INTRODUCTION TO BINARY LOGISTIC REGRESSION

Binary logistic regression is a type of regression analysis that is used to estimate the relationship between a dichotomous dependent variable and dichotomous-, interval-, and ratio-level independent variables.

Many different variables of interest are dichotomous – e.g., whether or not someone voted in the last election, whether or not someone is a smoker, whether or not one has a child, whether or not one is unemployed, etc.

These types of variables are often referred to as discrete or qualitative. Many discrete or qualitative variables can be thought of as events.

Dichotomous or dummy variables are usually coded 1, indicating “success” or “yes,” and 0, indicating “failure” or “no.” The mean of a dichotomous variable coded 1 and 0 is equal to the proportion of cases coded as 1, which can also be interpreted as a probability.

1 1 1 1 1 1 0 0 0 0
mean = 6 / 10 = .6 = the probability that any 1 case out of 10 has a score of 1

For quite a while, researchers used OLS regression to analyze dichotomous outcomes. This was based on the idea that predicted values ($\hat{y}$) – based on the regression results – generally range from 0 to 1 and are equivalent to predicted probabilities, predicted proportions, and predicted percents of “success” given values on the independent variables.

In other words, if we regressed a dummy variable, voted or not, on education and got the estimate b = .025, then we could say that a one-unit increase in education increases the probability of voting by .025. Equivalently, a one-unit increase in education increases the proportion voting by .025. Finally, a one-unit increase in education increases the percent voting by 2.5 percent.

Due to a number of conceptual and statistical problems, however, people no longer use OLS regression to analyze dichotomous dependent variables. There are a number of alternative approaches to modeling dichotomous outcomes including logistic regression, probit analysis, and discriminant function analysis.

Logistic regression is by far the most common, so that will be our main focus. Additionally, we will focus on binary logistic regression as opposed to multinomial logistic regression – used for nominal variables with more than 2 categories.

OLS Regression with a Dichotomous Dependent Variable

What is wrong with using OLS regression with dichotomous dependent variables? There are a number of problems.

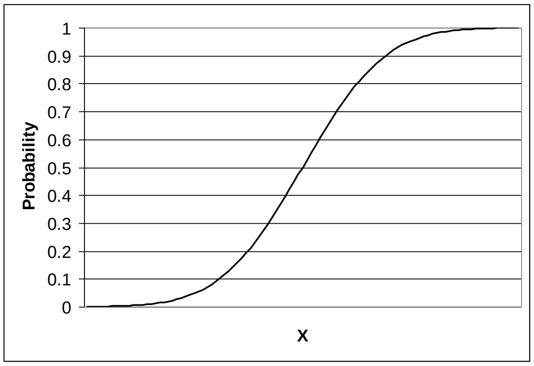
1. One of the regression assumptions that we discussed is that the dependent variable is quantitative (at least at the interval level), continuous (can take on any numerical value), and unbounded.

A person's score on the dependent variable is assumed to be a function of their score on each independent variable. Therefore, the dependent variable must be free to take on any value that is predicted by the combination of independent variables. If the dependent variable does not meet these requirements (e.g., it is dichotomous), then predicted scores on the dependent variable may lie outside possible limits. When you use OLS regression with a dichotomous dependent variable, predicted probabilities (based on the estimated OLS regression equation) are not bounded by the values of 0 and 1.

Why is this a problem? In the real world, probabilities can never be less than 0 and can never be greater than 1. With dummy dependent variables and OLS regression, it is not uncommon for predicted probabilities to be less than 0 and greater than 1. The likelihood of this increases as the difference between the number of successes and failures increases. In other words, if the split is 90% have a score of 1 and 10% have a score of 0, then you will probably experience impossible predicted probabilities.

2. Another OLS multiple regression assumption is that the relationship between Y and X is linear and additive in the population. Our estimates cannot be very good if we assume that the true relationship is linear and additive and we specify a linear and additive relationship when, in fact, in the population, the relationship is non-linear and/or non-additive

In many cases, it is not unreasonable to assume that the relationship between two variables is non-linear. The example that Pampel uses in the book is that of income and home ownership. A $10,000 increase in income probably increases the probability of owning a home more for someone with an initial income of $40,000 than someone with an initial income of $0. Also an additional $10,000 probably does not have much of an influence on the likelihood that a very rich person owns a home – e.g., earning $1,001,000 versus earning $1,000,000. So a more appropriate functional form (rather than a line) might be an s-shaped curve:



This type of a curve suggests that one-unit changes in the independent variable have different effects on the dependent variable at different levels of the independent variable. It takes a much larger increase in X to have the same effect on Y at extreme ends of the curve. One way to think about this is to consider the fact that the slope of this curve changes at different values of Y. In essence, there are a number of different perpendicular lines that one can draw.

3. Another regression assumption is that the error term is normally distributed. Remember that the error term summarizes all of the causes of the dependent variable not included in the model as well as errors in the functional form of the equation, measurement error, and the randomness in human behavior.

The assumption of normality allows you to do hypothesis testing. If the error term is not normally distributed, then we cannot use z (t) to find the probability under the curve.

The error term is not normally distributed when you use OLS regression with a dichotomous dependent variable because, for any value of X, there are only two possible values that the residuals can take. A residual is defined as the observed value on the dependent variable minus the predicted value given X.

Here's an example:
Consider 10 people with a value of 2 on the independent variable and an estimated regression equation of:
$\hat{y}_i = .03 + .48 * x_i$

The residual is equal to $y_i – \hat{y}_i$, where $\hat{y}_i = a + (b * x_i)$.
So after substituting, the residual is equal to $y_i – (a + (b * x_i))$.

For the value X = 2, there are only 2 possible values for the residual because there are only two possible observed values for Y (1 and 0).
$1 – (.03 + .48 * 2) = .01$
$0 – (.03 + .48 * 2) = -.99$

Thus, for any value of x, there are only two possible residuals so the distribution is not normal.

4. Another assumption is that of homoskedasticity – that the variance of the error term is constant across all values of the independent variables. Homoskedasticity means that the predicted values of the dependent variable are as good (or as bad) at all levels of the independent variable.

This is violated because the residuals vary with the value of x. The linear OLS regression consistently underestimates the slope at moderate levels of x and consistently overestimates the slope at extreme levels of x.

Heteroskedasticity leads to biased estimates of the standard errors, which we use in our t tests. Poor estimates increase the chance of drawing incorrect conclusions in hypothesis testing.

The Logit Transformation

So what can we do? As I mentioned earlier, many topics of interest are dichotomous. Logistic regression uses the logit transformation to linearize the non-linear relationship between X and the probability of Y. It does this through the use of odds and logarithms. So, the logit is a nonlinear function that represents the s-shaped curve. Let's look more closely at how this works. ['Generalized linear models' refers to a class of models that uses a link function to make estimation possible. The logit link function is used for binary logistic regression. Other link functions are used for other types of variables].

Probabilities express the likelihood of an event as a proportion of both occurrences and non-occurrences. In other words, probabilities are defined as the number of occurrences divided by the number of occurrences plus the number of non-occurrences. So, if you have a sample of 4,000 people and 3000 are married, the probability of being married is .75 (there is a 75% change of being married): 3000 / (3000 + 1000) = .75. Probabilities cannot be less than 0 and cannot be greater than 1. In other words, they are bounded by 0 and 1.

Odds, by contrast, are defined as the likelihood of occurrence divided by the likelihood of non-occurrence. Thus, the odds of being married for our example is: 3000 / 1000 = 3. What difference does dividing only by the number of non-occurrences make? It removes the upper limit of 1. But wait…that's not all…odds are also non-linear. Consider the examples in the Pampel text (p. 11):

The same change (a .1 increase in P) leads to increasingly large increases in the odds. Notice that the odds ratio is still bounded at the lower end. It is impossible for the odds to fall below 0. So, transforming the probabilities into odds has removed the upper limit.

P	.1	.2	.3	.4	**.5**	.6	.7	.8	.9
1-P	.9	.8	.7	.6	**.5**	.4	.3	.2	.1
Odds	.111	.250	.429	.667	**1**	1.500	2.333	4	9
Ln odds	-2.198	-1.386	-.846	-.405	**0**	.405	.847	1.386	2.197

The next step is to take the natural logarithm of the odds. Taking the natural log of the odds eliminates the floor of zero. Taking the natural log of a number above 0 and below 1 yields a negative number. Odds cannot be less than zero, but all odds less than 1 yield natural logs that are negative…the floor is gone. Taking the natural log of the number 1 yields 0. Finally, taking the natural logarithm of a number greater than 1 yields a positive number. However, notice that the distribution of logged odds is symmetrical around 0. The same size increase or decrease in the probabilities has the same absolute value in logged odds.

It is important to point out that the difference between the logged odds is not constant. For example, 2.197-1.386=.811 while 1.386-.847=.539. What does this mean? The logit transformation is stretching the distribution at extreme ends so the same one-unit change in X (the independent variable) leads to increasingly smaller gains in Y. In essence, it yields the s-shaped curve.

The linear relationship between X and the log odds is given by the following formula – cumulative logistic distribution function (you will see this in some statistical output):

Logged odds: $Ln\left(\frac{P_i}{1-P_i}\right) = b_0 + b_1 x_1 + b_2 x_2$

Notice that there is no error term in the model. The error term is not necessary. "The random component of the model is inherent in the modeling itself – the logit equation, for example, provides the expression for the probability that an event will occur. For each observation the occurrence or non-occurrence of that event comes about through a chance mechanism determined by this probability, rather than by a draw from a bowl of error terms" (Kennedy 1998, p. 234).

The right hand side of the equation looks like a normal linear regression equation, but the left hand side is the log odds rather than a probability. This equation can be re-written as follows:

Odds: $P_i / 1 - P_i = e^{b_o} * e^{b_1 x_{1i}} * e^{b_2 x_{2i}}$

Probabilities: $P_i = E(Y_i = 1 | X_{1i}, X_{2i}) = \frac{1}{1 + e^{-(b_0 + b_1 x_{1i} + b_2 x_{2i})}}$

Estimation

Fantastic…logistic regression allows us to estimate the relationship between dichotomous dependent variables and dichotomous, interval, and ratio independent variables. But…how does it work? Where do the estimates for β come from?

Logistic regression uses maximum likelihood estimation to generate estimates of β. Specifically, ML uses the observed data and probability theory to find the most likely or the most probable population value given the sample data (observations).

In logistic regression, the formula that is used to determine the population value most likely to yield the sample data is given by the likelihood function:

$$LF = \prod \left\{ P_i^{y_i} * (1 - P_i)^{1 - y_i} \right\}$$

The likelihood function is an expression for the likelihood of observing the pattern of occurrences (y=1) and non-occurrences (y=0) of an event in a given sample. In other words, it tells us the probability of getting our sample data from a population with probabilities equal to P_i.

The y_is above can be either 1 or 0, depending on the score of person i on the dependent variable. The P_is refer to predicted probabilities on the dependent variable (there is one for every person in the sample) given the person's scores on the independent variable(s). You plug predicted probabilities into this equation just like we plugged possible population proportions into the binomial probability distribution. The predicted probabilities are from a logistic regression model.

	y_i	P_i	$\{P_i^{yi} * (1-P_i)^{1-y_i}\}$	y_i	P_i	$\{P_i^{yi} * (1-P_i)^{1-y_i}\}$
	1	0.1	0.100000	1	0.8	0.800000
	1	0.2	0.200000	1	0.9	0.900000
	0	0.8	0.200000	0	0.1	0.900000
	0	0.9	0.100000	0	0.2	0.800000
LF			0.000400			0.518400

How does the software package do a logistic regression to get the predicted probabilities? It does this by using OLS regression to get a first set of estimates. These estimates yield predicted probabilities for each case, which are plugged into the likelihood function and generate a score for those estimates. The score is equal to the likelihood of obtaining the observed sample data (the combination of 1s and 0s) from a population with that particular set of slope estimates.

Next, based on sophisticated computer algorithms, it chooses a new estimate for β, which is used to generate a new set of predicted probabilities for each case. These are plugged into the likelihood function and generate a new score (a new probability) for this new estimate. This process is repeated over and over until the likelihood function is maximized – until increases in the number become extremely small with successive attempts. Believe me, you don't want to do this by hand.

In the book, Pampel talks about the log likelihood function. The log likelihood function is another version of the likelihood function – it is simply the log of the likelihood function. Why does the computer program use this instead of the likelihood function? Because it's easier to do addition than multiplication. By taking the log of the likelihood function, the different components of the equation can be added together instead of multiplied together. One rule of logarithms states that log (a*b) = log a + log b.

Summary

So to summarize, when we are interested in examining the relationship between a dichotomous dependent variable and dichotomous, interval, and ratio independent variables, we can use logistic regression. Logistic regression uses maximum likelihood to generate the estimates of the slopes. Logistic regression yields unbiased and efficient estimates of β and OLS regression does not.

BINARY LOGISTIC REGRESSION IN MULTILEVEL MODELING

The class of models for nominal, ordinal, and count dependent variables is known as generalized linear models. All use a link function to constrain estimates to fall within what is possible for that type of variable. The logit link function is used in binary logistic regression. This class of models can be estimated within a multilevel framework.

The level 1 model is:

$$n_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \ldots + \beta_{pj} X_{pij},$$

where n_{ij} is the logged odds: $\log\left(\frac{\phi_{ij}}{1-\phi_{ij}}\right)$

Notice that there is not an error term in the level 1 model. These models are already probabilistic so it would be redundant to include an error term at this level.

The level 2 model is (this could be the intercept or a slope):

$$\beta_{qj} = \gamma_{q0} + \sum_{s=1}^{Sq} \gamma_{qs} W_{sj} + u_{qj}$$

It is possible to estimate all of the same multilevel sub-models within multilevel binary logistic regression.

Example
Data are from the 2003 ISSP National Identity Module.

V49 Help minorities to preserve traditions: Some people say that it is better for a country if different racial and ethnic groups maintain their distinct customs and traditions. Others say that it is better if these groups adapt and blend into the larger society. Which of these views comes closer to your own?
0. It is better for society if groups maintain their distinct customs and traditions.
1. It is better if groups adapt and blend into the larger society.

			Freq.	Percent	Valid	Cum.
Valid	Maintain traditions	0	8515	39.08	47.46	47.46
	Adapt in society	1	9425	43.25	52.54	100
		Total	17940	82.33	100	
Missing		.	3850	17.67		
Total			21790	100		

cntryid	mean	N
d	0.443	1411
gb	0.804	828
a	0.572	835
h	0.387	905
i	0.627	969
irl	0.598	890
nl	0.706	1696
n	0.751	1240
s	0.806	1064
cz	0.503	899
slo	0.456	825
pl	0.466	986
bg	0.451	941
rus	0.184	1282
e	0.479	1077
lv	0.290	863
sk	0.380	1229
Total	0.525	17940

malem - female=0 male=1
agem is age measured in years
educm2 is education measured in years
EGP=Erikson, Goldthorpe, and Portocarero Nominal Class Categories
EGP123 (reference category) - higher and lower service and routine clerical and sales
EGP45 - independent and small employers
EGP711 - manual foremen skilled manual, semi-unskilled manual, farm workers, farmers, farm managers
EGP21 - students
EGP22 - unemployed
EGP2325 - homemakers, retirees, and others not in the labor force
WEUROPE – a dummy variable indicating the country is in Western Europe

The One-Way ANOVA with Random Effects Model

```
melogit adapt || cntryid: , intmethod(mcaghermite) intpoints(1)
```

```
Mixed-effects logistic regression               Number of obs      =     17940
Group variable:         cntryid                 Number of groups   =        17

                                                Obs per group: min =       825
                                                               avg =    1055.3
                                                               max =      1696

Integration method:     laplace

                                                Wald chi2(0)       =         .
Log likelihood = -11301.994                     Prob > chi2        =         .
------------------------------------------------------------------------------
       adapt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .1100984   .1855085     0.59   0.553    -.2534915    .4736883
-------------+----------------------------------------------------------------
cntryid      |
   var(_cons)|   .5800898   .2007773                      .2943624    1.143163
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =  2219.95 Prob>=chibar2 = 0.0000
```

The melogit command tends to run very slowly. This is due to the method of estimation that is used (adaptive quadrature within maximum likelihood). One of the sources of slowness is having large clusters (e.g., countries). Adding a bunch of random effects exacerbates the problem. Unfortunately, you often have to increase the number of integration points to get stable results. The higher the number of integration points leads to slower estimation. Rabe-Hesketh and Skrondal (2012) discuss this issue on page 523 and pages 537-541. They set the number of integration points to 30 for their examples.

The integration method in the output above (‘Laplace’) is selected by setting the integration points to 1. This method is less accurate, but faster (see Rabe-Hesketh and Skrondal 2012, p. 527). You have to use the mcaghermite integration method to set the integration points to 1.

Notice in the output that there is not a variance component at level 1. This is because there is not a level-one residual in logistic regression (it is probabilistic by nature).

One-way ANCOVA with Random Effects

```
melogit adapt malem agem EDUCM2 EGP45 EGP711 EGP21 EGP22 EGP2325 ||
cntryid: , intmethod(mcaghermite) intpoints(1)
```

```
Mixed-effects logistic regression               Number of obs      =     17419
Group variable:         cntryid                 Number of groups   =        17

                                                Obs per group: min =       787
                                                               avg =    1024.6
                                                               max =      1664

Integration method:     laplace

                                                Wald chi2(8)       =    208.16
Log likelihood = -10864.311                     Prob > chi2        =    0.0000
------------------------------------------------------------------------------
       adapt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       malem |   .0446208   .0343607     1.30   0.194    -.0227249    .1119665
        agem |   .0030156   .0013479     2.24   0.025     .0003739    .0056574
      EDUCM2 |   -.039286   .0052328    -7.51   0.000     -.049542   -.0290299
       EGP45 |   .2113438   .0792005     2.67   0.008     .0561136     .366574
      EGP711 |   .2534655   .0527061     4.81   0.000     .1501634    .3567675
       EGP21 |   .0146199   .0788345     0.19   0.853    -.1398928    .1691326
       EGP22 |   .0562829   .0729862     0.77   0.441    -.0867674    .1993331
     EGP2325 |    .243141   .0525976     4.62   0.000     .1400517    .3462303
       _cons |   .2397844   .2154053     1.11   0.266    -.1824022     .661971
-------------+----------------------------------------------------------------
cntryid      |
   var(_cons)|    .618678   .2140414                      .3140351    1.218853
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =  2247.52 Prob>=chibar2 = 0.0000
```

The coefficients are logit coefficients (i.e., logged odds).

Examples of interpretation:

- Each one year increase in age increases the logged odds of agreeing that it is better if groups adapt and blend into the larger society by .003.
- The logged odds of agreeing that it is better if groups adapt and blend into the larger society are .253 higher for blue-collar workers (EGP711 - manual foremen skilled manual, semi-unskilled manual, farm workers, farmers, farm managers) compared to white-collar workers (EGP123 - higher and lower service and routine clerical and sales)

You can calculate the odds coefficient by raising *e* to the power of the logged odds coefficient: e^{Logit}. For example (malem): $e^{.0446208} = 1.045631$

Alternatively, you can use the following command after running the original melogit command above (the ‘or’ stands for odds ratio): `melogit, or`

```
Mixed-effects logistic regression               Number of obs      =     17419
Group variable:         cntryid                 Number of groups   =        17

                                                Obs per group: min =       787
                                                               avg =    1024.6
                                                               max =      1664

Integration method:     laplace

                                                Wald chi2(8)       =    208.16
Log likelihood = -10864.311                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       adapt | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       malem |   1.045631   .0359286     1.30   0.194     .9775313    1.118475
        agem |    1.00302   .0013519     2.24   0.025     1.000374    1.005673
      EDUCM2 |   .9614757   .0050312    -7.51   0.000     .9516652    .9713874
       EGP45 |   1.235337   .0978394     2.67   0.008     1.057718    1.442783
      EGP711 |   1.288483   .0679109     4.81   0.000     1.162024    1.428704
       EGP21 |   1.014727   .0799955     0.19   0.853     .8694514    1.184277
       EGP22 |   1.057897   .0772118     0.77   0.441     .9168903    1.220589
     EGP2325 |   1.275248    .067075     4.62   0.000     1.150333    1.413728
       _cons |   1.270975   .2737747     1.11   0.266     .8332661     1.93861
-------------+----------------------------------------------------------------
cntryid      |
   var(_cons)|    .618678   .2140414                      .3140351    1.218853
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =  2247.52 Prob>=chibar2 = 0.0000
```

Examples of interpretation:

- Each one year increase in age increases the odds of agreeing that it is better if groups adapt and blend into the larger society by .3 percent.
- Each one year increase in education decreases the odds of agreeing that it is better if groups adapt and blend into the larger society by 3.9 percent.
- The odds of agreeing that it is better if groups adapt and blend into the larger society are 28.8 percent higher for blue-collar workers (manual foremen skilled manual, semi-unskilled manual, farm workers, farmers, farm managers) compared to white-collar workers (higher and lower service and routine clerical and sales)

Intercepts as Outcomes

```
melogit adapt malem agem EDUCM2 EGP45 EGP711 EGP21 EGP22 EGP2325
weurope || cntryid: , intmethod(mcaghermite) intpoints(1)
```

```
Mixed-effects logistic regression               Number of obs      =     17419
Group variable:         cntryid                 Number of groups   =        17

                                                Obs per group: min =       787
                                                               avg =    1024.6
                                                               max =      1664

Integration method:     laplace

                                                Wald chi2(9)       =    224.90
Log likelihood = -10858.081                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       adapt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       malem |   .0441588   .0343608     1.29   0.199     -.023187    .1115046
        agem |   .0030315   .0013478     2.25   0.024     .0003899    .0056731
      EDUCM2 |   -.039165   .0052317    -7.49   0.000     -.049419    -.028911
       EGP45 |   .2106622   .0791975     2.66   0.008      .055438    .3658864
      EGP711 |    .254381   .0527049     4.83   0.000     .1510813    .3576808
       EGP21 |   .0146583   .0788239     0.19   0.852    -.1398337    .1691504
       EGP22 |   .0564473   .0729826     0.77   0.439    -.086596     .1994905
     EGP2325 |   .2427728   .0525932     4.62   0.000      .139692    .3458535
     weurope |    1.13972   .2661713     4.28   0.000      .618034    1.661406
       _cons |  -.3659092   .2175315    -1.68   0.093    -.7922631    .0604446
-------------+----------------------------------------------------------------
cntryid      |
   var(_cons)|    .295026   .1029406                      .1488877     .584604
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =  1055.17 Prob>=chibar2 = 0.0000
```

```
melogit, or
```

```
Mixed-effects logistic regression               Number of obs      =     17419
Group variable:         cntryid                 Number of groups   =        17

                                                Obs per group: min =       787
                                                               avg =    1024.6
                                                               max =      1664

Integration method:     laplace

                                                Wald chi2(9)       =    224.90
Log likelihood = -10858.081                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       adapt | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       malem |   1.045148   .0359121     1.29   0.199     .9770797    1.117959
        agem |   1.003036   .0013519     2.25   0.024      1.00039    1.005689
      EDUCM2 |    .961592   .0050308    -7.49   0.000     .9517822     .971503
       EGP45 |   1.234495   .0977689     2.66   0.008     1.057003    1.441791
      EGP711 |   1.289663   .0679716     4.83   0.000     1.163091    1.430009
       EGP21 |   1.014766   .0799879     0.19   0.852     .8695028    1.184298
       EGP22 |   1.058071   .0772207     0.77   0.439     .9170475    1.220781
     EGP2325 |   1.274779   .0670447     4.62   0.000      1.14992    1.413196
     weurope |   3.125894   .8320233     4.28   0.000     1.855277    5.266713
       _cons |   .6935658   .1508724    -1.68   0.093     .4528189    1.062309
-------------+----------------------------------------------------------------
cntryid      |
   var(_cons)|    .295026   .1029406                      .1488877     .584604
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =  1055.17 Prob>=chibar2 = 0.0000
```

Suggested Readings

Hoffmann, John P. 2004. *Generalized Linear Model: An Applied Approach*. Boston: Pearson.

Pampel, Fred C. 2000. *Logistic Regression: A Primer*. Thousand Oaks, CA: Sage.

Rabe-Hesketh, Sophia and Anders Skrondal. 2012. *Multilevel and Longitudinal Modeling Using Stata, Volume II: Categorical Responses, Counts, and Survival (3rd Edition)*. College Station, TX: Stata Press.

Raudenbush, Stephen W. and Anthony S. Bryk. 2002. *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd Edition). Chapter 10: "Hierarchical Generalized Linear Models." Thousand Oaks, CA: Sage Publications.