Scalar Measures of Fit: Pseudo R^2 and Information Measures (AIC & BIC)

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/

Last revised January 12, 2020

First we present the results for an OLS regression and a similar logistic regression. incbinary is a dichotomized version of income where the higher half of the cases are coded 1 and the bottom half are coded 0. The rest of the handout refers to these results often.

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-reg, clear
. reg income educ

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  963.54
       Model |  26490.0257     1  26490.0257            Prob > F      =  0.0000
    Residual |  13691.2236   498  27.4924168            R-squared     =  0.6593
-------------+------------------------------           Adj R-squared =  0.6586
       Total |  40181.2493   499  80.5235456            Root MSE      =  5.2433

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.830332   .0589651    31.04   0.000     1.714481    1.946183
       _cons |   3.702833   .8106362     4.57   0.000     2.110145    5.295522
------------------------------------------------------------------------------

. fitstat

                           |     regress
---------------------------+-------------
Log-likelihood             |
                    Model  |   -1536.945
          Intercept-only   |   -1806.106
---------------------------+-------------
Chi-square                 |
       Deviance (df=498)   |    3073.890
---------------------------+-------------
R2                         |
                        R2 |       0.659
                Adjusted R2 |       0.659
                   McFadden |       0.149
       McFadden (adjusted)  |       0.148
              Cox-Snell/ML  |       0.659
    Cragg-Uhler/Nagelkerke  |       0.660
---------------------------+-------------
IC                         |
                       AIC |    3077.890
          AIC divided by N |       6.156
                BIC (df=2) |    3086.319

. logit incbinary educ

Iteration 0:   log likelihood = -346.57359
[Other 3 iterations deleted]

Logistic regression                               Number of obs   =        500
                                                  LR chi2(1)      =      48.17
                                                  Prob > chi2     =     0.0000
Log likelihood = -322.48937                       Pseudo R2       =     0.0695

------------------------------------------------------------------------------
   incbinary |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1702216   .0266265     6.39   0.000     .1180347    .2224086
       _cons |  -2.245047   .3645915    -6.16   0.000    -2.959633    -1.53046
------------------------------------------------------------------------------
```

```
. fitstat

                           |       logit
---------------------------+-------------
Log-likelihood             |
                     Model |    -322.489
           Intercept-only  |    -346.574
---------------------------+-------------
Chi-square                 |
       Deviance (df=498)   |     644.979
                LR (df=1)  |      48.168
                  p-value  |       0.000
---------------------------+-------------
R2                         |
                McFadden   |       0.069
     McFadden (adjusted)   |       0.064
      McKelvey & Zavoina   |       0.122
            Cox-Snell/ML   |       0.092
  Cragg-Uhler/Nagelkerke   |       0.122
                   Efron   |       0.087
                Tjur's D   |       0.089
                   Count   |       0.660
        Count (adjusted)   |       0.320
---------------------------+-------------
IC                         |
                     AIC   |     648.979
        AIC divided by N   |       1.298
              BIC (df=2)   |     657.408
---------------------------+-------------
Variance of                |
                       e   |       3.290
                  y-star   |       3.749
```

Translating `fitstat` into our earlier notation, Log-Likelihood Intercept Only is LL_0, Log-Likelihood Model is LL_M, Deviance (df=498) is $DEV_M = -2*LL_M$, and LR (df=1) is L^2 or else G_M.

R² Analogs. Several Pseudo R^2 measures are logical analogs to OLS R^2 measures. McFadden's R^2 is perhaps the most popular Pseudo R^2 of them all, and it is the one that Stata is reporting when it says Pseudo R2. However, `fitstat` also reports several over pseudo R^2 statistics. The formulas and rationale for each of these is presented in **Appendix A**. Personally, I just use McFadden all the time (Tjur's R^2 is also growing in popularity, but it only works with binary dependent variable, not ordinal), but you should be clear on what statistic you are using and you should also be aware of the other statistics in case you encounter them.

Information Measures. A different approach to assessing the fit of a model and for comparing competing models is based on measures of information. As the multiplicity of Pseudo R^2 statistics suggests, there is considerable controversy as to which (if any) of these measures should be used. Further, the use of chi-square statistics as goodness of fit measures has been criticized.

- When sample sizes are large, it is much easier to accept (or at least harder to reject) more complex models because the chi-square test statistics are designed to detect any departure between a model and observed data. That is, adding more terms to a model will always improve the fit, but with a large sample it becomes harder to distinguish a "real" improvement in fit from a substantively trivial one.
- Likelihood-ratio tests therefore often lead to the rejection of acceptable models, and models become less parsimonious than they need to be.

Therefore, as Long, Raftery and others note, information measures – in particular BIC and AIC – have become increasingly popular. Some key features of these measures:

- BIC and AIC statistics are appropriate for many types of statistical methods, e.g. `regress`; they aren't just limited to logistic regression.
- The basic idea is to compare the relative plausibility of two models rather than to find the absolute deviation of observed data from a particular model.
- Unlike many Pseudo R^2 measures, the information measures have penalties for including variables that do not significantly improve fit. Particularly with large samples, the information measures can lead to more parsimonious but adequate models.
- Another strength is that you can compare the fits of different models, even when the models are not nested. This is particularly useful when you have competing theories that are very different. For example, some theories of crime say that criminal behavior is deviant and linked to the offender's psychological, social and family circumstances. Other theories say that criminal activity is a rational choice determined by its costs and benefits relative to other (legitimate) opportunities.
- There are different formulas for these measures. It really doesn't matter which you use, so long as you are consistent. The smaller the value of the statistic (or the more negative the value is) the better the fit of the model.

We will primarily focus on the BIC statistic. The AIC (Akaike's Information Criterion) is discussed in **Appendix B**.

BIC. The Bayesian Information Criterion (BIC) assesses the overall fit of a model and allows the comparison of both nested and non-nested models. It is based on a Bayesian comparison of models. Suppose you have two models. Under the assumption that you have no prior preference for one model over the other, BIC identifies the model that is more likely to have generated the observed data.

The formula for the BIC statistic reported by Stata (there are other formulas; see Appendix A) is

$$BIC_{Stata} = DEV_M + \ln(N) * P$$

where P is the number of parameters estimated (including the constant).

For the original OLS example above,

$$BIC_{Stata} = DEV_M + \ln(N) * P = 3073.89 + \ln(500) * 2 = 3073.89 + 6.215 * 2 = 3086.319$$

For the original logistic regression example,

$$BIC_{Stata} = DEV_M + \ln(N) * P = 644.979 + \ln(500) * 2 = 644.979 + 6.215 * 2 = 657.408$$

The BIC (and also AIC) statistics reported by Stata use formulas that are simpler and perhaps easier to understand and interpret than are other formulas, so I can see why Stata uses them. I also like the fact that the Stata versions give positive values rather than negative values. Appendix C discusses these. Any of the BIC statistics can be used to compare models, regardless

of whether they are nested or not. Further, it really doesn't matter which one you use, since $BIC_1 - BIC_2 = BIC'_1 - BIC'_2 = BIC_{Stata1} - BIC_{Stata2}$ (where the subscripts refer to the two models you are comparing). Just be consistent with whichever one you use.

The model with the smaller BIC or BIC' or BIC_{Stata} is preferred, i.e. if $BIC_1 - BIC_2 < 0$, model 1 is preferred. If $BIC_1 - BIC_2 > 0$, the second model is prefered. Why? If you look at the formula for BIC_{Stata}, ln(N) * P increases as you add more variables, while DEV_M goes down. Therefore, in order for additional parameters to be worth adding to the model, they must produce at least enough of a decrease in DEV_M to offset the increase in ln(N) * P.

How much one model is preferred over the other depends on the magnitude of the difference. Raftery proposed the following guidelines:

Absolute difference	Evidence
0-2	Weak
2-6	Positive
6-10	Strong
>10	Very Strong

Nested models. We'll expand on our previous logistic regression example to illustrate the use of BIC and AIC comparisons (and also show how `fitstat` can make things a little easier when doing this).

```
. quietly logit incbinary educ
. quietly fitstat, save
. quietly logit incbinary educ jobexp i.black
. fitstat, diff
```

	Current	Saved	Difference
Log-likelihood			
Model	-242.471	-322.489	80.019
Intercept-only	-346.574	-346.574	0.000
Chi-square			
D (df=496/498/-2)	484.941	644.979	-160.038
LR (df=3/1/2)	***208.206***	***48.168***	***160.038***
p-value	0.000	0.000	0.000
R2			
McFadden	0.300	0.069	0.231
McFadden (adjusted)	0.289	0.064	0.225
McKelvey & Zavoina	0.523	0.122	0.400
Cox-Snell/ML	0.341	0.092	0.249
Cragg-Uhler/Nagelkerke	0.454	0.122	0.332
Efron	0.334	0.087	0.248
Tjur's D	0.345	0.089	0.256
Count	0.840	0.660	0.180
Count (adjusted)	0.680	0.320	0.360

```
IC                      |
                   AIC |     492.941      648.979     -156.038
         AIC divided by N |       0.986        1.298       -0.312
           BIC (df=4/2/2) |     509.799      657.408     -147.609
-------------------------+------------------------------------------
Variance of              |
                       e |       3.290        3.290        0.000
                  y-star |       6.891        3.749        3.142

Note: Likelihood-ratio test assumes saved model nested in current model.

Difference of  147.609 in BIC provides very strong support for current model.
```

As we see, there is very strong evidence for adding jobexp and black to the model. The AIC is smaller when you add the 2 variables 492.941 versus 648.979). The difference in the BICs is much bigger than 10 (509.79 versus 647.408, a difference of -147.609). And, of course, the LR chi-square contrast between the two is very large, 160.038 with 2 d.f. All the various pseudo R^2 measures go up (of course, most have to when you add variables, but McFadden's Adj R^2 and the Adj Count R^2 go up too).

`fitstat` is nice because it explicitly tells you which model is better supported (and how strongly) and computes all these differences between model statistics, but it isn't essential. The `lrtest` command can give you the same information:

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-reg, clear
. quietly logit incbinary educ
. est store constrained
. quietly logit incbinary educ jobexp i.black
. est store unconstrained
. lrtest constrained unconstrained, stats

Likelihood-ratio test                                 LR chi2(2)  =    160.04
(Assumption: constrained nested in unconstrained)     Prob > chi2 =    0.0000

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
 constrained |    500   -346.5736   -322.4894      2     648.9787     657.408
unconstrai~d |    500   -346.5736   -242.4705      4      492.941    509.7994
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

Also useful is the `estat ic` (information criteria) post-estimation command.

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-reg, clear
. quietly logit incbinary educ jobexp i.black
. estat ic

Akaike's information criterion and Bayesian information criterion

-----------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
           . |        500 -346.5736  -242.4705       4     492.941   509.7994
-----------------------------------------------------------------------------
               Note: N=Obs used in calculating BIC; see [R] BIC note.
```

Another Example. Lets go back to our earlier example with grades. First we enter gpa and psi and then we enter tuce:

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-logit.dta, clear
. quietly logit grade gpa i.psi
. quietly fitstat, save
. quietly logit grade gpa i.psi tuce
. fitstat, diff
```

	Current	Saved	Difference
Log-likelihood			
Model	-12.890	-13.127	0.237
Intercept-only	-20.592	-20.592	0.000
Chi-square			
D (df=28/29/-1)	25.779	26.253	-0.474
LR (df=3/2/1)	***15.404***	***14.930***	***0.474***
p-value	0.002	0.001	0.491
R2			
McFadden	0.374	0.363	0.012
McFadden (adjusted)	***0.180***	***0.217***	***-0.037***
McKelvey & Zavoina	0.544	0.520	0.024
Cox-Snell/ML	0.382	0.373	0.009
Cragg-Uhler/Nagelkerke	0.528	0.515	0.013
Efron	0.426	0.407	0.019
Tjur's D	0.429	0.415	0.014
Count	0.813	0.813	0.000
Count (adjusted)	0.455	0.455	0.000
IC			
AIC	***33.779***	***32.253***	***1.526***
AIC divided by N	1.056	1.008	0.048
BIC (df=4/3/1)	***39.642***	***36.650***	***2.992***
Variance of			
e	3.290	3.290	0.000
y-star	7.210	6.856	0.354

```
Note: Likelihood-ratio test assumes saved model nested in current model.

Difference of    2.992 in BIC provides positive support for saved model.
```

Most of the measures which can go down when variables are added (AIC, BIC, McFadden's Adj R^2) do go down, while the other Pseudo R^2 measures go up very little. The LR chi-square contrast between the models is not significant. The more parsimonious model that does not include TUCE is preferred.

Non-Nested Models. This isn't the best example…but suppose we had two theories, one of which said that income was a function of characteristics determined at birth (e.g. race) and another theory that said income was a function of achieved characteristics, e.g. education and job experience. If we viewed these as dueling theories, we could do something like this:

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-reg, clear
. quietly logit incbinary i.black
. quietly fitstat, save
```

```
. quietly logit incbinary educ jobexp
. fitstat, diff
```

	Current	Saved	Difference
Log-likelihood			
Model	-256.251	-301.713	45.462
Intercept-only	-346.574	-346.574	0.000
Chi-square			
D (df=497/498/-1)	512.503	603.426	-90.923
LR (df=2/1/1)	180.644	89.721	90.923
p-value	0.000	0.000	0.000
R2			
McFadden	0.261	0.129	0.131
McFadden (adjusted)	0.252	0.124	0.128
McKelvey & Zavoina	0.442	0.248	0.194
Cox-Snell/ML	0.303	0.164	0.139
Cragg-Uhler/Nagelkerke	0.404	0.219	0.185
Efron	0.285	0.160	0.125
Tjur's D	0.301	0.160	0.141
Count	0.660	0.660	0.000
Count (adjusted)	0.320	0.320	0.000
IC			
AIC	***518.503***	***607.426***	***-88.923***
AIC divided by N	1.037	1.215	-0.178
BIC (df=3/2/1)	***531.147***	***615.855***	***-84.709***
Variance of			
e	3.290	3.290	0.000
y-star	5.896	4.376	1.520

Note: Likelihood-ratio test assumes saved model nested in current model.

Difference of 84.709 in BIC provides very strong support for current model.

Note that the model with education and job experience fits the data much better than the model with race only. It has smaller AIC and BIC values. Such comparisons can be very useful when the theories are very different and you can't just compare them via a series of nested models. But, the same cases do need to be analyzed throughout.

Conclusion. When presenting results, I think it is generally a good idea to present the McFadden's Pseudo R^2 statistic; the model chi-square and degrees of freedom; and personally I like to see the BIC and/or AIC statistics included as well.

Sources/Additional Reading. Long and Freese (2003, 2006, 2014), Long (1997) and Powers & Xie (2000, 2008) are major sources for these notes (the worked examples are mine but a lot of the text comes directly from their books.) Also, the 1995 volume of Sociological Methodology contains several chapters on Bayesian model selection that are used in this handout. The simple models presented here do not begin to do justice to the BIC measures. Adrian Raftery's "Bayesian Model Selection in Social Research," from Sociological Methodology, V. 25 (1995), pp. 111-163, does a superb job of discussing BIC. He points out the problems with traditional methods of hypothesis testing and how the use of BIC can help to address them. The first few pages and the last few pages cover the highlights, but the entire article is highly recommended. The volume also included some interesting responses to Raftery; Robert Hauser in particular praises the use of BIC in model selection.

Appendix A: Pseudo R^2 Measures

NOTE: Paul Allison has a good discussion of the merits of different measures at

http://www.statisticalhorizons.com/r2logistic

As noted earlier, McFadden's Pseudo R^2 is the measure reported by Stata. But, there are several others, many of which are based on some so-called logical analog with the R^2 reported in OLS. We will use the results from the OLS and Logistic models reported on p. 1 of this handout. A key thing to realize is that the different formulas for OLS R^2 all give the same results, but the logical analogs for Pseudo R^2 do not.

OLS Regression	Logistic Regression
1A. Percentage of Explained Variation: $R^2 = \frac{SSR}{SST} = \frac{26490}{40181} = .6593$ $= 1 - \frac{SSE}{SST} = 1 - \frac{13691}{40181} = .6593$ Note: Different authors use different notation. I use SSR = Regression Sum of Squares SSE = Error Sum of Squares SST = Total Sum of Squares	1A. McFadden's R^2 (perhaps the most popular Pseudo R^2 of them all, and the one that Stata is reporting when it says Pseudo R2): $R^2 = \frac{G_M}{Dev_0} = \frac{48.168}{-2*-346.574} = .069$ $= 1 - \frac{LL_M}{LL_0} = 1 - \frac{-322.489}{-346.574} = .069$ $= 1 - \frac{Dev_M}{Dev_0} = 1 - \frac{644.979}{693.148} = .069$
	1B. Efron's R^2 (another logical analog to Percentage of Explained Variation): $R^2 = 1 - \frac{\sum(y_i - \hat{\pi}_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{114.17}{125} = .087$ Note: You know the denominator will be 125 (because the mean of incbinary is .5) but I had to do some additional runs to get the numerator, i.e. I computed the predicted probabilities ($\hat{\pi}_i$) and then computed the squared residuals.
1C. Adjusted R^2: $R^2 = 1 - \frac{SSE/(N-K)}{SST/(N-1)} = 1 - \frac{MSE}{MST}$ $= 1 - \frac{27.4924168}{80.5235456} = .6586$ Where K = the # of parameters (including the intercept). This will not necessarily go up as more variables are added.	1C. McFadden's Adjusted R^2: $R^2 = 1 - \frac{LL_M - K}{LL_0} = 1 - \frac{-322.489 - 2}{-346.574} = .064$ Where K = the # of parameters (including the intercept). Like Adjusted R^2 in OLS, this will not necessarily go up as more variables are added.

2. Ratio of Var(Y) and Var($\hat{Y}$): $R^2 = \frac{Var(\hat{Y})}{Var(Y)} = 1 - \frac{Var(\varepsilon_Y)}{Var(Y)} = 1 - \frac{27.437}{80.524} = .6593$	2. McKelvey & Zavoina's R^2: $R^2 = 1 - \frac{Var(\varepsilon_{Y^*})}{Var(Y^*)} = 1 - \frac{3.29}{3.749} = .122$
3A. Transformation of the Likelihood Ratio: $R^2 = 1 - \left[\frac{L_0}{L_M}\right]^{2/N} = 1 - \left[e^{LL_0 - LL_M}\right]^{2/N}$ $= 1 - \left[e^{-1806.106+1536.945}\right]^{2/500} = 1 - \left[e^{-269.161}\right]^{2/500}$ $= 1 - e^{-1.076644} = 1 - .3407 = .6593$ I had to do the above simplifications to get my calculator to handle it without giving me errors! A computer can handle it better. Or, more simply, $R^2 = 1 - \exp(-G_M / N) = 1 - \exp(-538.323/500)$ $= 1 - \exp(-1.076646) = 1 - .3407 = .6593$	3a. Maximum Likelihood R^2 (SPSS calls this the Cox-Snell R^2 and it is also called the geometric mean squared improvement per observation; You'll also see Maddala's name associated with this): $R^2 = 1 - \exp(-G_M / N) = 1 - \exp(-48.168/500)$ $= 1 - \exp(-.096336) = 1 - .908 = .092$ NOTE: This isn't just a logical analog to OLS; it is the exact same formula!
	3B. Craig and Uhler's R^2 (which SPSS calls Nagelkerke R^2! But either way it is an adjustment for the ML R^2/ Cox-Snell R^2, which makes it possible for the R^2 to have a maximum value of 1; otherwise it maxes out at the denominator shown below): $R^2 = \frac{1 - \left[\frac{L_0}{L_M}\right]^{2/N}}{1 - L_0^{2/N}} = \frac{1 - \left[\frac{\exp(-346.574)}{\exp(-322.489)}\right]^{2/500}}{1 - \exp(-346.574)^{2/500}}$ $= \frac{1 - [\exp(-24.085)]^{2/500}}{1 - \exp(-346.574)^{2/500}} = \frac{.0918448}{.75} = .122$

What should you use? Allison (http://www.statisticalhorizons.com/r2logistic) says "For years, I've been recommending the Cox and Snell R^2 over the McFadden R^2, but I've recently concluded that that was a mistake. I now believe that McFadden's R^2 is a better choice. However, I've also learned about another R^2 that has good properties, a lot of intuitive appeal, and is easily calculated. At the moment, I like it better than the McFadden R^2. But I'm not going to make a definite recommendation until I get more experience with it." That R^2 is discussed next.

Tjur's Coefficient of Discrimination. `fitstat` also produces a fairly new measure of pseudo R^2, Tjur's Coefficient of Discrimination, D. (Others call it Tjur's R^2). SAS (http://support.sas.com/kb/39/109.html) describes it as follows:

D is the difference in the average of the event probabilities between the groups of observations with observed events and nonevents.

These are the properties of D according to Tjur (2009):

Like R^2, D ranges from 0 to 1.

$D \geq 0$. D = 0 if and only if all estimated probabilities are equal — the model has no discriminatory power.

$D \leq 1$. D = 1 if and only if the observed and estimated probabilities are equal for all observations — the model discriminates perfectly.

D will not always increase when predictors are added to the model.

Allison (http://www.statisticalhorizons.com/r2logistic) says several good things about Tjur's measure (which he calls Tjur's R^2):

It has a lot of intuitive appeal, its upper bound is 1.0, and it's closely related to R^2 definitions for linear models. It's also easy to calculate.

The definition is very simple. For each of the two categories of the dependent variable, calculate the mean of the predicted probabilities of an event. Then, take the [absolute value of the] difference between those two means. That's it!

The motivation should be clear. If a model makes good predictions, the cases with events should have high predicted values and the cases without events should have low predicted values.

`fitstat` gave a value of .089 for the statistic. Here is how you can estimate it in our current example:

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-reg, clear
. quietly logit incbinary educ
. predict yhat
(option pr assumed; Pr(incbinary))

. ttest yhat, by(incbinary)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     250    .4554406     .007504    .1186482    .4406613      .47022
       1 |     250    .5445594    .0105659     .167062    .5237494    .5653694
---------+--------------------------------------------------------------------
combined |     500          .5    .0067736    .1514629    .4866917    .5133083
---------+--------------------------------------------------------------------
    diff |           -.0891188    .0129595               -.1145808   -.0636568
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -6.8767
Ho: diff = 0                                     degrees of freedom =      498

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

As we would expect/hope, those who had zeros on incbinary had a lower average predicted probability (.4554) than did those who had ones on incbinary (.5445). That isn't a huge difference though, so Tjur's D is only .089. In this case (but not always), if you add a few more variables to the model, D gets bigger:

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-reg, clear
. quietly logit incbinary educ jobexp i.black
. predict yhat
(option pr assumed; Pr(incbinary))
. ttest yhat, by(incbinary)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     250    .3276272     .019042    .3010807    .2901233    .3651312
       1 |     250    .6723728    .0105909     .167457    .6515136     .693232
---------+--------------------------------------------------------------------
combined |     500          .5    .0133416    .2983265    .4737874    .5262126
---------+--------------------------------------------------------------------
    diff |           -.3447456    .0217891               -.3875555   -.3019356
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -15.8219
Ho: diff = 0                                     degrees of freedom =      498

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Note that, if everybody in group 0 had a predicted probability of 0, and everybody in group 1 had a predicted probability of 1, the difference in the mean predicted probabilities would be 1 and Tjur's R^2 would equal 1. Conversely, if the average predicted probability was the same in both groups, Tjur's R^2 would equal 0. So, the better the model does at predicting the outcomes, the higher Tjur's R^2 will be.

Allison also notes some possible limitations of Tjur's R^2:

> One possible objection to the Tjur R^2 is that, unlike Cox-Snell and McFadden, it's not based on the quantity being maximized, namely, the likelihood function.* As a result, it's possible that adding a variable to the model could reduce the Tjur R^2. But Kvalseth (1985) argued that it's actually preferable that R^2 not be based on a particular estimation method. In that way, it can legitimately be used to compare predictive power for models that generate their predictions using very different methods. For example, one might want to compare predictions based on logistic regression with those based on a classification tree method.
>
> *Another potential complaint is that the Tjur R^2 cannot be easily generalized to ordinal or nominal logistic regression*. For McFadden and Cox-Snell, the generalization is straightforward.

Pseudo R^2's Using Observed Versus Predicted Values: Count R^2 and Adjusted Count R^2

. These are two other Pseudo R^2 measures and are not based on an analog with OLS. To understand them, let's first present a little additional information from our previous logistic regression (you can also give the command `estat clas`):

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-reg, clear
. quietly logit incbinary educ
. lstat

Logistic model for incbinary

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       170            90  |        260
     -     |        80           160  |        240
-----------+--------------------------+-----------
   Total   |       250           250  |        500
```

In the table, the diagonal cases are the ones that were correctly classified, i.e. (170 + 160) = 330 were correctly classified, i.e. 66%. Thus, the Count R^2 is

$$\text{Count } R^2 = \frac{\text{Number Correctly Classified}}{\text{Total Number of Cases}} = \frac{330}{500} = .66$$

Note, however, that that number may not mean a whole lot. You can always correctly predict at least 50% of the cases by choosing the outcome category with the largest percentage of observed cases, e.g. if 60% of the cases are successes you will be correct 60% of the time if you pick all the cases to be a success (better yet, if 90% of the cases are failures, you will be right 90% of the time by picking everyone to fail!). The Adjusted Count R^2 adjusts for this by subtracting the largest row marginal from both the denominator and the numerator. In this case, there are an even number of successes and failures, 250; but if instead there had been 300 successes and 200 failures, you would subtract 300 from numerator and denominator. If there had been 100 successes and 400 failures you would subtract 400 from both the numerator and denominator.

$$\text{Adj Count } R^2 = \frac{\text{\# Correctly Classified - Max(Observed \# Successes, Observed \# Failures)}}{\text{Total \# of Cases - Max(Observed \# Successes, Observed \# Failures)}} = \frac{330 - 250}{500 - 250} = .32$$

Appendix B: Akaike's Information Criterion (AIC)

Akaike's Information Criterion (AIC) is defined as

$$AIC = Dev_M + 2P$$

Where P = the number of parameters in the model (including the intercept). In the OLS example above,

$$AIC = Dev_M + 2P = 3073.89 + 4 = 3077.89$$

and in the logistic regression example,

$$AIC = Dev_M + 2P = 644.979 + 4 = 648.979$$

The smaller the deviance, the better the model fits. As you add more parameters, the fit will improve; adding 2P to the deviance is a penalty for increasing the number of parameters. Since the number of observations affects the deviance, we divide by N to obtain the per-observation contribution to the adjusted deviance. All else being equal, smaller values suggest a better fitting model.

AIC is often used to compare models across different samples or to compare non-nested models that cannot be compared with the LR test. All else being equal, the model with the smaller AIC is considered the better fitting model.

Some authors prefer to report AIC/N. You can then compare AIC statistics across samples even when the sample sizes differ. `fitstat` reports these alternative AICs as AIC and AIC/N. AIC is also the AIC statistic reported by Stata. The formula can be written simply as

$$AIC / N = \frac{Dev_M + 2P}{500} = \frac{644.979 + 4}{500} = 1.298$$

Appendix C: Alternative Formulas for BIC

There are a couple of different formulas for BIC. Again, the formula Stata uses is

$$BIC_{Stata} = DEV_M + \ln(N) * P$$

where P is the number of parameters estimated (including the constant).

I like this formula because it produces positive values which makes it easier to interpret. But, it doesn't matter that much which formula you use so long as you are clear and consistent. When comparing two models, the differences in their BIC values will be the same no matter which formula you use. If you are having trouble replicating previous work it may be because a different formula was used. Note too that even within Stata different formulas sometimes get used, e.g. the user-written `bicdrop1` uses different formulas (i.e. the ones below) than BIC_{Stata}.

For example, one popular alternative (which `fitstat` call BIC (deviance) because it uses the deviance in the calculation; but others will just call it BIC) is

$$BIC_{Dev} = Dev_M - df_M \ln N$$

where the df = N - # of parameters (including the intercept). In the OLS example above,

$$BIC_{Dev} = Dev_M - df_M \ln N = 3073.890 - 498 * \ln 500 = 3073.89 - 498 * 6.2146 = -20.985$$

and for the logistic regression

$$BIC_{Dev} = Dev_M - df_M \ln N = 644.979 - 498 * \ln 500 = 644.979 - 498 * 6.2146 = -2449.896$$

If BIC_{Dev} is positive, the saturated model (i.e. the model with one parameter for every case; the BIC_{Dev} for a saturated model will equal 0) is preferred (i.e. the more complex model is better). When BIC_{Dev} is negative, the current model is preferred. The more negative the BIC_{Dev}, the better the fit.

Another version of BIC is called BIC' and is based on the Model Chi-Square, with the degrees of freedom equal to the number of regressors (intercept not included).

$$BIC' = -G_M + df'_M \ln N$$

For the OLS example above,

$$BIC' = -G_M + df'_M \ln N = -538.323 + 1 * \ln 500 = -538.323 + 6.2146 = -532.108$$

For the logistic regression example,

$$BIC' = -G_M + df'_M \ln N = -48.168 + 1 * \ln 500 = -48.168 + 6.2146 = -41.953$$

If BIC' is positive, the null model is preferred (i.e. the model with only the constant; it will have a BIC' value of 0). A positive BIC' implies that your model has too many variables in it. If BIC' is negative, then the current model is preferred over the null model (and the more negative BIC' is, the better). Basically, BIC' tests whether the model fits the data sufficiently well enough to justify the number of parameters that are used.

You can get all three of the different BIC measures via the following:

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-reg, clear
. quietly reg income educ
. fitstat, ic

                           |     regress
---------------------------+-------------
AIC                        |
                       AIC |    3077.890
           (divided by N)  |       6.156
---------------------------+-------------
BIC                        |
               BIC (df=2)  |    3086.319
 BIC (based on deviance)   |     -20.985
    BIC' (based on LRX2)   |    -532.108

. quietly logit incbinary educ
. fitstat, ic


                           |       logit
---------------------------+-------------
AIC                        |
                       AIC |     648.979
           (divided by N)  |       1.298
---------------------------+-------------
BIC                        |
               BIC (df=2)  |     657.408
 BIC (based on deviance)   |   -2449.896
    BIC' (based on LRX2)   |     -41.954
```

The first BIC statistic is the BIC reported by Stata while the other two BICs use the formulas presented in this appendix.

To confirm that it doesn't matter which BIC formula you use so long as you always use the same one,

```
. quietly logit incbinary i.black
. quietly fitstat, save ic
. quietly logit incbinary educ jobexp
```

```
. fitstat, diff ic

                           |     Current        Saved   Difference
---------------------------+-----------------------------------------
AIC                        |
                       AIC |     518.503      607.426      -88.923
           (divided by N)  |       1.037        1.215       -0.178
---------------------------+-----------------------------------------
BIC                        |
          BIC (df=3/2/1)   |     531.147      615.855      -84.709
 BIC (based on deviance)   |   -2576.157    -2491.449      -84.709
    BIC' (based on LRX2)   |    -168.215      -83.507      -84.709

Difference of   84.709 in BIC provides very strong support for current model.
```