogit if visartd ~=.

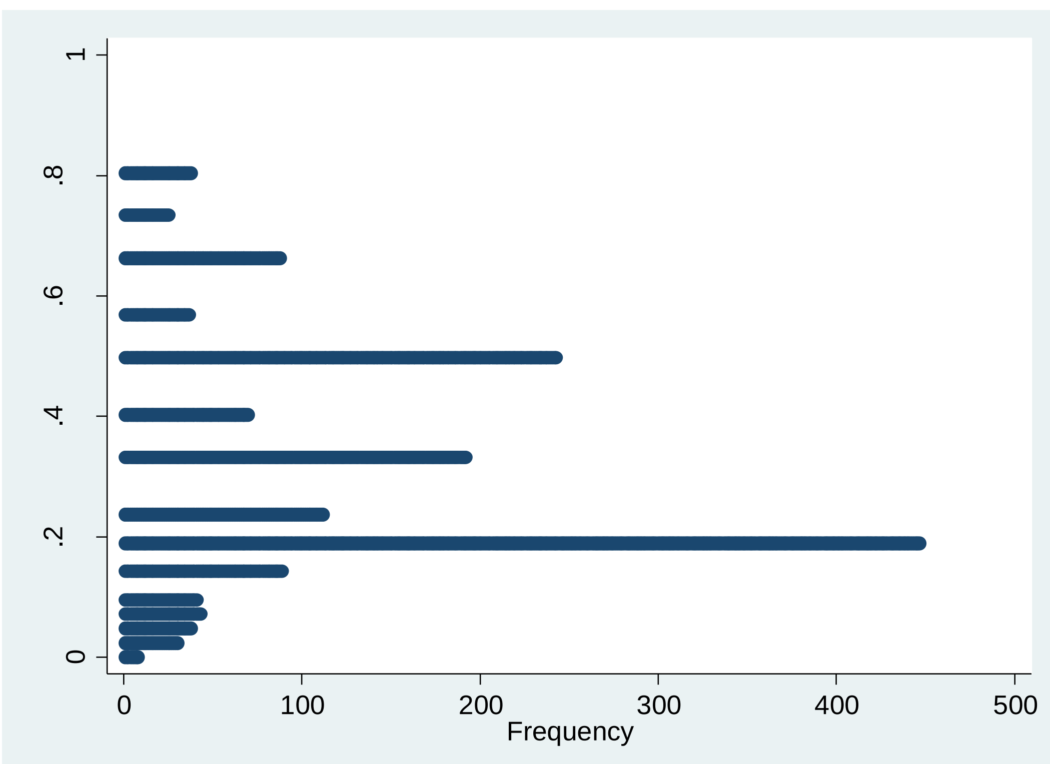
```
(option pr assumed; Pr(visartd))
(522 missing values generated)
```

label var predlogit "Logit: Pr(visartd)"

summarize predlogit

```
Variable |     Obs        Mean    Std. Dev.       Min        Max
---------+-----------------------------------------------------
 logitpr |    1501    .3137908    .1917059    .0032464   .7987462
```

dotplot predlogit, ylabel(0(.2)1)



A MULTIVARIATE EXAMPLE: Capital Punishment

Let's estimate a multivariate logistic regression equation, with several continuous and categoric (dummy) independent variables. The 2008 GSS asked: "Do you favor or oppose the death penalty for persons convicted of murder?" The responses to cappun were "Yes" = 1 and "No" = 2. For independent variables, use educ, polviews, region, race, relig.

1. Recode cappun into a 1-0 dichotomy, where 1 = favors capital punishment:

```
recode cappun (1=1)(2=0), generate(procappun)
codebook procappun
```

```
        range:  [0,1]                          units:  1
unique values:  2                          missing .:  121/2023
   tabulation:  Freq.   Numeric  Label
                  639         0
                 1263         1
                  121         .
```

2. Check all frequencies and missing values for educ, polviews, region, race, relig. Higher scores in polviews indicate more conservative respondents. The last three nonordered discrete measures must be recoded as dummy variables.

3. Create dummy variables and <u>always</u> check the new values and frequencies:

```
recode region (5/7=1)(nonmiss=0), generate(south)
recode race (2=1)(nonmiss=0), generate(black)
recode relig (2=1)(nonmiss=0), generate(catholic)
codebook educ polviews south black catholic
```

4. Run a logistic regression equation of procappun with five predictors:

logit procappun educ polviews south black catholic
logistic procappun educ polviews south black catholic, coef
logistic procappun educ polviews south black catholic

```
Iteration 0:   log likelihood = -1166.7275
Iteration 1:   log likelihood =  -1083.954
Iteration 2:   log likelihood = -1083.0898
Iteration 3:   log likelihood = -1083.0893
Iteration 4:   log likelihood = -1083.0893
```

```
Logistic regression                               Number of obs   =       1823
                                                  LR chi2(5)      =     167.28
                                                  Prob > chi2     =     0.0000
Log likelihood = -1083.0893                       Pseudo R2       =     0.0717

------------------------------------------------------------------------------
   procappun |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0448776   .0178932    -2.51   0.012    -.0799477   -.0098075
    polviews |   .3268739   .0376149     8.69   0.000     .2531501    .4005976
       south |   .1652793   .1123133     1.47   0.141    -.0548507    .3854093
       black |  -1.160563   .1504345    -7.71   0.000    -1.455409   -.8657162
    catholic |  -.3480156   .1242622    -2.80   0.005    -.5915649   -.1044662
       _cons |   .1711663   .3103656     0.55   0.581    -.4371391    .7794717
------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------
   procappun | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .9561145    .017108    -2.51   0.012     .9231646    .9902404
    polviews |   1.386627   .0521578     8.69   0.000     1.288077    1.492717
       south |   1.179723   .1324985     1.47   0.141     .9466264    1.470216
       black |   .3133099   .0471326    -7.71   0.000      .233305    .4207501
    catholic |   .7060879     .08774    -2.80   0.005     .5534605    .9008052
------------------------------------------------------------------------------
```

5. Run the fit statistics:

fitstat

```
Measures of Fit for logistic of procappun
Log-Lik Intercept Only: -1166.727  Log-Lik Full Model:  -1083.089
D(1817):                 2166.179  LR(5):                 167.276
                                   Prob > LR:               0.000
McFadden's R2:              0.072  McFadden's Adj R2:       0.067
Maximum Likelihood R2:      0.088  Cragg & Uhler's R2:      0.121
McKelvey and Zavoina's R2:  0.119  Efron's R2:              0.090
Variance of y*:             3.736  Variance of error:       3.290
Count R2:                   0.677  Adj Count R2:            0.047
AIC:                        1.195  AIC*n:                2178.179
BIC:                   -11476.291  BIC':                 -129.735
```

6. And the HL statistic:

```
Logistic model for procappun, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.4527 |    77 |  68.7 |   106 | 114.3 |   183 |
  |     2 | 0.5371 |    85 |  90.1 |    97 |  91.9 |   182 |
  |     3 | 0.6017 |    94 | 105.5 |    91 |  79.5 |   185 |
  |     4 | 0.6439 |   119 | 113.6 |    62 |  67.4 |   181 |
  |     5 | 0.6962 |   114 | 122.2 |    68 |  59.8 |   182 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.7191 |   151 | 146.1 |    54 |  58.9 |   205 |
  |     7 | 0.7513 |   147 | 146.2 |    50 |  50.8 |   197 |
  |     8 | 0.7878 |   105 | 113.5 |    42 |  33.5 |   147 |
  |     9 | 0.8312 |   194 | 177.5 |    24 |  40.5 |   218 |
  |    10 | 0.9097 |   120 | 122.6 |    23 |  20.4 |   143 |
  +--------------------------------------------------------+

       number of observations =      1823
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =        19.55
                  Prob > chi2 =         0.0122
```

What are your substantive interpretations of these results?

ANALYZING PREDICTED PROBABILITIES

Use Stata predict to create predicted probabilities (predprob) for every case in the multivariate equation:

predict predprob, pr

List these predprob values and the observed values of procappun, using only respondents with no missing values. Here are the paired values for the first 20 respondents:

```
     +----------------------+
     | predprob   procap~n |
     |---------------------|
  1. | .4845967          1 |
  2. | .6413228          0 |
  3. |  .434059          0 |
  4. | .4673772          0 |
  5. | .2943854          1 |
     |---------------------|
  6. | .3212664          1 |
  7. | .2900129          0 |
  8. | .3664897          0 |
  9. | .2121607          1 |
 10. | .4785634          1 |
     |---------------------|
 11. | .2500084          0 |
 12. | .4482606          0 |
 13. | .5851597          0 |
 14. | .4013262          0 |
 17. | .7562764          1 |
     |---------------------|
 18. | .8247694          1 |
 19. | .3616945          0 |
 20. | .6279662          0 |
 21. | .6438503          1 |
 22. | .7765829          1 |
     |---------------------|
```

Although several predictions were good (respondents in yellow), other were erroneous (respondents in green). (Two of the first 20 cases in the survey had missing values and were excluded.)

Stata's adjust command calculates the predicted probabilities for different groups, as identified by categories of variable(s). The example below shows the predicted procappun of men and women.

Caution: By default, the adjust command uses the entire sample, not just the selected cases in the preceding logistic regression. Instruct Stata to adjust the probabilities by sex only for the sampled cases:

adjust if e(sample), pr by(sex)

```
----------------------------
Respondents sex |         pr
----------------+-----------
           male |     .68235
         female |     .66841
----------------------------
```

Predicted probabilities can be calculated for various combinations of attributes. For example, what are the probabilities of supporting capital punishment by polviews and gender among southerners with a college degree?

adjust south=1 educ=16, pr by(polviews sex)

```
-----------------------------------------------------------------
Dependent var: procappun  Equation: procappun   Command: logistic
   Variables left as is: black, catholic
Covariates set to value: south = 1, educ = 16
-----------------------------------------------------------------

-----------------------------
think of self as     |
liberal or           | respondents sex
conservative         |    male   female
---------------------+-----------------
   extremely liberal | .409949  .424054
             liberal | .506429  .505286
    slightly liberal | .574187  .566406
            moderate | .659906  .654442
slghtly conservative | .750288  .739022
        conservative | .813595  .807876
extrmly conservative | .846651  .819780
---------------------------------------
```

How does support for capital punishment differ across the polviews spectrum? Are the sex differences constant or do they interact with polviews?

Use both logit and odds forms of the equation to calculate expected values for a person who: has 12 years of education, is very conservative (=7), lives in the South, and is black and non-Catholic. Show that the two results yield identical values.

General forms:

$$\hat{L}_i = \alpha + \sum \beta_k X_k$$

$$\frac{p_1}{p_0} = \exp^{a+\Sigma\beta X} = (\exp \alpha)^1 (\exp \beta_1)^{X_1} (\exp \beta_2)^{X_2} (\exp \beta_3)^{X_3} \ldots$$

Specific equations (include the exponentiated constant):

$$\hat{L}_i = 0.17 - 0.04 X_E + 0.33 X_P + 0.17 X_S - 1.16 X_B - 0.35 X_C$$

$$\frac{p_1}{p_0} = (1.19)^1 (0.96)^{X_E} (1.39)^{X_P} (1.18)^{X_S} (0.31)^{X_B} (0.71)^{X_C}$$

Substitute & solve:

$$\hat{L}_i = 0.17 - 0.04(12) + 0.33(7) + 0.17(1) - 1.16(1) - 0.35(0)$$

$$\hat{L}_i = 0.17 - 0.48 + 2.31 + 0.17 - 1.16 - 0 = 1.01$$

$$\frac{p_1}{p_0} = (1.19)^1 (0.96)^{12} (1.39)^7 (1.18)^1 (0.31)^1 (0.71)^0$$

$$\frac{p_1}{p_0} = (1.19)(0.61)(10.0)(1.18)(0.31)(1.0) = 2.66$$

Results are same (discrepancy due to cumulative rounding errors):

$$\exp(\hat{L}_i) = \exp(1.01) = 2.75 \cong 2.66 = \frac{p_1}{p_0}$$

STANDARDIZING LOGISTIC COEFFICIENTS

Long and Freese (2006) created a listcoef command, which can be added to Stata, that facilitates interpretation of logistic regression coefficients. To locate their program, open Stata and enter this command:

findit postado

Click on the link shown in the new window and let Stata install the program on your computer. After estimating a logistic regression equation, type:

listcoef, help

```
logit (N=1823): Factor Change in Odds
  Odds of: 1 vs 0
------------------------------------------------------------------
 procappun |      b         z     P>|z|    e^b    e^bStdX    SDofX
-----------+------------------------------------------------------
      educ | -0.04488  -2.508  0.012   0.9561   0.8741    2.9977
  polviews |  0.32687   8.690  0.000   1.3866   1.6025    1.4427
     south |  0.16528   1.472  0.141   1.1797   1.0828    0.4812
     black | -1.16056  -7.715  0.000   0.3133   0.6726    0.3418
  catholic | -0.34802  -2.801  0.005   0.7061   0.8640    0.4201
------------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
     e^b = exp(b) = factor change in odds for unit increase in X
 e^bStdX = exp(b*SD of X) = change in odds for SD increase in X
   SDofX = standard deviation of X
```

The values in the first four columns are identical to those in the usual logistic regression outputs (where the Odds Ratio = e^b; also = exp(b)). The coefficients in the fourth column can be interpreted as the effect on the odds of the dependent variable Y for a 1-unit difference or change in independent variable X. In the fifth column, the effect is expressed as the effect on the odds of the dependent variable Y for a 1-standard deviation difference or change in independent variable X. (The computation formula appears in the table footnotes.) For example, the odds in favor of the death penalty are 0.9561 lower per year of education, and 0.8741 lower per standard deviation of educ.

For a more insightful transformation, change the multiplicative X-standardized effects above into percentage effects, using the command:

listcoef, percent

```
logit (N=1823): Percentage Change in Odds
  Odds of: 1 vs 0
----------------------------------------------------------------
 procappun |      b         z     P>|z|       %      %StdX      SDofX
-----------+----------------------------------------------------
      educ | -0.04488  -2.508  0.012     -4.4     -12.6     2.9977
  polviews |  0.32687   8.690  0.000     38.7      60.3     1.4427
     south |  0.16528   1.472  0.141     18.0       8.3     0.4812
     black | -1.16056  -7.715  0.000    -68.7     -32.7     0.3418
  catholic | -0.34802  -2.801  0.005    -29.4     -13.6     0.4201
----------------------------------------------------------------
```

The effect of a one-standard deviation difference or change in polviews (60.3%) is almost five times as great as the impact of a one-standard deviation difference/change in educ (-12.6%), in the opposite direction.

What do the standardized percentage effects reveal about the relative impacts of the other predictors on visartd?

MULTINOMIAL LOGISTIC REGRESSION

Logistic regression with a binary (dichotomous) dependent variable is a special instance of nonlinear regression involving a multicategory dependent variable, the multinomial logistic regression. A multinomial model is used when the dependent variables has more than two categories that cannot be ranked. For example, workers' employment statuses might be classified as working full-time, working part-time, laid-off, unemployed, and not in the labor force. Artificially forcing all observations into an employed-unemployed dichotomy could be more concealing than revealing. Fortunately, the logistic regression estimation techniques discussed above can be extended to analyze M nonordered discrete categories.

I illustrate multinomial logistic regression using the 2008 GSS to analyze the respondents' 2004 Presidential election choices among three alternatives: vote for Bush, vote for Kerry, or don't vote. (The 25 respondents who voted for Nader were treated as missing data). I created a three-category pres3 variable from two GSS questions about the 2004 presidential election: (1) "Do you remember for sure whether or not you voted in that election?" (vote04); (2) "Did you vote for Kerry or Bush?" (pres04). Stata's syntax in "replace-if" statements uses a double equal sign (==) inside the parentheses:

```
generate pres3 = .
replace pres3 = 0 if (vote04 == 2)
replace pres3 = 1 if (pres04 == 1)
replace pres3 = 2 if (pres04 == 2)
label variable pres3 "three category presidential vote 2004"
label define presvote  0 "Nonvoter" 1 "Kerry" 2 "Bush"
label values pres3 presvote
```

```
codebook pres3
type:  numeric (byte)
 label:  presvote
 range:  [0,2]              units:  1
 unique values:  3    missing .:  305/2023
 tabulation:  Freq.   Numeric  Label
                539         0  Nonvoter
                580         1  Kerry
                599         2  Bush
                305         .
```

The probability that the i̲th observation occurs in the j̲th category of a multicategory dependent variable is designation $\underline{p}_{ij}$. Thus, the nonvoters Kerry voters, and Bush voters are coded 0, 1, and 2, respectively, and their probabilities are symbolized $\underline{p}_{i0}$, $\underline{p}_{i1}$, and $\underline{p}_{i2}$. Probabilities are defined as relative frequencies, so that their sum across the M̲ categories must always equal unity: $\sum_{j=1}^{M} p_{ij} = 1$. Thus, in the 2008 GSS data, $p_{i0} + p_{i1} + p_{i2} = .314 + .338 + .349 = 1.00$.

In a logistic regression equation, the expected probabilities depend in nonlinear ways on the set of K̲ independent variables that predict them. The relationship is given by a multivariate logistic distribution function:

$$p_{ij} = \frac{e^{\alpha + \Sigma \beta_{kj} X_{kji}}}{\sum_{j=1}^{J} e^{\alpha + \Sigma \beta_{kj} X_{kji}}}$$

where
$\underline{p}_{ij}$ = the probability that the i̲th case is in the j̲th category of the dependent variable.

The triple subscripts indicate the i̲th observation on the k̲th predictor variable in the logistic equation for the j̲th category the multicategory dependent variable. To solve these equations for unique parameter estimates, a linear constraint must be placed on the set of βs pertaining to the k̲th predictor. A conventional constraint is that they sum to zero: $\sum_{j=1}^{M} \beta_{kj} = 0$. Just as with dummy-variable predictors in a regression equation, the M̲ categories of a multicategory dependent variable have only M̲ - 1 degrees of freedom. In addition to requiring that the βs for the K̲ predictors sum to 1.00, we can also specify that all coefficients in the M̲th equation equal zero. Then, each estimated coefficient β_{kj} reveals the effect of predictor $\underline{X}_k$ on the odds of respondent i̲ being in the j̲th dependent variable category <u>relative</u> <u>to</u> the omitted category M̲. Which dependent variable category we designate as our reference, or baseline, group is arbitrary.

For the multinomial logit model, the nonlinear transformations cannot assure that the probabilities will add to 1.00. But, as the next section demonstrates, the natural logarithms of the ratios of the probabilities for each category relative to the reference category must sum to 1.00 as required.

The table below displays the parameter estimates for the trichotomous voting example, where the reference category is nonvoters. The Kerry multinomial logit coefficients indicate the effects of the independent variables on voting for Kerry vs. nonvoting, while the Bush coefficients indicate the effects of these predictors on voting for Bush vs. nonvoting.

```
recode partyid (7=1), generate(party7) label(Dem to Repub identifier)
recode region (5/7=1)(nonmiss=0), generate(south)
recode race (2=1)(nonmiss=0), generate(black)
codebook polviews partyid south black educ

mlogit pres3 polviews partyid south black educ, baseoutcome(0)
```

```
Multinomial logistic regression                   Number of obs   =       1656
                                                  LR chi2(8)      =    1039.66
                                                  Prob > chi2     =     0.0000
Log likelihood = -1294.0274                       Pseudo R2       =     0.2866
------------------------------------------------------------------------------
       pres3 |     Coef.   Std. Err.      z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Nonvoter     |  (base outcome)
-------------+----------------------------------------------------------------
Kerry        |
    polviews |  -.0968942  .0550182   -1.76   0.078   -.2047278    .0109394
     partyid |  -.4289445  .0449496   -9.54   0.000   -.5170441   -.3408448
       south |  -.3381738  .1481419   -2.28   0.022   -.6285266   -.0478210
       black |   .2315242  .1825567    1.27   0.205   -.1262803    .5893288
        educ |   .2498392  .0262964    9.50   0.000    .1982993    .3013792
       _cons | -2.0176450  .4301744   -4.69   0.000  -2.8607710  -1.1745190
-------------+----------------------------------------------------------------
Bush         |
    polviews |   .4108660   .0606779    6.77   0.000    .2919395    .5297924
     partyid |   .4327645   .0416732   10.38   0.000    .3510865    .5144425
       south |  -.0076531   .1498175   -0.05   0.959   -.3012899    .2859837
       black |  -1.355903   .3416369   -3.97   0.000  -2.0254990   -.6863072
        educ |    .164274   .0268857    6.11   0.000    .1115798    .2169698
       _cons |  -5.262655   .4687691  -11.23   0.000  -6.1814260  -4.3438850
------------------------------------------------------------------------------
```

Logically enough, several coefficients have opposite signs. For example, the +0.43 **partyid** parameter for Bush vote means that Republican identifiers were more likely to vote for him than to stay home on election day, while the -0.43 parameter for Kerry vote means that Republicans were less likely to vote for him than to stay home (conversely, the Democratic voters were more likely to vote for Kerry than to stay home). Similarly, political conservatives (**polviews**) were more likely to vote for Bush (0.41) than not to vote; however, the nonsignificant -0.10 **polviews** coefficient for Kerry means political conservatives were not more likely to stay home than to vote for him. The positive **educ** coefficients for both types voters indicate that more educated respondents were more likely to vote for either candidate than to stay at home. The negative coefficient for **south** indicates that Southerners were less likely to vote for Kerry than to stay home. But, the **south** coefficient for Bush voting is not significant. The negative coefficient for **black** indicates that blacks were less likely to vote for Bush than to stay home. But the **black** coefficient for Kerry voting is not significant.

To exponentiated the mlogit coefficients, add "rr" (for relative risk ratio) to the end of the command:

mlogit pres3 polviews partyid south black educ, baseoutcome(0) rr

pres3	RRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Nonvoter	(base outcome)					
Kerry						
polviews	.907652	.0499373	-1.76	0.078	.8148691	1.0109990
partyid	.6511961	.029271	-9.54	0.000	.5962805	.7111693
south	.7130713	.1056357	-2.28	0.022	.5333771	.9533044
black	1.26052	.2301164	1.27	0.205	.8813677	1.802778
educ	1.283819	.0337598	9.50	0.000	1.219327	1.351722
Bush						
polviews	1.508123	.0915097	6.77	0.000	1.339022	1.69858
partyid	1.541513	.0642398	10.38	0.000	1.42061	1.672706
south	.9923761	.1486753	-0.05	0.959	.7398632	1.331071
black	.2577144	.0880448	-3.97	0.000	.131928	.5034317
educ	1.178538	.0316858	6.11	0.000	1.118043	1.242307

Interpretation of a relative risk ratio is similar to the odds ratio, except the comparison is to the reference category.

Below is the model fit output. **How well do the four predictors explain voting turnout and choice of candidate?**

fitstat

```
Measures of Fit for mlogit of pres3
Log-Lik Intercept Only: -1813.859  Log-Lik Full Model:   -1294.027
D(1637):                 2588.055  LR(10):                1039.663
                                   Prob > LR:                0.000
McFadden's R2:              0.287  McFadden's Adj R2:        0.277
Maximum Likelihood R2:      0.466  Cragg & Uhler's R2:       0.525
Count R2:                   0.463  Adj Count R2:             0.183
AIC:                        1.586  AIC*n:                 2624.055
BIC:                    -9544.663  BIC':                  -965.548
```

Stata doesn't have a command to produce a classification table for mlogit. However, you compute the number of cases correctly predicted in each category by following these steps. First, find the predicted probabilities of each choice for every case, and list the results:

predict prednovote predkerry predbush if e(sample), pr
list prednovote predkerry predbush

Here are cases #61 to #70, which have a mixture of missing values, and differential predictions for all three choices:

```
     +--------------------------------+
     | predno~e   predke~y   predbush |
     |--------------------------------|
 61. | .0570762   .9249492   .0179746 |
 62. | .0617898   .9088636   .0293466 |
 63. | .3525854   .6355622   .0118524 |
 64. | .5696362   .3849729   .0453909 |
 65. |        .          .          . |
     |--------------------------------|
 66. | .2353433   .7017469   .0629097 |
 67. |        .          .          . |
 68. | .1992154    .024945   .7758396 |
 69. | .2080186   .1197146   .6722668 |
 70. | .1039949   .8423758   .0536294 |
     |--------------------------------|
```

Next, create three new binary variables, coded 0 or 1 for every case, according to whether the three predicted probabilities are < .50 or ≥ .50:

```
generate nvbin = 0
replace nvbin = 1 if (prednovote >= .50)
replace nvbin = . if (prednovote == .)
generate kerrybin = 0
replace kerrybin = 1 if (predkerry >= .50)
replace kerrybin = . if (predkerry == .)
generate bushbin = 0
replace bushbin = 1 if (predbush >= .50)
replace bushbin = . if (predbush == .)
```

List the results to see whether the replacements were made correctly, and also show the respondent's reported vote decision (pres3). Finally, crosstabulate the binary predictions with the actual vote.

```
list pres3 prednovote nvbin predkerry kerrybin predbush bushbin
table pres3 nvbin
table pres3 kerrybin
table pres3 bushbin
```

```
     +-------------------------------------------------------------------------+
     |    pres3   predno~e   nvbin   predke~y   kerrybin    predbush   bushbin |
     |-------------------------------------------------------------------------|
 61. |    Kerry   .0570762       0   .9249492          1    .0179746         0 |
 62. |    Kerry   .0617898       0   .9088636          1    .0293466         0 |
 63. | Nonvoter   .3525854       0   .6355622          1    .0118524         0 |
 64. | Nonvoter   .5696362       1   .3849729          0    .0453909         0 |
 65. |        .          .       .          .          .           .         . |
     |-------------------------------------------------------------------------|
 66. | Nonvoter   .2353433       0   .7017469          1    .0629097         0 |
 67. |        .          .       .          .          .           .         . |
 68. |     Bush   .1992154       0    .024945          0    .7758396         1 |
 69. |     Bush   .2080186       0   .1197146          0    .6722668         1 |
 70. |    Kerry   .1039949       0   .8423758          1    .0536294         0 |
     |-------------------------------------------------------------------------|
```

```
----------------------      ----------------------      ----------------------
2004 vote |    nvbin        2004 vote | kerrybin        2004 vote | bushbin
          |     0      1              |    0      1               |    0      1
----------+-----------      ----------+-----------      ----------+-----------
 Nonvoter |  372    125      Nonvoter |  379    118      Nonvoter |  405     92
    Kerry |  534     34         Kerry |  192    376         Kerry |  527     41
     Bush |  573     17          Bush |  546     44          Bush |  150    440
----------------------      ----------------------      ----------------------
```

Which predicted vote choice had the highest percentage correct? Lowest?

Box 9.2 Multicategory Probabilities Relative to a Reference Category

---

For M ≥ 2 discrete nonordered categories of a dependent variable and K ≥ 1 predictor variables, let any arbitrarily chosen baseline or reference category M have the logit probability

$$p(Y_i = M) = \frac{1}{1 + \sum_{j=1}^{M-1} e^{Z_{im}}}$$

where Z_{im} represents $\alpha + \sum_{k=1}^{K} \beta_{jk} X_{jki}$ for the m = M - 1 other categories of the dependent variable (the subscript i stands for the ith individual observation).

Given an mth dependent variable category, its logit relative to the Mth baseline category is:

$$\log_e \left(\frac{p(Y_i = m)}{p(Y_i = M)} \right) = Z_{im}$$

Exponentiate this expression and rearrange as follows:

$$\frac{p(Y_i = m)}{p(Y_i = M)} = e^{Z_{im}}$$

$$\text{Therefore, } p(Y_i = m) = (p(Y_i = M))(e^{Z_{im}})$$

Now, substituting the first equation in this box into the immediately preceding equation and carrying out the multiplication results in the following equation for the probability that the ith observation falls into the mth category of the dependent variable:

$$p(Y_i = m) = \left(\frac{1}{1 + \sum_{j=1}^{M-1} e^{Z_{im}}} \right) \left(e^{Z_{im}} \right) = \frac{e^{Z_{im}}}{1 + \sum_{j=1}^{M-1} e^{Z_{im}}}$$

Next, normalize the denominator of the preceding equation by setting α and all the βs in the M̲th baseline equation equal to 0. Because in general $e^0 = 1$, so $e^{Z_{iM}} = 1$ when all the $\underline{\alpha}$ and $\underline{\beta}$ parameters in the M̲th equation are set to zero. Consequently, we can replace the 1 in the denominator with this exponential term:

$$p(Y_i = m) = \frac{e^{Z_{im}}}{1 + \sum_{j=1}^{M-1} e^{Z_{im}}} = \frac{e^{Z_{im}}}{e^{Z_{iM}} + \sum_{j=1}^{M-1} e^{Z_{im}}} = \frac{e^{Z_{im}}}{\sum_{j=1}^{M} e^{Z_{im}}}$$

because the denominator now sums across all M̲ equations. Thus, the probability that observation $\underline{Y}_i$ is in the m̲th category is expressed relative to the sum over all M̲ categories.

Finally, also apply the probability formula to the M̲th category where all parameters were set to zero:

$$p(Y_i = M) = \frac{e^0}{e^0 + \sum_{j=1}^{M-1} e^{Z_{im}}} = \frac{1}{1 + \sum_{j=1}^{M-1} e^{Z_{im}}} = \frac{1}{\sum_{j=1}^{M} e^{Z_{im}}}$$

When the probabilities for all M̲ categories are added, their sum equals 1.00. That is,

$$\sum_{j=1}^{M} p_i = \sum_{j=1}^{M-1}\left(\frac{e^{Z_{im}}}{\sum_{j=1}^{M} e^{Z_{im}}}\right) + \frac{1}{\sum_{j=1}^{M} e^{Z_{im}}}$$

$$= \frac{\sum_{j=1}^{M-1} e^{Z_{im}} + 1}{\sum_{j=1}^{M} e^{Z_{im}}} = \frac{\sum_{j=1}^{M} e^{Z_{im}}}{\sum_{j=1}^{M} e^{Z_{im}}} = 1.00$$

ORDERED LOGIT

Some categorical variables are ordered but cannot be considered continuous measures, thus rendering OLS linear regression problematic. Examples include subjective social class (lower, working, middle, upper), behavioral frequencies (none, little, some, many), and most attitude items (strong disagree, disagree, neither, agree, strongly agree). The ordered logit model, also called ordinal regression (McKelvey and Zavoina 1975)*, does not require an assumption of equal distances between the set of ordered categories.

The dependent variable is conceptualized as a continuous latent variable (y*) ranging from –∞ to +∞. For a single independent variable, the structural equation is:

$$y_i^* = \alpha + \beta X_i + \varepsilon_i$$

The measurement model divides y* into *J* ordinal categories:

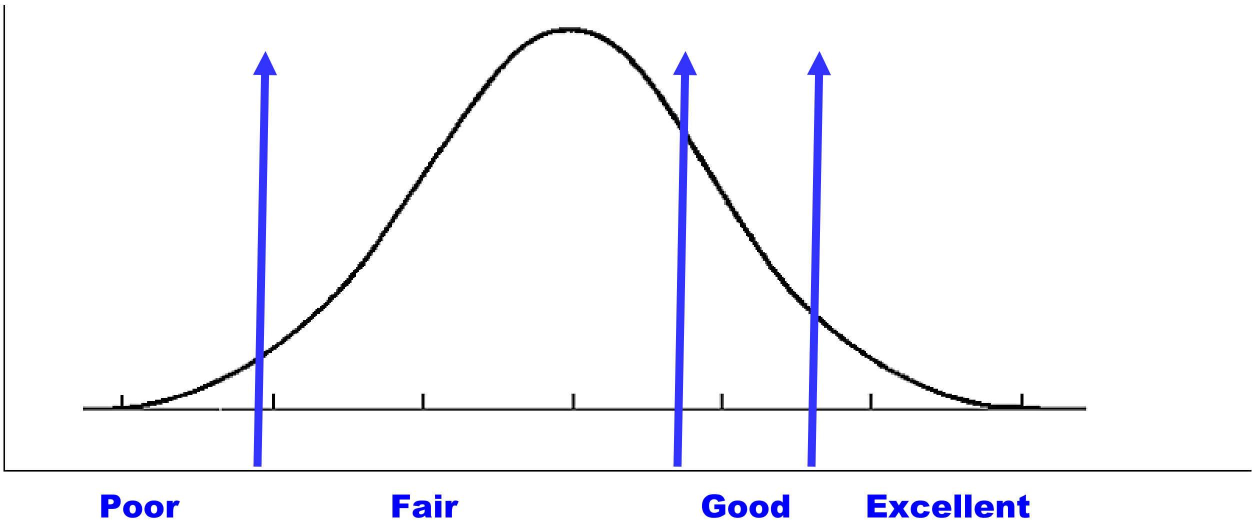
$$y_i = m \quad if \ \tau_{m-1} \leq y_i^* < \tau_m \qquad \text{for m = 1 to J}$$

where the tau cutpoints (thresholds) are estimated by the program. The measurement model assumes that

$$\tau_0 = -\infty \quad and \quad \tau_J = +\infty$$

*** McKelvey, R.D. and W. Zavoina. 1975. "A Statistical Model for the Analysis of Ordinal Dependent Variables." *Journal of Mathematical Sociology* 4:103-120.**

Suppose the distribution below represents an unobserved attitude towards some entity (e.g., "How well is President Obama doing his job?"). If a respondent's attitude falls below a particular unobserved threshold, then she will likely choose a corresponding category on a provided response scale comprised of four ordered categories: "excellent, good, fair, poor." (Although OLS regression assumes the dependent variable is normally distributed, it's not required for the ordered logit model.)



In contrast to logistic regression, which estimates the probability that Y=1, the ordered logit model examines the probability of falling into a particular range.

To illustrate Stata's ologit program, I analyze a 2008 GSS attitude item, natrace: "Are we spending too much, too little, or about the right amount on improving the conditions of Blacks." Here's the frequency distribution:

```
          range:  [1,3]                         units:  1
  unique values:  3                         missing .:  1142/2023
     tabulation:  Freq.   Numeric  Label
                    350         1  too little
                    420         2  about right
                    111         3  too much
                   1142         .
```

Use Stata's ologit program to regress natrace on six independent variables (to obtain odds ratios instead of coefficients, add ", or" to the end of this command line):

ologit natrace black educ age female partyid polviews

```
Ordered logistic regression                       Number of obs   =        841
                                                  LR chi2(6)      =     184.19
                                                  Prob > chi2     =     0.0000
Log likelihood = -728.43465                       Pseudo R2       =     0.1122
------------------------------------------------------------------------------
     natrace |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black | -2.175909   .2600274   -8.37 0.000   -2.685553 -1.666264
        educ |  -.032554   .0236124   -1.38 0.168    -.078834   .013725
         age |  -.002502   .0041122   -0.61 0.543    -.010561   .005557
      female |  -.225914   .1405413   -1.61 0.108    -.50137    .049541
     partyid |   .095954   .0398634    2.41 0.016     .017823   .174085
    polviews |   .308361   .0553721    5.57 0.000     .199833   .416888
-------------+----------------------------------------------------------------
       /cut1 |   .043502   .4565776                  -.851373   .938377
       /cut2 |  2.786168   .4714708                  1.862102  3.710234
------------------------------------------------------------------------------
```

Instead of a constant, ologit reports two cutpoints (thresholds), which can be used to compute the probability of a case falling into a particular interval on the dependent variable. Fully standardize all coefficients:

listcoef, std help

```
----------------------------------------------------------------------
natrace |        b         z     P>|z|    bStdX     bStdY    bStdXY      SDofX
--------+-------------------------------------------------------------
  black | -2.17591    -8.368   0.000  -0.7745   -1.0401   -0.3702     0.3559
   educ | -0.03255    -1.379   0.168  -0.0982   -0.0156   -0.0469     3.0171
    age | -0.00250    -0.609   0.543  -0.0431   -0.0012   -0.0206    17.2377
 female | -0.22591    -1.607   0.108  -0.1129   -0.1080   -0.0540     0.4999
partyid |  0.09595     2.407   0.016   0.1975    0.0459    0.0944     2.0585
polviews|  0.30836     5.569   0.000   0.4547    0.1474    0.2173     1.4745
----------------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
   bStdX = x-standardized coefficient
   bStdY = y-standardized coefficient
  bStdXY = fully standardized coefficient
   SDofX = standard deviation of X
```

Which independent variable has the largest effect on natrace? Which have the least impacts?

III. OTHER TOPICS

This section examines other topics applicable to multivariate models. It includes: (1) nonlinear independent variables; (2) dummy variables; (3) interaction terms; (4) comparing separate regression equations.

NONLINEAR INDEPENDENT VARIABLES

In logistic regression, a parameter estimate (b) depicts a linear relation between the logit of a dichotomous dependent variable and an independent variable. That is, for each unit increase in X, the logit increases (or decreases) by b-units. Although these effects can be transformed into nonlinear relationships by exponentiation, the logged relationship remains linear. By recoding the independent variable, we can test whether significant nonlinear effects on the logit occur. Three methods, which can also be applied to OLS regression, use power terms, logarithmic transformations, and spline coding.

1. POWER TERMS

A common method of assessing nonlinear effects of a continuous independent is to include one or more power transformations of the predictors. For example, consider a logistic regression of Republican party identification (partyid recoded as partyid2 where "strong Republican," "not strong Republican," and "Independent, near Republican" = 1, "Independent and Democrat" = 0) on years of education (educ):

```
recode partyid(0/3=0)(4/6=1)(7=.), generate (partyid2)
logit partyid2 educ
```

$$\hat{L}_i = -1.499 + 0.061\,X_{educ}$$
$$(0.222) \quad (0.016)$$

where the standard errors are in parentheses. The positive sign of the coefficient of education indicates that each year of schooling increases the log-odds of identifying with the Republican party. At what level of

significance can you reject the null hypothesis that education is unrelated to party in the population?

Next, create a squared term for education and include it to the equation with the linear predictor:

```
generate educ2=educ*educ
logit partyid2 educ educ2
```

The results show that both predictors are significant, but have opposite signs:

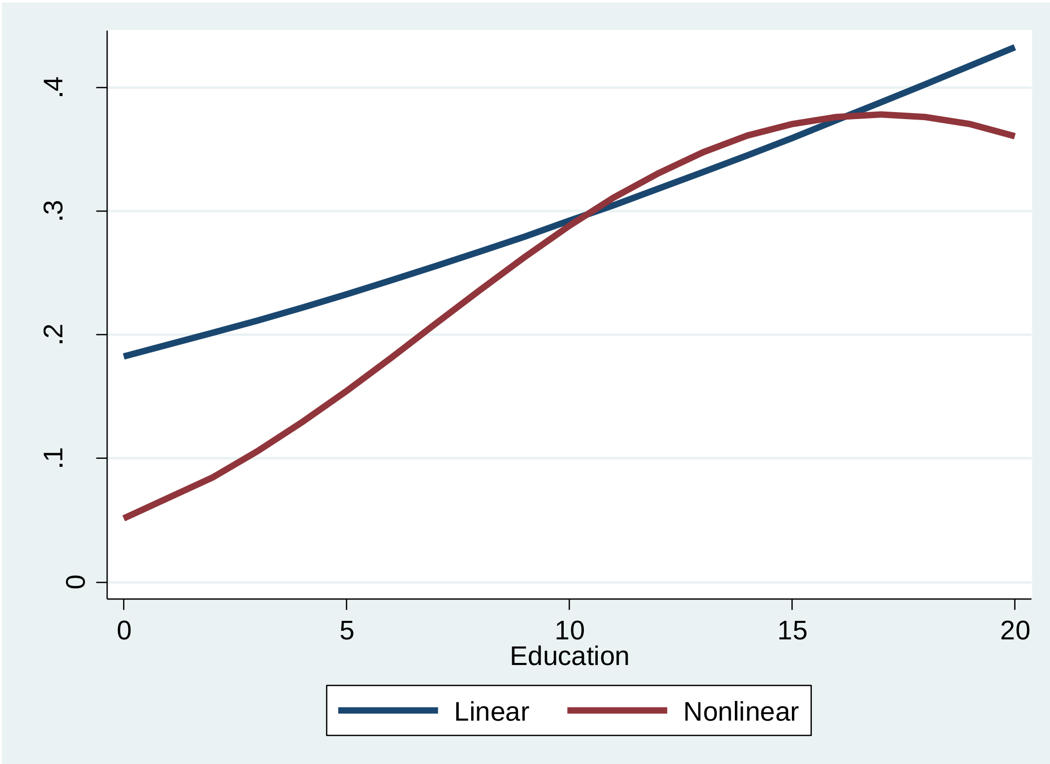
$$\hat{L}_i = -2.914 + 0.285\,X_{educ} - 0.008\,X^2_{educ^2}$$
$$(0.674) \quad (0.100) \qquad (0.004)$$

The difference in –2LLs for the two equations is also significant: G^2 = (2503.9) - (2497.0) = 6.9 for <u>df</u> = 1, p < .01.

Substantively, the second equation shows that although the logit of Republican identification increases linearly with education, such support increasingly falls off as education approaches its highest levels.

By graphing the predicted probabilities for both equations across the 21 years of educ, we can easily visualize how the combined linear and nonlinear effects relate to Republican party identification.

```
logit partyid2 educ
predict repub_linear
logit partyid2 educ educ2
predict repub_nonlinear
sort educ
twoway (line repub_linear educ) (line repub_nonlinear educ),
        ytitle(Probability Republican) xtitle(Education) legend(order(1
        "Linear" 2 "Nonlinear"))
```



Increasing Republicanism occurs only among people who completed schooling before obtaining a college degree. Among people with college degrees (16 years) and more, Republican identification does not increase but slightly decreases. Because the linear equation's parameter estimates are heavily influenced by the large numbers of cases occurring in the middle-range of educ, it failed to detect the downward-curving right tail.

2. LOGARITHMIC INDEPENDENT VARIABLES

The independent variables in a logistic regression equation can also be transformed using the logarithmic function. Such specifications involve nonlinearities in the dependent and independent variables' relationships. To illustrate, I estimated a natural log relationship between women's ages and multiple children ever-born. I dichotomized between 2 or more children ever-born (= 1) versus one or none (= 0). My hypothesis is older woman are likely to have had multiple childbirths. However, in using the natural log of marital age, I expect the probability of plural motherhood to increase more slowly with age. The logistic equation specification is

$$\hat{L}_{\mathrm{i}} = \alpha + \beta \ln X_{age}$$

where ln X_{age} is the natural logarithm (base e) of age in years. The β coefficient has a positive sign, consistent with my hypothesis that older women are more likely than younger women to have two or more children.

```
recode childs(0/1=0)(nonmiss=1), generate(kidsbin)
generate ageln=ln(age)
```

The logistic regression equation for N=1,085 women:

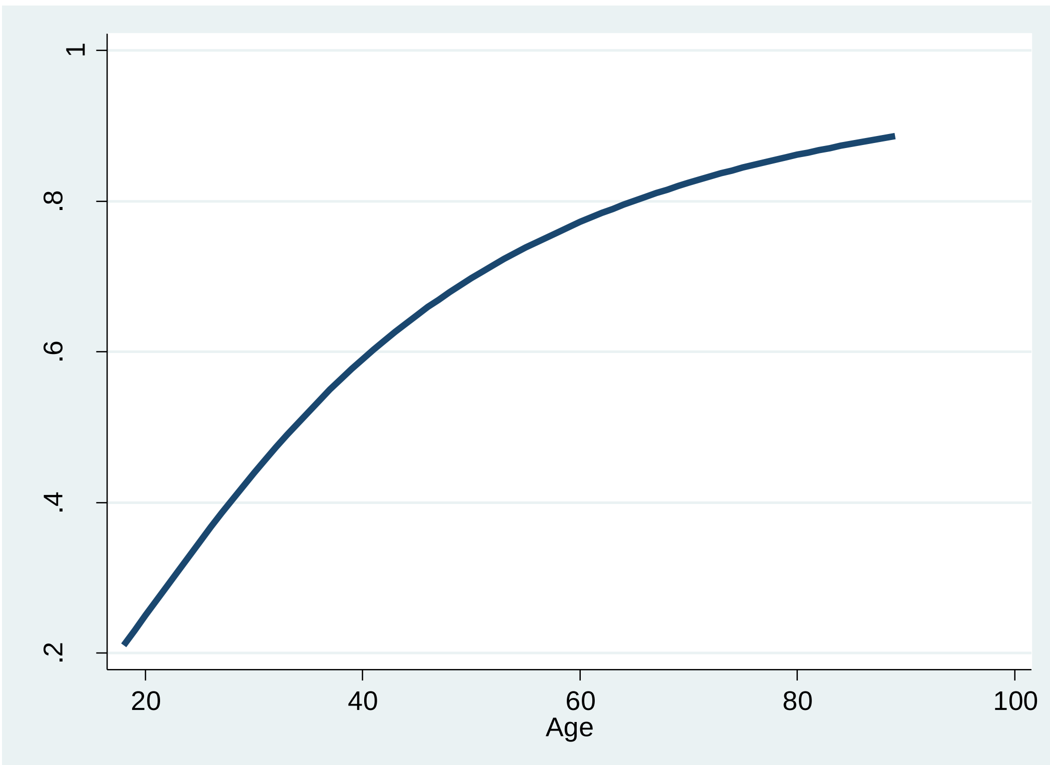
```
logit kidsbin ageln if sex==2
```

$$\hat{L}_{\mathrm{i}} = -7.42 + 2.11 \ln X_{age}$$
$$(0.680) \quad (0.18)$$

The expected logit is not constant across the (log-transformed) age variable. For example, the expected logit of multiple births for a woman age 18 years is -7.42 + (2.11) (2.89) = -1.32, while a woman of 20 years has a expected logit of -7.42 +(2.11)(3.00) = -1.10, a difference of 0.22 across that three-year interval. Women ages 37 and 40 years have a smaller difference (0.20 and 0.36 = 0.16), while the difference between women ages 57 and 60 years is still smaller (1.11 and 1.22 = .11). Clearly, a woman's expected odds of multiple childbirth increase with age, but at a decreasing rate.

A graph of the predicted probabilities also shows a nonlinear relationship between age and the probability of multiple childbirths:

```
predict probkids
sort age
twoway (line probkids age), ytitle(Probability Multiple Births)
    xtitle(Age)
```



3. SPLINES

The effect of a continuous independent variable on a dependent variable may not be uniformly linear or curvilinear across its range. Among other variables, age, education, and income may exhibit threshold effects, "kinks," and other unusual discontinuities. One way to determine whether such departures occur is by spline-coding the predictor. In effect, one or more new independent variables are constructed having this general form:

newvar = oldvar - x if oldvar > k and 0 otherwise

where k is the threshold value above which a shift in slope is expected.

Here's an example of spline coding for schooling. hischool is coded for completing high school plus additional years, while college counts the number of years from a BA through grad school. A coefficient for one of these splines, controlling for the linear effect of educ, could be interpreted as the impact of earning diploma on the dependent measure.

educ	hischool	college
0	0	0
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0
9	0	0
10	0	0
11	0	0
12	1	0
13	2	0
14	3	0
15	4	0
16	5	1
17	6	2
18	7	3
19	8	4
20	9	5

To illustrate, I analyze the conditional effects of college education on the logit of partyid2, where 1 = Republican, 0 = Other.

Create college, a spline-coded variable for the number of years of education starting from 16 up to 20 (i.e., college graduation and post-BA schooling).

```
recode educ(0/15=0), generate(coll)
recode coll(0=0)(16=1)(17=2)(18=3)(19=4)(20=5), generate(college)
```

College	Freq.
0	1,431
1	322
2	51
3	124
4	38
5	52

As shown at the beginning of this section, a logistic regression of partyid2 on educ produces a highly significant positive effect on the log-odds of Republican identification. Here's that linear effect on the logit again:

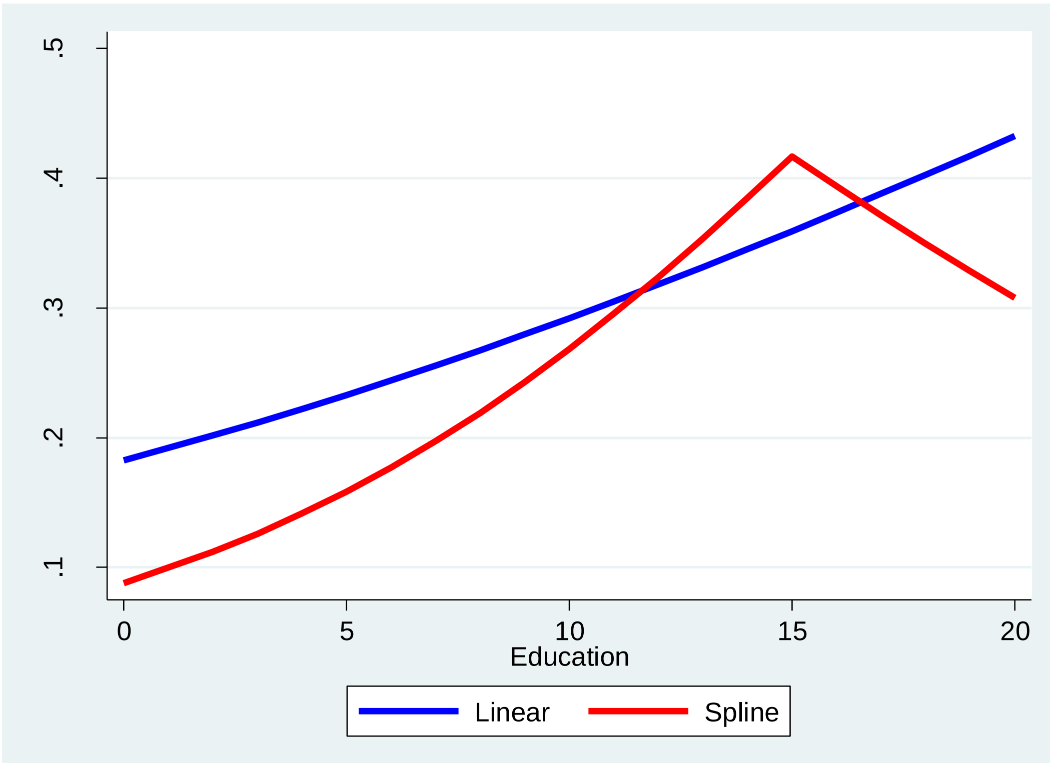
$$\hat{L}_i = -1.499 \; + \; 0.061\, X_{educ}$$
$$(0.222) \qquad (0.016)$$

Next, enter both educ and the spline-coded college variable, which results in two highly significant effects, with opposite signs:

$$\hat{L}_i = -2.337 \; + \; 0.133\, X_{educ} \; - \; 0.228\, X_{college}$$
$$(0.336) \quad (0.026) \qquad\qquad (0.064)$$

The positive linear effect of education on Republican identification is much steeper in the second equation, but is more than offset in the higher end by a strong negative effect from years 16 to 20. The joint effect of these education variables on the probability of Republican identification can be seen in the figure below, which graphs both equations:

```
logit partyid2 educ
predict repub_linear
logit partyid2 educ college
predict repub_spline
sort educ
twoway (line repub_linear educ) (line repub_spline educ),
        ytitle(Probability Republican) xtitle(Education) legend(order(1
        "Linear" 2 "Spline"))
```



The spline plot shows that the expected probability of Republican identification increases with each year of educ, reaching a peak at 15 years. Then, as the college effect takes over, Republican identification decreases markedly through the next five years. Although a somewhat similar pattern occurred in the power equation above, the spline analysis indicates that the reversal at the highest education levels is even more dramatic.

DUMMY VARIABLES

Principles of dummy variable predictors for logistic regression are the same as for OLS regression with continuous dependent variables (see pages 271-275 in *SSDA*). This example analyzes a dichotomous dependent variable happyd ("very happy" = 1 and "pretty happy, not too happy" = 0), recoded from the 2008 GSS variable happy. The marital status variable has five unordered categories:

```
marital          marital status
        range:  [1,5]                          units:  1
unique values:  5                          missing .:  5/2023
   tabulation:  Freq.   Numeric  Label
                  972         1  married
                  164         2  widowed
                  281         3  divorced
                   70         4  separated
                  531         5  never married
                    5         .
```

The five dummy dichotomies will have the pattern below in relation to marital. Every respondent is coded "1" only one dummy for his/her marital status, "0" on all other dummies:

marital:	married	widowd	divorced	separated	unmarried
1. married	1	0	0	0	0
2. widowed	0	1	0	0	0
3. divorced	0	0	1	0	0
4. separated	0	0	0	1	0
5. never married	0	0	0	0	1

Use these Stata generate commands to create the five dummy variables:

```
generate married = marital == 1 if marital < .
generate widowd = marital == 2 if marital < .
generate divorced = marital == 3 if marital < .
generate separated = marital == 4 if marital < .
generate unmarried = marital == 5 if marital < .
```

Always run a codebook on new variables to see whether the commands worked properly. Here's one:

```
married
        range:  [0,1]      units:  1
unique values:  2      missing .:  5/2023
            tabulation:  Freq.  Value
                          1046  0
                           972  1
                             5  .
```

These five dummies are <u>linearly dependent</u>: If you know a person's codes on any four of the dummies, you also know his/her code on the fifth dummy. Consequently, all five dummies cannot be used together as predictors in a multivariate equation. Instead, one dummy must be omitted: it serves as a "reference" or "baseline" category against which to judge the effects of the remaining *K* - 1 dummy predictors.

Let's choose divorced as the reference category (so the four coefficients have positive signs) and estimate a logistic regression with HAPPYD as the dependent dichotomy:

logistic happyd married widowd separated unmarried, coef

```
Logistic regression                    Number of obs   =       2010
                                       LR chi2(4)      =     118.03
                                           Prob > chi2 =     0.0000
Log likelihood = -1163.6958              Pseudo R2     =     0.0483
---------------------------------------------------------------------
      happyd |      Coef.   Std. Err.       z     P>|z|
-------------+-------------------------------------------------------
     married |   1.025357     .16305      6.29    0.000
      widowd |  -.0625204   .2493096     -0.25    0.802
   separated |  -.2933499   .3608044     -0.81    0.416
   unmarried |  -.0553692   .1859149     -0.30    0.766
       _cons |  -1.386294   .1494036     -9.28    0.000
---------------------------------------------------------------------
```

The B-coefficients are interpreted relative to divorced persons (i.e., the omitted divorced dummy has an implicit B = 0). Because all four dummy variable coefficients have positive signs, persons in these four marital categories have higher predicted logit values for "very happy" than do the divorced respondents. However, only married dummy has significantly higher log-odds of very happiness ($p < .001$) in the population. The separated and widowd respondents do not differ significantly from divorced persons in their log-odds of being very happy.

Suppose we had chosen a different reference category. Re-run the logistic regression omitting married and adding divorced:

```
Logistic regression                    Number of obs    =    2010
                                       LR chi2(4)       =  118.03
                                           Prob > chi2 =  0.0000
Log likelihood = -1163.6958              Pseudo R2 =  0.0483
-----------------------------------------------------------------
      happyd |       Coef.    Std. Err.       z     P>|z|
-------------+---------------------------------------------------
    divorced |   -1.025357      .16305     -6.29    0.000
      widowd |   -1.087878    .2099945     -5.18    0.000
   separated |   -1.318707    .3348466     -3.94    0.000
   unmarried |   -1.080726    .1284787     -8.41    0.000
       _cons |   -.3609372    .0652984     -5.53    0.000
-----------------------------------------------------------------
```

The overall fit statistics remain unchanged. But the B's for the dummy set are now all negative and all significant! Does this mean we must change our interpretation of how marital status affects happiness, according to whichever baseline/reference we choose?

Compare the two outputs: Which marital group is happiest? Which is least happy? (HINT: Compare the B’s for married and divorced in the two equations.) Are the other 3 categories similar to one another and closer to the happiest or least happy category? Do both equations yield the same substantive interpretations? Can you show how to translate the coefficients in the first equation into those in the second equation & vice versa?

INTERACTION EFFECTS

Interaction effects can yield important insights into complex conditional relationships among three or more variables. We'll examine three approaches in the logistic regression situation: (1) the ANCOVA method for interaction of a continuous independent variable and a dummy variable on a dependent variable; (2) the centered product term method for the interaction of two continuous independent variables; and (3) the comparison of parameters for separate subsample equations.

1. ANCOVA

The analysis of covariance (ANCOVA) model in OLS and logistic regression refers to an equation that includes both continuous and dummy independent variables. Their parameters estimates are additive effects; that is, the effect of each predictor is same (constant) regardless of the values of the other independent variables. For example, estimate a logistic regression of fepresch ("A preschool child is likely to suffer if his or her mother works..") on a continuous variable age and a dummy variable female. The dependent variable is dichotomized into "Agree" = 1 and "Disagree" = 0.

```
recode fepresch (1/2=1)(3/4=0), generate(fepreschd)
recode sex (2=1)(1=0), generate(female)
logistic fepreschd age female, coef
```

```
Logistic regression                    Number of obs    =     1300
                                       LR chi2(2)       =    50.66
                                       Prob > chi2      =   0.0000
Log likelihood = -825.82664                Pseudo R2    =   0.0298
------------------------------------------------------------------
   fepreschd |      Coef.    Std. Err.        z      P>|z|
-------------+----------------------------------------------------
         age |    .0172548   .0035105      4.92     0.000
      female |   -.616306    .1179667     -5.22     0.000
       _cons |  -1.073554    .1846264     -5.81     0.000
------------------------------------------------------------------
```

Each year of age strongly increases the predicted logit of a traditional sex-role response by 0.017 and the female dummy has a negative effect (-0.616), meaning that women are slightly less traditional than men. The additive nature of this specification means that the effect of age is identical for both

genders: men and women express more traditional views by identical logit amounts per year of age.

We can test the hypothesis that the age-effect differs for men and women by forming an interaction term between gender and age. First, multiply these two values to create the interaction term femage. Because the men are coded "0" on female, all male respondents are given the value "0" on the interaction term:

```
generate femage = age*female
```

Include the two "main-effect" predictors plus their interaction term in another logistic regression:

```
logistic fepresch age female femage, coef
```

```
Logistic regression                        Number of obs   =       1300
                                           LR chi2(3)      =      54.87
                                           Prob > chi2     =     0.0000
Log likelihood = -823.71735                Pseudo R2       =     0.0322
-----------------------------------------------------------------------
   fepreschd |      Coef.   Std. Err.       z     P>|z|
-------------+---------------------------------------------------------
         age |   .0248372   .0051567      4.82    0.000
      female |    .078602   .3583515      0.22    0.826
      femage |  -.0144894   .0070774     -2.05    0.041
       _cons |  -1.431303   .2574344     -5.56    0.000
-----------------------------------------------------------------------
```

The femage interaction has a small negative effect, while the main effect of female has vanished. We can see the differing effects for age and gender by combining the four parameters to create two predictor equations. For men, the effect of age on attitude involves only the constant and age coefficients:

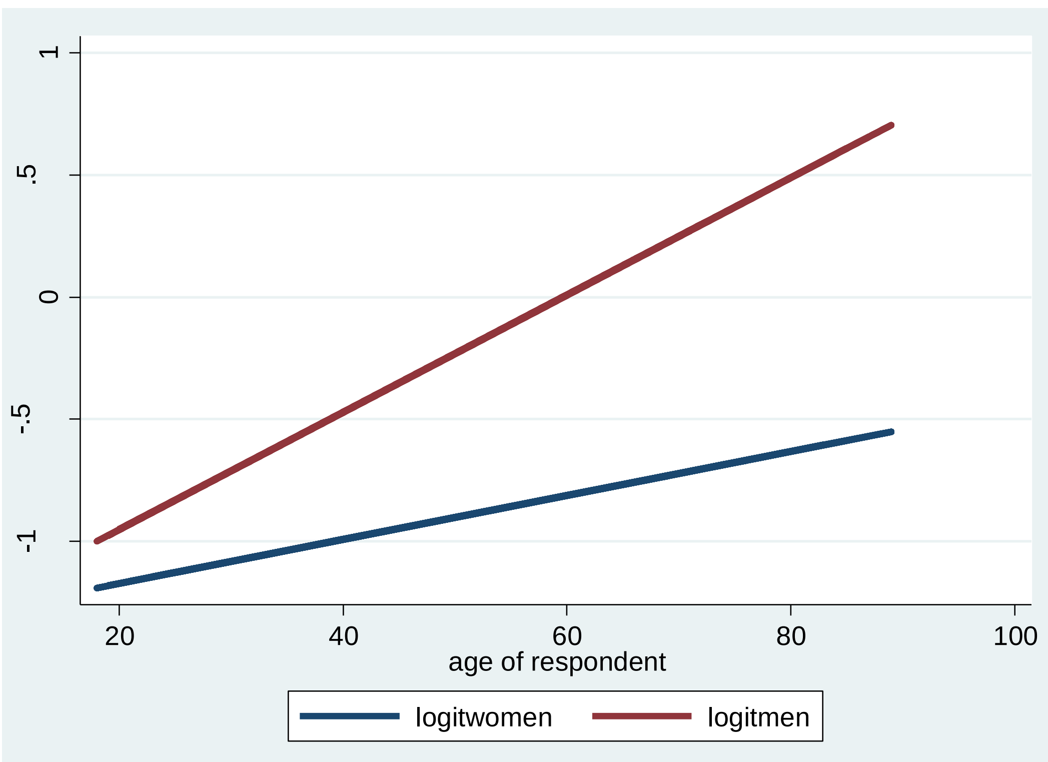
$$\hat{L}_i = -1.431 + 0.024\,X_{age} + 0.079\,D_{female} - 0.015\,X_{femage}$$
$$= -1.431 + 0.024\,X_{age} + 0.079\,(0) - 0.015\,(0)$$
$$= -1.431 + 0.024\,X_{age}$$

The equation for women combines all four parameters:

$$\hat{L}_i = -1.431 + 0.024\,X_{age} + 0.079\,D_{female} - 0.015\,X_{femage}$$
$$= -1.431 + 0.024\,X_{age} + 0.079\,(1) - 0.015\,X_{(1)age}$$
$$= -1.352 + 0.009\,X_{AGE}$$

The graph shows how the expected logits vary with age for both genders, revealing that endorsing the traditional response to fepreschd rises about 267% more per year of age for men than for women (0.024 versus 0.009). Younger men are slightly more traditional than younger women, but older men are much more traditional than older women!

```
generate logitwomen = -1.352 + 0.009*age if female ==1
generate logitmen = -1.431 + 0.024*age if female ==0
twoway (line logitwomen age) (line logitmen age)
```



2. CENTERED PRODUCT TERMS

<u>Bilinear interaction</u> has been the traditional OLS regression technique for estimating the interaction effect of two continuous independent variables on a dependent measure. By extension, the method can also be applied to logistic regression. The procedure involves multiplying two predictors, then adding this product term to the equation along with the original measures (whose parameters are referred to as the "main effects"):

$$\hat{L}_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 X_2)$$

If the third coefficient is significant, it indicates that the main effect of X_1 on the dependent variable is <u>conditional on</u> (i.e., varies according to) the level of X_2. Similarly, the effect of X_2 is conditioned by the level of X_1. More below on the interpretation of interaction effects.

The bilinear multiplicative method frequently generates substantial multicollinearity, that is, high correlations among variables. As a result the parameter estimates may be accompanied by huge standard errors, which makes meaningful substantive interpretations difficult. A preferred solution is to <u>center</u> the two continuous predictors before computing their product term.* This procedure usually reduces the magnitude of the correlations between the multiplicative interaction term and its component predictors.

> In Stata regress postestimation, the command estat vif will produce variable inflation factor (VIF) scores for the independent variales that indicate the presence of multicollinearity. VIF values close to 1.00 indicate that multicollinearity is not problematic. A comparison of VIFs in equations using noncentered versus centered specifications can reveal dramatic reductions in multicollinearity with the latter method. VIF is not available for logistic regression.

The centering procedure is simple: Form the deviation of both variable's scores from their respective means, then multiply them and store the product in a third variable. Use all three centered measures as predictors of a dichotomous dependent variable.

---

* Pages 30-33 in Jaccard, James, Robert Turrisi and Choi K. Wan. 1990. *Interaction Effects in Multiple Regression.* Newbury Park, CA: Sage.

An example illustrating this method is the age-education interaction effect on happyd. Find the means of age and educ; create variables centered around each mean; create the interaction term by multiplying these centered variables; estimate a logistic regression equation with all three predictors.

summarize educ age

```
Variable |     Obs         Mean     Std. Dev.      Min        Max
---------+-----------------------------------------------------
    educ |    2018     13.43211     3.078964         0         20
     age      2013      47.7084     17.35084        18         89
```

generate educctr = (educ - 13.43211)
generate agectr = (age – 47.7084)
generate educage = educctr*agectr
logistic happyd educctr agectr educage, coef

```
Logistic regression     umber of obs    =    2000
                            LR chi2(3) =   22.51
                           Prob > chi2 =  0.0001
Log likelihood = -1202.7509  Pseudo R2 =  0.0093
------------------------------------------------
 happyd |      Coef.    Std. Err.       z     P>|z|
--------+---------------------------------------
 agectr |    .0078516   .0028814      2.72    0.006
educctr |    .0621825   .0167368      3.72    0.000
educage |   -.0016152   .0009355     -1.73    0.084
  _cons |   -.8850285   .0497711    -17.78    0.000
------------------------------------------------
```

Although main effects of age and education are highly significant, the centered interaction term is significant at p < .05 only for a one-tailed hypothesis.

The substantive interpretation of a centered bilinear interaction effect is facilitated by: (1) choosing one of the two main predictors as “moderator variable”; (2) selecting low-, medium-, and high-scores of the second predictor; and (3) calculating the differing slopes of the second predictor.

I chose educctr as the moderator: what is the effect of agectr (i.e., of different generations) on happyd, holding constant the level of education? Given that one standard deviation of educ = 3.08 years, three plausible educctr scores are – 3.08 (i.e., 10.35 years of education), 0.00 (13.43 years), and +3.08 (16.51 years). Here are the slopes of centered agectr at those three centered educctr levels:

$$\hat{L}_i = -0.8850 + 0.0622\ X_{eductr}\ +\ 0.0079\ X_{agectr}\ -\ 0.0016\ X_{agect*eductr}$$

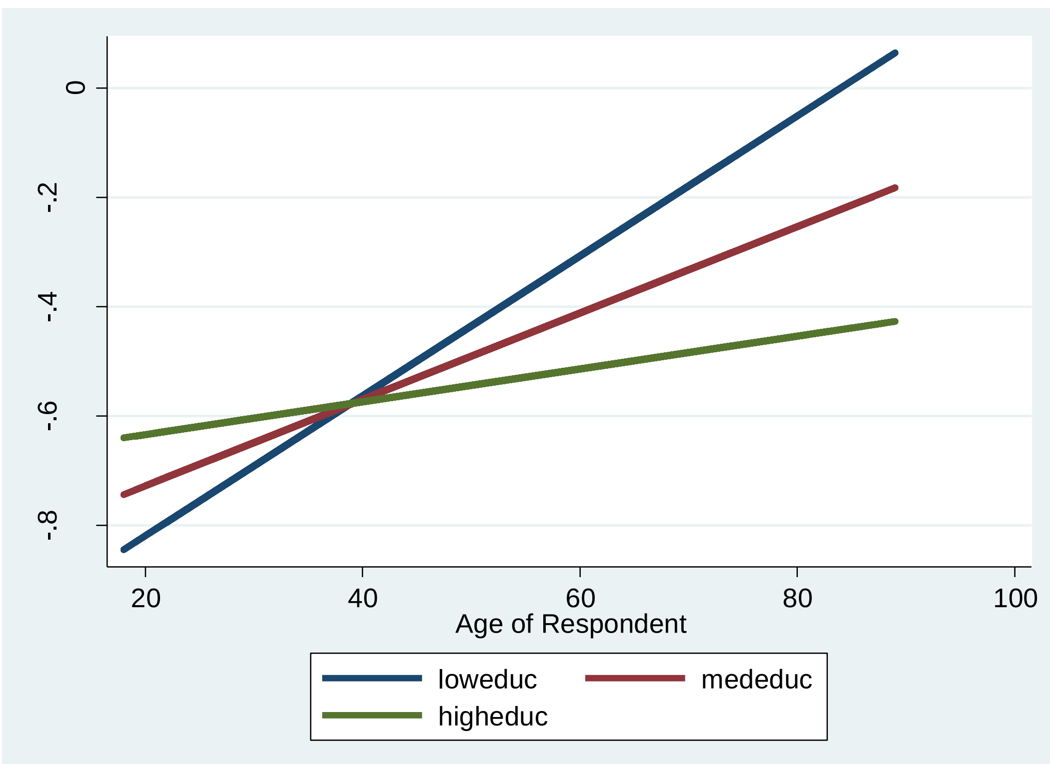
$$For\ educctr = -3.08:\quad \beta_{agectr} = 0.0079 + (-0.0016)(-3.08) = 0.0128$$

$$For\ educctr = \ \ 0.00:\quad \beta_{agectr} = 0.0079 + (-0.0016)(0) = 0.0079$$

$$For\ educctr = +3.08:\quad \beta_{agectr} = 0.0079 + (-0.0016)(+3.08) = 0.0030$$

Thus, for a one standard deviation increase in educ, the slope of the age effect on the log-odds of happyd falls by -0.0049. The figure below plots the expected happyd logit (= adjusted intercept + age slope) for the three education levels. To aid interpretation, I plotted these lines above the original, noncentered age values. What are *your* interpretations about how the age-education interaction affects the log-odds of happyd?

```
generate loweduc = -1.0766 + 0.0128*age
generate mededuc = -0.8850 + 0.0079*age
generate higheduc = -0.6934 + 0.0030*age
twoway (line loweduc age) (line mededuc age)(line higheduc  age),
ytitle(Logit of Happyd) xtitle(Age of Respondent)
```



3. COMPARING SUBSAMPLE EQUATIONS

Another approach to interaction involves estimating separate equations for samples from two populations, then performing a t-squared test of the difference in their estimated slope parameters, using the formula:

$$\mathbf{t}^2 = \frac{(\mathbf{b}_{\mathbf{k}_1} - \mathbf{b}_{\mathbf{k}_2})^2 - (\beta_{\mathbf{k}_1} - \beta_{\mathbf{k}_2})^2}{\mathbf{se}^2_{\mathbf{b}_{\mathbf{k}_1}} + \mathbf{se}^2_{\mathbf{b}_{\mathbf{k}_2}}}$$

where the squared difference in corresponding regression parameters is divided by the sum of their squared standard errors. Squaring a t-score is equivalent to chi-square for 1 degree of freedom (i.e., a Wald statistic). If you set a region of rejection at $\alpha = .05$, then the critical value of $t^2 = (1.96)^2 = 3.84$ for a two-tailed alternative to a null hypothesis that the two β's are equal in the populations.

Test whether party identification, education, and Southern residence effects on the logit of conservative political views differ for men and women.

```
logit polviewsd partyid educ south if female==1
logit polviewsd partyid educ south if female==0
```

	WOMEN		MEN	
	B	se	B	se
constant	-1.329	0.367	-2.243	0.383
partyid	0.490	0.038	0.467	0.040
educ	-0.068	0.026	0.006	0.026
south	0.345	0.154	0.256	0.159
(N)	(576)		(610)	

Neither the partyid nor south effects differ by gender. But, for the educ parameters, the test statistic is:

$$t^2 = \frac{((-0.068) - (0.006))^2}{(0.026)^2 + (0.026)^2} = 4.05$$

Hence, we can reject a two-tailed null hypothesis that the education effect on conservatism differs for men and women, with the probability of a Type I error (false rejection error) p < .05.

IV. MODELS FOR COUNTS

This section briefly discusses multivariate equations with dependent variables that are neither continuous nor dichotomous. Many dependent variables are counts, nonnegative integers for the number of activities, events, or occurrences. For example, how many children live in a household; number of automobiles per family; how many new firms started in an industry; the number of trips to national parks. Models for count data include Poisson regression, negative binomial regression, zero-inflated count models (both Poisson & NB regression), zero-truncated count models, hurdle models, and random-effects count models. Time allows for examining only the first three models. The last topic in this module is the censored regression or Tobit model.

1. POISSON REGRESSION

In many social analyses, the dependent variable is better conceived as a <u>discrete count</u> of the number of occurrences over an observation period, rather than a measure of continuous variation or as a simple dichotomous choice. For example: the ideal number of children; the number of automobiles owned; how many acquaintances with AIDS; voluntary association memberships; number of traffic accidents; number of major earthquakes. Once again, an OLS regression approach is unsatisfactory. The Poisson regression (named for a French mathematician, not a fish) uses a probability density function whose expected mean and variance are equal:

$$P(y \mid \mu) = \frac{e^{-\mu}\mu^{y}}{y!}$$

where y is the observed count, μ is the expected count (and variance), and <u>y!</u> is the factorial of the discrete number of events (e.g., 3! = (3)(2)(1) = 6).

A classic application appeared in Ladislaus von Bortkiewicz's *The Law of Small Numbers* (1898), a table showing the number of deaths from mule-kicks in 10 Prussian cavalry corps over 20 years of observation (200 corps-years):*[*]

Y_i	n_i
0	109
1	65
2	22
3	3
4	1

*** From page 292 in James S. Coleman. 1964. *Introduction to Mathematical Sociology*. Glencoe, IL: Free Press.**

Applying the formula $\hat{p}_y = (e^{-\mu}\mu^y)/y!$, where μ is the sample mean (0.61 deaths per corps-year), yields the following estimated frequencies: no deaths = 108.7; one death = 66.3; two = 20.2; three = 4.11; and four = 0.6. These values appear to approximate the observed data very closely.

Stata's Poisson regression program models the natural log of μ as a function of K independent variables:

$$\ln \mu = \sum_{j=1}^{K} \beta_j X_{ji}$$

Equivalently, by exponentiating both sides:

$$\mu = e^{\sum \beta_j X_{ji}}$$

The 2008 GSS respondents were asked to tell "the names of the people who usually live in this household." Here's the distribution of hompop:

table hompop

number of persons in household	Freq.
1	523
2	701
3	322
4	277
5	125
6	54
7	13
8	6
9	1
11	1

The Poisson regression equation of hompop on six predictors:

poisson hompop educ age female black catholic south

```
Poisson regression                                Number of obs   =       1999
                                                  LR chi2(6)      =     243.29
                                                  Prob > chi2     =     0.0000
Log likelihood = -3317.537                        Pseudo R2       =     0.0354
```

hompop	Coef.	Std. Err.	z	P>|z|
educ	-.0146108	.0046827	-3.12	0.002
age	-.0125042	.000848	-14.75	0.000
female	.0489142	.0283084	1.73	0.084
black	-.0477458	.0419144	-1.14	0.255
catholic	.0366189	.0338193	1.08	0.279
south	.069378	.0294924	2.35	0.019
_cons	1.642186	.080971	20.28	0.000

Three two-tailed and one one-tailed null hypotheses can be rejected at $p < .05$ or lower. The positive coefficients indicate a higher rate (large household size) for women and Southern residents, while negative coefficients indicate lower household sizes for older and better-educated respondents.

Request the exponentiated Poisson coefficients and their standardized values with this command:

listcoef educ age female black catholic south, help

```
poisson (N=1999): Factor Change in Expected Count
 Observed SD: 1.4166955
----------------------------------------------------------------
   hompop |        b          z     P>|z|     e^b    e^bStdX    SDofX
----------+-----------------------------------------------------
     educ |  -0.01461    -3.120   0.002   0.9855    0.9560   3.0827
      age |  -0.01250   -14.745   0.000   0.9876    0.8052  17.3274
   female |   0.04891     1.728   0.084   1.0501    1.0247   0.4987
    black |  -0.04775    -1.139   0.255   0.9534    0.9837   0.3440
 catholic |   0.03662     1.083   0.279   1.0373    1.0156   0.4223
    south |   0.06938     2.352   0.019   1.0718    1.0339   0.4806
----------------------------------------------------------------
      b = raw coefficient
      z = z-score for test of b=0
  P>|z| = p-value for z-test
e^b     = exp(b)=factor change in expected count for unit increase in X
e^bStdX = exp(b*SD of X)=change in expected count for SD increase in X
  SDofX = standard deviation of X
```

As in logistic regression, exponentiated coefficients help with interpretation. Similar to odds ratios, they're called "incidence rate ratios." Recall that exponentiated coefficients are multiplicative, raising or lowering the odds proportionally depending on whether they're above or below 1.000. Thus, being female increases the expected count of people in the household by (1.050 – 1.000)(100%) = +5.0% relative to males. Each year of educ reduces the expected incidence by (0.985 -1.000)(100%) = -1.5%.

Stata calculates percent changes for a unit of predictor X and for one standard deviation of X. With the latter, age has the largest impact (-19.5%):

listcoef educ age female black catholic south, percent help

```
poisson (N=1999): Percentage Change in Expected Count
----------------------------------------------------------------
   hompop |        b          z     P>|z|      %      %StdX      SDofX
----------+-----------------------------------------------------
     educ |  -0.01461    -3.120   0.002   -1.5      -4.4      3.0827
      age |  -0.01250   -14.745   0.000   -1.2     -19.5     17.3274
   female |   0.04891     1.728   0.084    5.0       2.5      0.4987
    black |  -0.04775    -1.139   0.255   -4.7      -1.6      0.3440
 catholic |   0.03662     1.083   0.279    3.7       1.6      0.4223
    south |   0.06938     2.352   0.019    7.2       3.4      0.4806
----------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
       % = percent change in expected count for unit increase in X
   %StdX = percent change in expected count for SD increase in X
   SDofX = standard deviation of X
```

Another way to obtain the incidence rate ratios is by appending "irr" at the end of the Poisson command:

poisson hompop educ age female black catholic south, irr

```
------------------------------------------------------
   hompop |        IRR   Std. Err.       z     P>|z|
----------+-------------------------------------------
     educ |   .9854954   .0046148    -3.12    0.002
      age |   .9875736   .0008375   -14.75    0.000
   female |    1.05013   .0297275     1.73    0.084
    black |   .9533761   .0399602    -1.14    0.255
 catholic |   1.037298   .0350806     1.08    0.279
    south |   1.071841   .0316112     2.35    0.019
------------------------------------------------------
```

Stata will calculate the predicted count values for each R. Use these commands, where "e(sample)" means the effective sample, excluding cases with missing values:

predict prhompop if e(sample), n
list hompop prhompop if e(sample)

```
       +-------------------+
       | hompop   prhompop |
       |-------------------|
    1. |      2   2.436998 |
    2. |      1   2.116504 |
    3. |      2   2.263246 |
    4. |      3   2.852501 |
    5. |      3   2.732911 |
       |-------------------|
    6. |      2    1.64375 |
    7. |      5   3.462719 |
    8. |      5   2.767299 |
    9. |      1   2.399398 |
   10. |      1   2.251551 |
```

Some predictions fit the observations closely (#3,4) but others are well of the mark (#7,8). Overall, the equation fits the data poorly (see Pseudo-R^2).

Exposure Time

To this point, an implicit assumption is that everyone is at risk of an event occurring for the same amount of time. In the example, each R had the same period in which to acquire households. More typically, different Rs are observed for different <u>exposure times</u>. Young people have less time in

which to accumulate events; for example, young criminals commit fewer crimes than older ones; young academic publish fewer papers than older ones; veteran soldiers with more time in combat zones receive more wounds than newbies.

Different exposure times can be incorporated into count models. Modify a multiplicative Poisson regression equation to include the natural log of the exposure time:

$$\mu_i t_i = e^{\sum_{j=1}^{K} \beta_j X_{ji} + \ln(t_i)}$$

where t_i is the exposure time for case *i*.

Suppose we assume that R's age is a reasonable proxy for exposure time. Then add the "exposure(varname)" option to the Poisson command:

poisson hompop educ female black catholic south, exposure(age)

```
Poisson regression                        Number of obs  =     1999
                                          LR chi2(5)     =    31.84
                                          Prob > chi2    =   0.0000
Log likelihood = -4163.3342               Pseudo R2      =   0.0038
------------------------------------------------------------------------------
      hompop |      Coef.   Std. Err.       z     P>|z|
-------------+----------------------------------------------------------------
        educ |  -.0014149   .0044842    -0.32    0.752
      female |   .0338835   .0283118     1.20    0.231
       black |   .1391718   .0415367     3.35    0.001
    catholic |   .0806812   .0338356     2.38    0.017
       south |   .1064363   .0295357     3.60    0.000
       _cons |  -3.012744   .0677837   -44.45    0.000
         age | (exposure)
------------------------------------------------------------------------------
```

To show how exposure(varname) operates, the same results occur if the natural log of age, lnage, is added as an independent variable and constrained to 1:

generate lnage=ln(age)
constraint define 1 lnage=1
poisson hompop lnage educ female black catholic south,
constraint(1)

```
Poisson regression                       Number of obs   =      1999
                                           Wald chi2(5)  =     32.40
Log likelihood = -4163.3342              Prob > chi2     =    0.0000
 ( 1)  [hompop]lnage = 1
-----------------------------------------------------------------------
      hompop |      Coef.   Std. Err.        z     P>|z|
-------------+---------------------------------------------------------
       lnage |          1           .        .         .
        educ |  -.0014149    .0044842    -0.32     0.752
      female |   .0338834    .0283118     1.20     0.231
       black |   .1391718    .0415367     3.35     0.001
    catholic |   .0806812    .0338356     2.38     0.017
       south |   .1064363    .0295357     3.60     0.000
       _cons |  -3.012744    .0677837   -44.45     0.000
-----------------------------------------------------------------------
```

The substantive results change when exposure time is held constant. Now the education and gender effects are no longer significant, while black, catholic, and south are all associated with larger household sizes.

2. NEGATIVE BINOMIAL REGRESSION

Poisson regression also makes an assumption of "equidispersion" – that the mean and variance of the dependent variable are identical. Few real data can meet this requirement; more often, "overdispersion" occurs – the variance is much larger than the mean. The result is underestimated standard errors and false rejection of the null hypothesis. Overdispersion most often occurs because of highly skewed dependent variables, with many more zeros than expected. For example, most academics have no publications, but a few have extremely high article counts.

To correct for overdispersion, the negative binomial regression model (NBRM) adds an error term that is presumed uncorrelated with the X's:

$$\mu = e^{\sum_{j=1}^{K} \beta_j X_{ji} + \varepsilon_i)}$$

To identify the model, the expected value of the error = 1, equivalent to an expected value of 0 in the logistic regression model. The error term is unknown, but by assuming it is has a gamma distribution the model becomes mathematically tractable. The NBRM is:

$$P(y \mid x) = \frac{\Gamma(y+\alpha^{-1})}{y!\Gamma(\alpha^{-1})} \left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu} \right)^{\alpha^{-1}} \left(\frac{\mu}{\alpha^{-1}+\mu} \right)^{y}$$

where Γ is the gamma function. The parameter, α, determines the degree of dispersion in the predictions, with larger values indicating a greater spread in the data.

The NBRM example below analyzes the number of R's sex partners in the past 12 months (partners12) as the dependent count variable. Here is the distribution:

```
generate p12=partners
recode p12(0=0)(1=1)(2=2)(3=3)(4=4)(5=8)(6=15)(7=60)(8=120)
      (nonmissing=.), generate(partners12)
table partners12
----------------------
RECODE of |
p12       |      Freq.
----------+-----------
        0 |        415
        1 |      1,092
        2 |        120
        3 |         55
        4 |         25
        8 |         33
       15 |         11
       60 |          3
      120 |          2
----------------------
```

Run the NBRM and store its estimates in a file for comparison to Poisson regression estimates:

```
nbreg partners12 educ female black catholic south
estimates store NBRM
Negative binomial regression    Number of obs  = 1750
                                 LR chi2(5)    = 90.72
Dispersion     = mean            Prob > chi2 = 0.0000
Log likelihood = -2800.3183      Pseudo R2    = 0.0159
-----------------------------------------------------
  partners12 |     Coef.   Std. Err.      z    P>|z|
-------------+---------------------------------------
        educ |  -.0114919   .0106799    -1.08   0.282
      female |  -.4540601   .0600383    -7.56   0.000
       black |   .2800478   .0855524     3.27   0.001
    catholic |  -.2263316   .0752474    -3.01   0.003
       south |   .0878847   .0623771     1.41   0.159
       _cons |   .6786853   .1557313     4.36   0.000
-------------+---------------------------------------
    /lnalpha |  -.1981557   .0550963
-------------+---------------------------------------
       alpha |   .8202421   .0451923
-----------------------------------------------------
Likelihood-ratio test of alpha=0:
chibar2(01) = 2243.78 Prob>=chibar2 = 0.000
```

poisson partners12 educ female black catholic south
estimates store PRM
estimates table PRM NBRM, b(%9.3f) t label varwidth(32)
drop(lnalpha:_cons) stats(alpha N)

```
-----------------------------------
   Variable |     PRM         NBRM
------------+----------------------
       educ |    -0.017       -0.011
            |     -2.39        -1.08
     female |    -0.473       -0.454
            |    -11.52        -7.56
       race |     0.307        0.280
            |      5.63         3.27
   catholic |    -0.244       -0.226
            |     -4.56        -3.01
      south |     0.072        0.088
            |      1.72         1.41
   Constant |     0.761        0.679
            |      7.49         4.36
------------+----------------------
      alpha |                  0.820
          N |      1750         1750
-----------------------------------
                         legend: b/t
```

The corresponding parameter estimates for both models are close, but the t-test values for the NBRM are consistently smaller than the Poisson model. As discussed above, when overdispersion occurs, the Poisson standard errors are biased downward and the t-test values are inflated.

When α = 0, the negative binomial equation is identical to the Poisson regression. Thus, comparing the two models' results permits a one-tailed test of the overdispersion hypothesis:

$H_0: \alpha = 0$
$H_1: \alpha > 0$

The test statistic "chibar2(01)" is at the bottom of the NBRM output. Its computation also requires the log likelihood for the corresponding Poisson model (which is -3922.21):

$$G^2 = 2(\ln L_{NBRM} - \ln L_{Poisson})$$
$$= 2(-2800.32 - (-3922.21)) = 2243.78$$

Clearly the null hypothesis must be rejected! The conclusion is that overdispersion very likely occurs in the partners12; therefore the negative binomial regression model estimates are preferred to the Poisson estimates.

As a general principle, count data should be analyzed by both Poisson regression and NBRM and alpha tested for overdispersion. If the null hypothesis is not rejected, report the Poisson regression results. It makes fewer assumptions than NBRM, which assumes a gamma-distributed error term.

3. ZERO-INFLATED COUNT MODELS

If the dependent variable has many zeros, if may be highly skewed. In such instances, NBRM is preferred to Poisson regression because of overdispersion. However, in the presence of enormous numbers of zeros, NBRM tends to under-predict zeros and hence not fit the data well. For example, number of arrests last year are mostly 0 for a sample of the general population. For such data structures, zero-inflated Poisson or NBRM are better.

Zero-inflated count models assume two latent (unobserved) groups: (1) in the "Always Zero" group (Group A) individuals have a count of 0 with probability = 1 (i.e., certainty); (2) in the "Not Always Zero" group (Group -A) respondents may have a zero count but have a nonzero probability of a positive count. For example, people with no computer spend 0 hours visiting Websites, but for people with computers, the hours may range from 0 to 80 or more per week. Zero-inflate count models are estimates in three stages: the probability of being in Group A is modeled with a logit regression; the counts in Group -A are modeled with either a Poisson regression or NBRM; and the two groups are mixed according to their proportions in the population to determine the overall rate.

In this example, the number of hours worked per week at 0 for people with no jobs, and recoded into 10-hour intervals for employed people:

```
recode hrs1(0=0)(1/19=1)(20/29=2)(30/39=3)(40/49=4)(50/59=5)
        (60/69=6)(70/79=7)(80/89=8), generate(hourswork)
replace hourswork = 0 if (wrkstat > 2)
```

```
tabulation:  Freq.   Value
               809   0
                66   1
               106   2
               161   3
               552   4
               156   5
               107   6
                23   7
                32   8
                11   .
```

To estimate a zero-inflated Poisson model or NBRM, the dependent variable is followed by the set of independent variables predicting the number of hours worked by Rs in the Not Always Zero group, then by a list of the inflation variables that predict whether R is in the Always Zero group. The sets of predictors can be identical but don't have to be.

zip hourswork educ age female black south catholic,
inflate(educ age female black south catholic)

```
Zero-inflated Poisson regression                  Number of obs =   1988
                                                  Nonzero obs   =   1190
                                                  Zero obs      =    798
Inflation model = logit                           LR chi2(6)    =  52.93
Log likelihood  = -3333.374                       Prob > chi2   =  0.000
------------------------------------------------------------------------
   hourswork |      Coef.   Std. Err.          z       P>|z|
-------------+----------------------------------------------------------
hourswork    |
        educ |   .0037302   .0053495        0.70       0.486
         age |  -.0005563   .0011879       -0.47       0.640
      female |  -.2168681   .0310574       -6.98       0.000
       black |  -.0407135   .0454134       -0.90       0.370
       south |   .0276841   .0318736        0.87       0.385
    catholic |   .0066872   .0365337        0.18       0.855
       _cons |   1.425691    .095506       14.93       0.000
-------------+----------------------------------------------------------
inflate      |
        educ |  -.1088842   .0180589       -6.03       0.000
         age |   .0482937    .003432       14.07       0.000
      female |   .7739544   .1076439        7.19       0.000
       black |  -.1761624   .1609218       -1.09       0.274
       south |  -.0776404   .1124911       -0.69       0.490
    catholic |  -.3512207   .1321932       -2.66       0.008
       _cons |  -1.661118   .3189834       -5.21       0.000
------------------------------------------------------------------------
```

Which variables predict R's membership in the Always Zero group? Do they differ from the predictors of how many hours worked by Rs in the Not Always Zero group?

4. CENSORED REGRESSION

Sometimes the distribution of a continuous dependent variable is <u>censored</u>. Information is not available about those cases having values above and/or below a particular threshold. All Rs are assigned the value of the threshold, even if some may be far above or below that value. For example, the 2008 GSS records its highest income category as "$150,000 or over," which encompasses people earning precisely that amount and multimillionaires.

Three types of censoring can occur, depending on where the threshold is located on the continuous dependent variable's scale:

1. **Right-censored (upper limit): No precision among cases above the threshold (the income example above), or the threshold is an artificial constraint on higher values. "Attendance at Twins home games" is constrained by Target Field's capacity of 40,000 seats.**

2. **Left-censored (lower limit): The lower threshold is a qualitative barrier to the continuous measure. "Price paid for automobile" is 0 for Rs who didn't purchase a car.**

3. **Double-censored: Both upper and lower thresholds exist. SAT and GRE scores are bounded between 200 and 800.**

Applying OLS regression to such dependent variables produces biased parameter estimates if the censored cases are excluded or are given an imputed value. To analyze censored data requires a multivariate model that explicitly takes the censored cases into account.

Tobit analysis – more accurately, the censored regression model – is a multivariate method for limited dependent variables that permits unbiased parameter estimates by including the censored cases, while treating them differently from cases with observed variation. The method was proposed decades ago by James Tobin, a subsequent Nobel prize economist, and given the name "tobit" (for "Tobin's probit") by Arthur Goldberger.[*]

*** Tobin, James. 1958. "Estimation of Relationships for Limited Dependent Variables." *Econometrica* 26:24-36.**

The tobit model predicts a "latent" (unobserved) value of a censored dependent variable as a linear function of one or more predictors, with a normally distributed error term:

$$\mathbf{y}_{\mathbf{i}}^{*} = \boldsymbol{\beta}\mathbf{X}_{\mathbf{i}} + \boldsymbol{\varepsilon}_{\mathbf{i}}$$

The expected score of the ith observed case depends on whether it is an uncensored case:

$$\hat{\mathbf{y}}_{\mathbf{i}} = \hat{\mathbf{y}}_{\mathbf{i}}^{*} = \mathbf{b}\mathbf{X}_{\mathbf{i}}$$

or a censored case:

$$\hat{\mathbf{y}}_{\mathbf{i}} = 0_{\mathbf{i}}$$

The example below analyzes the 2008 GSS occupational prestige scores. About 40% of Rs were not in the labor force and could have been dropped. However, let's include those cases by coding them 0 and designating them as left-censored in the tobit command. (If the cases are right-censored, use "ul(#)" to indicate the threshold value; for double-censoring, "ll(#)" and "ll(#)" are both used.)

```
generate prest=prestg80
replace prest=0 if (wrkstat >2)
tobit prest educ age female black south catholic, ll(0)
Tobit regression                 Number of obs   =    1976
                                    LR chi2(6)   =  409.52
                                    Prob > chi2  =  0.0000
Log likelihood = -6436.6167         Pseudo R2    =  0.0308
-----------------------------------------------------
       prest |      Coef.   Std. Err.      t    P>|t|
-------------+---------------------------------------
        educ |   3.305408   .2798588    11.81   0.000
         age |  -.7349894   .0512137   -14.35   0.000
      female |  -12.19587   1.637457    -7.45   0.000
       black |   1.430946   2.436438     0.59   0.557
       south |   1.477284   1.720747     0.86   0.391
    catholic |   5.215537   1.970816     2.65   0.008
       _cons |   11.56908   4.829994     2.40   0.017
-------------+---------------------------------------
      /sigma |   33.64138   .7722422
-----------------------------------------------------
 Obs. summary: 798 left-censored obs  at prst<=0
              1178 uncensored observations
                 0 right-censored observations
```

Tobit coefficients are interpreted the same way as OLS regression coefficients. Which predictors increase or decrease the prestige of R's job and by how many points per unit of X?

The "sigma" coefficient at the bottom of the tobit output is the estimated standard error of the regression and is comparable to the root mean squared error in an OLS regression. However, some statisticians argue that it has no substantive interpretation.