Logistic Regression, Part III: Hypothesis Testing, Comparisons to OLS

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised January 17, 2022

This handout steals heavily from Linear probability, logit, and probit models, by John Aldrich and Forrest Nelson, paper # 45 in the Sage series on Quantitative Applications in the Social Sciences; and Applied Logistic Regression Analysis Second Edition by Scott Menard, paper # 106 in that series. WARNING: As Menard more or less points out, notation is wildly inconsistent across authors and programs when it comes to Logistic regression. I'm trying to more or less follow Menard, but you'll have to learn to adapt to whatever the author or statistical program happens to use.

OVERVIEW. In this handout, we'll examine hypothesis testing in logistic regression and make comparisons between logistic regression and OLS. A separate handout provides more detail about using Stata. The optional appendices to this handout also provide more details. Appendix A shows more logical analogs between logistic regression and OLS regression. Appendix B explains what the Log Likelihood is and how it is calculated. Appendix C elaborates further on calculating the Model Chi-Square.

Using the same data as before, here is part of the output we get in Stata when we do a logistic regression of Grade on Gpa, Tuce and Psi.

```
. use https://www3.nd.edu/~rwilliam/statafiles/logist.dta, clear
. logit  grade gpa tuce psi

Iteration 0:   log likelihood =  -20.59173
Iteration 1:   log likelihood = -13.496795
Iteration 2:   log likelihood = -12.929188
Iteration 3:   log likelihood = -12.889941
Iteration 4:   log likelihood = -12.889633
Iteration 5:   log likelihood = -12.889633

Logistic regression                                     Number of obs   =         32
                                                        LR chi2(3)      =      15.40
                                                        Prob > chi2     =     0.0015
Log likelihood = -12.889633                             Pseudo R2       =     0.3740
[Rest of output deleted]
```

GLOBAL TESTS OF PARAMETERS. In OLS regression, if we wanted to test the hypothesis that all β's = 0 versus the alternative that at least one did not, we used a global F test. In logistic regression, we use a *likelihood ratio chi-square test* instead. Stata calls this LR chi2. The value in this case is 15.40. This is computed by contrasting a model which has no independent variables (i.e. has the constant only) with a model that does. There are three degrees of freedom in this case because three coefficients (other than the constant) were estimated, e.g. one for each independent variable in the model. You can calculate this using the information from iteration 0 (the constant only model; we will call the log-likelihood from this model LL_0) and the final iteration (we'll call this log-likelihood LL_M):

$$\text{LR Chi-Square} = -2 * (LL0 - LLM) = -2 * (-20.592 + 12.889) = 15.40.$$

Another common notation refers to the Deviances of the models. $Dev_0 = -2 * LL_0$ and $Dev_M = -2 * LL_M$. Think of the deviances as reflecting the extent to which the model fails to perfectly

predict the observed outcomes. In this case $Dev_0 = -2 * -20.59173 = 41.18$ and $Dev_M = -2 * -12.89 = 25.78$. So,

$$\text{LR Chi-Square} = Dev_0 - Dev_M = 41.18 - 25.78 = 15.40.$$

If the null hypothesis is true, i.e. if all coefficients (other than the constant) equal 0 then the model chi-square statistic has a chi-square distribution with k degrees of freedom (k = number coefficients estimated other than the constant). In this case the model chi-square is highly significant suggesting that at least one variable has an effect that differs from 0.

Common notations for the model chi-square include Model χ^2, L^2, G_M.

INCREMENTAL TESTS / LIKELIHOOD RATIO CHI-SQUARE TESTS. There is also an analog to the incremental F test. Just like with OLS, we can compare constrained and unconstrained models, i.e. *nested models*. For example, we might be interested in contrasting a model with X1, X2, and X3, with a model that has the same three variables plus X4 & X5. We refer to the first model as the constrained model because, by not including X4 and X5, we in effect constrain their effects to equal 0. For example, X1, X2, and X3 might be demographic variables, and we might want to see whether attitudinal measures X4 and X5 tell us anything more than the demographic variables do.

In logistic regression we use an incremental chi-square square statistic instead of an incremental F statistic. (More commonly, you see phrases like chi-square contrasts.) The difference between the deviances of constrained and unconstrained models has a chi-square distribution with degrees of freedom equal to the number of constraints. The simplest formula is

$$L^2 = \text{Model } L^2_{\text{ Unconstrained}} - \text{Model } L^2_{\text{ Constrained}}, \text{ d.f.} = \text{number of constraints}$$

The notation L^2 is used to signify that this is a Likelihood Ratio Chi Square test (as opposed to, say, a Pearson Chi-Square test, which has less desirable properties). Again, notation is wildly inconsistent across authors. G^2 is another notation sometime used.

In Stata, we can get incremental and global LR chi-square tests easily by using the `estimates store` and `lrtest` command. In the following the `quietly` option suppresses a lot of the intermediate information, but don't use it if you want to see those results.

```
. quietly logit  grade gpa
. est store m1
. quietly logit  grade gpa tuce
. est store m2
. quietly logit  grade gpa tuce psi
. est store m3
. lrtest m1 m2

Likelihood-ratio test                                 LR chi2(1)  =      0.43
(Assumption: m1 nested in m2)                         Prob > chi2 =    0.5096

. lrtest m2 m3

Likelihood-ratio test                                 LR chi2(1)  =      6.20
(Assumption: m2 nested in m3)                         Prob > chi2 =    0.0127
```

TESTS OF INDIVIDUAL PARAMETERS. Testing whether any individual parameter equals zero proceeds pretty much the same way as in OLS regression. You can, if you want, do an incremental LR chi-square test. That, in fact, is the best way to do it, since the Wald test referred to next is biased under certain situations. For individual coefficients, Stata reports z values, which is b/s_b.

```
. logit  grade gpa tuce psi, nolog

Logistic regression                               Number of obs   =         32
                                                  LR chi2(3)      =      15.40
                                                  Prob > chi2     =     0.0015
Log likelihood = -12.889633                       Pseudo R2       =     0.3740

------------------------------------------------------------------------------
       grade |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         gpa |   2.826113   1.262941     2.24   0.025     .3507938    5.301432
        tuce |   .0951577   .1415542     0.67   0.501    -.1822835    .3725988
         psi |   2.378688   1.064564     2.23   0.025       .29218    4.465195
       _cons |  -13.02135   4.931325    -2.64   0.008    -22.68657    -3.35613
------------------------------------------------------------------------------
```

With Stata, you can also continue to use the `test` command. The `test` command does Wald tests, which aren't as good as LR tests but which may be adequate in large samples, e.g.

```
. * Test whether effects of gpa and tuce are both 0
. test gpa tuce

 ( 1)  gpa = 0
 ( 2)  tuce = 0

           chi2(  2) =    6.35
         Prob > chi2 =    0.0418

. * Test whether effects of gpa and psi are equal
. test gpa = psi

 ( 1)  gpa - psi = 0

           chi2(  1) =    0.11
         Prob > chi2 =    0.7437
```

R^2 ANALOGS. As Menard points out in Applied Logistic Regression Analysis, Second Edition, several people have tried to come up with the equivalent of an R^2 measure for logistic regression. No one of these measures seems to have achieved widespread acceptance yet. One of the simplest and most popular formulas is McFadden's Pseudo R^2:

$$\text{Pseudo } R^2 = \text{Model } L^2 / DEV_0 = 1 - DEV_M/DEV_0 = 1 - LL_M/LL_0$$

This statistic will equal zero if all coefficients are zero. It will come close to 1 if the model is very good. In the present case, for the model with gpa, psi and tuce included,

$$\text{Pseudo } R^2 = \text{Model } L^2 / \text{ DEV}_0 = 15.404/41.183 = .374$$

OTHER WAYS OF ASSESSING "GOODNESS OF FIT." There are other ways to assess whether or not the model fits the data. For example, there is the *classification table*. The command in Stata is `estat class` (you can also just use `lstat`)

```
. quietly logit  grade gpa tuce psi
. estat class

Logistic model for grade

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         8             3  |         11
     -     |         3            18  |         21
-----------+--------------------------+-----------
   Total   |        11            21  |         32

Classified + if predicted Pr(D) >= .5
True D defined as grade != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   72.73%
Specificity                     Pr( -|~D)   85.71%
Positive predictive value       Pr( D| +)   72.73%
Negative predictive value       Pr(~D| -)   85.71%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   14.29%
False - rate for true D         Pr( -| D)   27.27%
False + rate for classified +   Pr(~D| +)   27.27%
False - rate for classified -   Pr( D| -)   14.29%
-------------------------------------------------
Correctly classified                        81.25%
-------------------------------------------------
```

In the classification table, cases with probabilities ≥ .50 are predicted as having the event, other cases are predicted as not having the event. Ideally, you would like to see the two groups have very different estimated probabilities. In this case, of the 21 people who did not get A's, the model correctly predicted 18 would not but said that 3 would. Similarly, of the 11 who got A's, the model was right on 8 of them.

In this case, if you had no useful information about people, the smartest strategy would be to guess that nobody got an A – and you would be right for 21 of the 32 cases, or 65.625% of the time. (In other words, if you have to make a wild guess, guess that everyone will have the value of the category with the highest frequency on the dependent variable.) Using the logistic regression model, however, 81.25% of the cases, or 26 of the 32, are correctly classified. Thus, thanks to the model, 5 additional cases are classified correctly. If the classification table classifies more cases correctly than just guessing that every case falls into the category with the highest frequency, then the table has provided something of value.

The classification table has limits though. From the classification table, you can't tell how great the errors are. The 6 misclassified cases may have been within one or two percentage points of being classified correctly, or they may have been way off. For "rare" events, e.g. when only 10% of the sample has a value of 1 on the dependent variable, the table may not be at all useful, because every case can get classified as NOT experiencing the event. A 30% predicted probability for a case may be relatively high, but still not high enough to get the case classified as a 1. Menard goes on at some length about other possible classification/prediction strategies.

The handout on **Measures of Fit** discusses several other measures that may be useful at times. For example, the Adjusted Count R^2 may be a useful supplement to the Classification Table, because it helps to quantify exactly how much the classification of cases has been improved because of the model.

DIAGNOSTICS. It can also be useful to run various diagnostics. These help to indicate areas or cases for which the model is not working well. Menard lists several statistics for looking at residuals. Menard also briefly discusses some graphical techniques that can be useful. Also see Hamilton's Statistics with Stata for some ideas.

In Stata, you can again use the `predict` command to compute various outliers. As was the case with OLS, Stata tends to use different names than SPSS and does some computations differently. Cases 2 and 27 seem to be the most problematic.

```
. * Generate standardized residuals
. predict p
(option pr assumed; Pr(grade))
. predict rstandard, rstandard
. extremes rstandard p grade gpa tuce psi
```

obs:	rstandard	p	grade	gpa	tuce	psi
27.	-2.541286	.8520909	0	3.51	26	1
18.	-1.270176	.5898724	0	3.12	23	1
16.	-1.128117	.5291171	0	3.1	21	1
28.	-.817158	.3609899	0	3.53	26	0
24.	-.7397601	.3222395	0	3.57	23	0

19.	.8948758	.6354207	1	3.39	17	1
30.	1.060433	.569893	1	4	21	0
15.	1.222325	.481133	1	2.83	27	1
23.	2.154218	.1932112	1	3.26	25	0
2.	3.033444	.1110308	1	2.39	19	1

Appendix A (Optional): More Comparisons with OLS

There are many similarities between OLS and Logistic Regression, and some important differences. I'll try to highlight the most crucial points here.

OLS and its extensions	**Logistic Regression**
Estimated via least squares	Estimated via Maximum Likelihood.
Y is continuous, can take on any value	Y can only take on 2 values, typically 0 and 1
X's are continuous vars. Categorical variables are divided up into dummy variables	Same as OLS
X's are linearly related to Y; in the case of the LPM, X's are linearly related to P(Y=1)	X's are linearly related to log odds of event occurring. Log odds, in turn, are nonlinearly related to P(Y = 1).
Y's are statistically independent of each other, e.g., don't have serial correlation, don't include husbands and their wives as separate cases	Same as OLS
Robust standard errors can be used when error terms are not independent and identically distributed.	Same as OLS. Stata makes this easy (just add a `robust` parameter), SPSS does not.
There can be no perfect multicollinearity among the X's. High levels of multicollinearity can result in unstable sample estimates and large standard errors	Same as OLS. Techniques for detecting multicollinearity are also similar. In fact, as Menard points out, you could just run the corresponding OLS regression, and then look at the correlations of the IVs, the tolerances, variance inflation factors, etc. Or, use Stata's `collin` command.
Missing data can be dealt with via listwise deletion, pairwise deletion, mean substitution, multiple imputation	Pairwise deletion isn't an option. Can't do "mean substitution" on the DV. Otherwise, can use techniques similar to those that we've described for OLS.
Global F test is used to test whether any IV effects differ from 0. d.f. = K, N-K-1	Model chi-square statistic (also known as Model L^2 or G^2 or G_M) is used for same purpose. D.F. = number of IVs in the model = K.
Incremental F test is used to test hypotheses concerning whether subset of coefficients = 0. If you specify variables in blocks, the F change statistic will give you the info you need.	LR Chi-square statistic is used. $DEV_{Constrained} - DEV_{Unconstrained}$ Model L^2 $_{Unconstrained}$ – Model L^2 $_{Constrained}$

T test or incremental F test is used to test whether an individual coefficient = 0	Can use a LR chi square test (preferable) or Wald statistic (probably usually ok, but not always).
Incremental F tests or T tests can be used to test equalities of coefficients within a model, equalities across populations, interaction effects.	Same basic procedures, substituting LR chi square tests for F tests.
Wald tests (as produced by the `test` command in stata) will produce the same results as incremental F tests. A nice thing about Wald tests is that they only require the estimation of the unconstrained model.	Wald tests can be performed, but they will generally NOT produce exactly the same results as LR tests. LR tests (which require the estimation of constrained and unconstrained models) are preferable, although in practice results will often be similar.
Can have interaction effects. Centering can sometimes make main effects easier to interpret. If you center the continuous vars, then the main effect of an IV like race is equal to the difference in the predicted values for an "average" Black person and an "average" White person.	NOT quite the same as OLS. You can use interaction terms, but there are potential problems you should be aware of when interpreting results. See Allison (1999) or Williams (2009, 2010) for discussions. If you center, then the main effect of an IV like race is equal to the difference in the log odds for an "average" Black person and an "average" White person.
Can do transformations of the IVs and DV to deal with nonlinear relationships, e.g. X^2, ln(X), ln(Y).	Same as OLS for the IVs, but you of course can't do transformations of the dichotomous DV.
Can plot Y against X, examine residuals, plot X against residuals, to identify possible problems with the model	Similar to OLS. Can examine residuals.
Can do mindless, atheoretical stepwise regression	Similar to OLS
R^2 tells how much of total variance is "explained".	Numerous Pseudo R^2 stats have been proposed. If you use one, make clear which one it is.
Can look at standardized betas.	There is actually a reasonable case for using standardized coefficients in logistic regression. Long & Freese's `spost13` routines include the `listcoef` command, which can do various types of standardization.
Can do path analysis. Can decompose association. Can estimate recursive and nonrecursive models. Programs like LISREL and MPlus and Stata's sem command can deal with measurement error.	There is work going on in this area. Stata has the `gsem` and user-written `gllamm` commands. If you can afford it, probably the best program is `MPlus`.

OLS VERSUS LOGISTIC REGRESSION FOR HYPOTHESIS TESTING. There are a number of logical analogs between OLS and Logistic regression for hypothesis testing, i.e. the math is different but the functions served are similar.

OLS Regression	Logical Analog in Logistic Regression
Total Sums of Squares	$-2LL_0$, DEV_0, D_0
Error/ Residual Sums of Squares	$-2LL_M$, DEV_M, D_M
Regression/Explained Sums of Squares	Model Chi Square, L^2, G_M
Global F	Model Chi Square, L^2, G_M
Incremental F Test	Chi-Square Contrast/ Incremental chi-square contrast
Incremental F Test and Wald test of the same hypotheses give identical results	Chi-square contrast between models and a Wald test of the same hypotheses generally do NOT give exactly identical results.

Appendix B (Optional): Computing the log likelihood.

This is adapted from J. Scott Long's Regression Models for Categorical and Limited Dependent Variables.

Define p_i as the probability of observing whatever value of y was actually observed for a given observation, i.e.

$$p_i = \begin{cases} \Pr(y_i = 1 \mid x_i) & \text{if } y_i = 1 \text{ is observed} \\ 1 - \Pr(y_i = 1 \mid x_i) & \text{if } y_i = 0 \text{ is observed} \end{cases}$$

If the observations are independent, the likelihood equation is

$$L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} p_i$$

The likelihood tends to be an incredibly small number, and it is generally easier to work with the log likelihood. Ergo, taking logs, we obtain the log likelihood equation:

$$\ln L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \ln p_i$$

Before proceeding, let's see how this works in practice! Here is how you compute p_i and the log of p_i using Stata:

```
. use https://www3.nd.edu/~rwilliam/statafiles/logist.dta, clear
. quietly  logit grade gpa tuce psi
. * Compute probability that y = 1
. predict pi
(option p assumed; Pr(grade))
. * If y = 0, replace pi with probability y = 0
. replace pi = 1 - pi if grade == 0
(21 real changes made)
. * compute log of pi
. gen lnpi = ln(pi)

. list  grade pi lnpi, sep(8)

     +------------------------------+
     | grade         pi        lnpi |
     |------------------------------|
  1. |     0   .9386242   -.0633401 |
  2. |     1   .1110308   -2.197947 |
  3. |     0   .9755296   -.0247748 |
     |     --- Output deleted ---   |
 30. |     1    .569893   -.5623066 |
 31. |     1   .9453403   -.0562103 |
 32. |     1   .6935114   -.3659876 |
     +------------------------------+
```

So, this tells us that the predicted probability of the first case being 0 was .9386. The probability of the second case being a 1 was .111. The probability of the 3rd case being a 0 was .9755; and so on. The likelihood is therefore

$$L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{N} p_i = .9386 * .1110 * .9755 * ... * .6935 = .000002524$$

which is a really small number; indeed so small that your computer or calculator may have trouble calculating it correctly (and this is only 32 cases; imagine the difficulty if you have hundreds of thousands). Much easier to calculate is the log likelihood, which is

$$\ln L(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \ln p_i = -.0633 + -2.198 + ... + -.366 = -12.88963$$

Stata's `total` command makes this calculation easy for us:

```
. total lnpi

Total estimation                  Number of obs    =      32

------------------------------------------------------------------
             |      Total   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
        lnpi |  -12.88963   3.127734    -19.26869   -6.510578
------------------------------------------------------------------
```

If we do the same thing with the constant-only model the value is -20.59173 (which is what Stata reported as the LL for iteration 0). As reported in the main handout, LL_0 and LL_M can be used to compute the Model Chi-square as well as other statistics of interest.

Note: The maximum likelihood estimates are those values of the parameters that make the observed data most likely. That is, the maximum likelihood estimates will be those values which produce the largest value for the likelihood equation (i.e. get it as close to 1 as possible; which is equivalent to getting the log likelihood equation as close to 0 as possible).

Appendix C (Optional): Calculating the Model Chi-Square

The probability of the observed results given the parameter estimates is known as the *likelihood*. Since the likelihood is a small number less than 1, it is customary to use -2 times the log of the likelihood. -2LL is a measure of how well the estimated model fits the likelihood. A good model is one that results in a high likelihood of the observed results. This translates to a small number for -2LL (If a model fits perfectly, the likelihood is 1, and -2 times the log likelihood is 0).

-2LL is also called the Deviance, DEV, or simply D. Subscripts are often used to denote which model this particular deviance applies to. The smaller the deviance is, the better the model fits the data.

The "initial log likelihood function" is for a model in which only the constant is included. This is used as the baseline against which models with IVs are assessed. Stata reports LL_0, -20.59173, which is the log likelihood for iteration 0. $-2LL_0 = -2* -20.59173 = 41.18$.

$-2LL_0$, DEV_0, or simply D_0 are alternative ways of referring to the deviance for a model which has only the intercept. This is analogous to the Total Sums of Squares, SST, in OLS Regression.

When GPA, PSI, and TUCE are in the model, $-2LL_M = -2 * -12.889633 = 25.78$. We can refer to this as DEV_M or simply D_M.

The -2LL for a model, or DEV_M, indicates the extent to which the model fails to perfectly predict the values of the DV, i.e. it tells how much improvement is needed before the predictors provide the best possible prediction of the dependent variable. DEV_M is analogous to the Error Sums of Squares, SSE, in OLS regression.

The addition of these 3 parameters reduces -2LL by 15.40, i.e.
$DEV_0 - DEV_M = 41.183 – 25.779 = 15.40$. This is reflected in the *Model Chi-square,* which Stata labels as LR chi2.

The Model Chi-Square, also called Model L^2 or G_M, is analogous to the Regression (explained) Sums of Squares, SSR, in OLS regression. It is also the direct counterpart to the Global F Test in regression analysis. A significant value tells you that one or more betas differ from zero, but it doesn't tell you which ones.

$$G_M = L^2 = DEV_0 - DEV_M$$

The significance level for the model chi-square indicates that this is a very large drop in chi-square, ergo we reject the null hypothesis. The effect of at least one of the IVs likely differs from zero.

You can think of the Deviance as telling you how bad the model still is, while the Model L^2 , aka G_M tells you how good it is.