

## Scalar Measures of Fit: Pseudo R<sup>2</sup> and Information Measures (AIC & BIC)

Long and Freese (2003, 2006), Long (1997) and Powers & Xie (2000) 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 (see especially Raftery's chapter) that are used in this handout.

OLS R<sup>2</sup> Analogs. Several Pseudo R<sup>2</sup> measures are logical analogs to OLS R<sup>2</sup> measures. To show this, 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.

```
. use http://www.nd.edu/~rwilliam/xsoc73994/statafiles/glm-reg, clear
. reg income educ
```

Source	SS	df	MS	Number of obs =	500
Model	26490.0257	1	26490.0257	F( 1, 498) =	963.54
Residual	13691.2236	498	27.4924168	Prob > F =	0.0000
				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
```

Measures of Fit for regress of income

Log-Lik Intercept Only:	-1806.106	Log-Lik Full Model:	-1536.945
D(498):	3073.890	LR(1):	538.323
		Prob > LR:	0.000
R2:	0.659	Adjusted R2:	0.659
AIC:	6.156	AIC*n:	3077.890
BIC:	-20.985	BIC':	-532.108
BIC used by Stata:	3086.319	AIC used by Stata:	3077.890

```
. logit incbinary educ, nolog
```

Logit estimates	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 .2224085
_cons	-2.245047	.3645912	-6.16	0.000	-2.959632 -1.530461

. fitstat

Measures of Fit for logit of incbinary

Log-Lik Intercept Only:	-346.574	Log-Lik Full Model:	-322.489
D(498):	644.979	LR(1):	48.168
		Prob > LR:	0.000
McFadden's R2:	0.069	McFadden's Adj R2:	0.064
ML (Cox-Snell) R2:	0.092	Cragg-Uhler(Nagelkerke) R2:	0.122
McKelvey & Zavoina's R2:	0.122	Efron's R2:	0.087
Variance of y*:	3.749	Variance of error:	3.290
Count R2:	0.660	Adj Count R2:	0.320
AIC:	1.298	AIC*n:	648.979
BIC:	-2449.896	BIC':	-41.954
BIC used by Stata:	657.408	AIC used by Stata:	648.979

Now we'll compare formulas for  $R^2$  in OLS with the suggested analogs in logistic regression (As an exercise, try doing these same computations with another data set.). Remember that  $G_M$  is the model chi-square (538.23 in the OLS regression, 48.168 in the logistic regression).

OLS Regression	Logistic Regression
<p>1A. Percentage of Explained Variation:</p> $R^2 = \frac{SSR}{SST} = \frac{26490}{40181} = .6593$ $= 1 - \frac{SSE}{SST} = 1 - \frac{13691}{40181} = .6593$	<p>1A. McFadden's <math>R^2</math> (perhaps the most popular Pseudo <math>R^2</math> of them all, and the one that Stata is reporting when it says Pseudo R2):</p> $R^2 = \frac{G_M}{Dev_0} = \frac{48.168}{-2 * -346.574} = .069$ $= 1 - \frac{LL_M}{LL_0} = 1 - \frac{-322.489}{-346.574} = .069$ $= 1 - \frac{Dev_M}{Dev_0} = 1 - \frac{644.979}{693.148} = .069$
	<p>1B. Efron's <math>R^2</math> (another logical analog to Percentage of Explained Variation):</p> $R^2 = 1 - \frac{\sum (y_i - \hat{\pi}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{114.17}{125} = .087$ <p>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 (<math>\hat{\pi}_i</math>) and then computed the squared residuals.</p>

<p>1C. Adjusted <math>R^2</math>:</p> $R^2 = 1 - \frac{SSE/(N - K)}{SST/(N - 1)} = 1 - \frac{MSE}{MST}$ $= 1 - \frac{27.4924168}{80.5235456} = .6586$ <p>Where K = the # of parameters (including the intercept). This will not necessarily go up as more variables are added.</p>	<p>1C. McFadden's Adjusted <math>R^2</math>:</p> $R^2 = 1 - \frac{LL_M - K}{LL_0} = 1 - \frac{-322.489 - 2}{-346.574} = .064$ <p>Where K = the # of parameters (including the intercept). Like Adjusted <math>R^2</math> in OLS, this will not necessarily go up as more variables are added.</p>
<p>2. Ratio of <math>\text{Var}(Y)</math> and <math>\text{Var}(\hat{Y})</math>:</p> $R^2 = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)} = 1 - \frac{\text{Var}(\varepsilon_Y)}{\text{Var}(Y)} = 1 - \frac{27.437}{80.524} = .6593$	<p>2. McKelvey &amp; Zavoina's <math>R^2</math>:</p> $R^2 = 1 - \frac{\text{Var}(\varepsilon_{Y^*})}{\text{Var}(Y^*)} = 1 - \frac{3.29}{3.749} = .122$
<p>3A. Transformation of the Likelihood Ratio:</p> $R^2 = 1 - \left[ \frac{L_0}{L_M} \right]^{2/N} = 1 - \left[ e^{LL_0 - LL_M} \right]^{2/N}$ $= 1 - \left[ e^{-1806.106 + 1536.945} \right]^{2/500} = 1 - \left[ e^{-269.161} \right]^{2/500}$ $= 1 - e^{-1.076644} = 1 - .3407 = .6593$ <p>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,</p> $R^2 = 1 - \exp(-G_M / N) = 1 - \exp(-538.323 / 500)$ $= 1 - \exp(-1.076646) = 1 - .3407 = .6593$	<p>3a. Maximum Likelihood <math>R^2</math> (SPSS calls this the Cox-Snell <math>R^2</math> and it is also called the geometric mean squared improvement per observation; You'll also see Maddala's name associated with this):</p> $R^2 = 1 - \exp(-G_M / N) = 1 - \exp(-48.168 / 500)$ $= 1 - \exp(-.096336) = 1 - .908 = .092$ <p>NOTE: This isn't just a logical analog to OLS; it is the exact same formula!</p>

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):

$$R^2 = \frac{1 - \left[ \frac{L_0}{L_M} \right]^{2/N}}{1 - L_0^{2/N}} = \frac{1 - \left[ \frac{\exp(-346.574)}{\exp(-322.489)} \right]^{2/500}}{1 - \exp(-346.574)^{2/500}}$$

$$= \frac{1 - [\exp(-24.085)]^{2/500}}{1 - \exp(-346.574)^{2/500}} = \frac{.0918448}{.75} = .122$$

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:

```
. lstat
```

Logistic model for incbinary

Classified	True		Total
	D	~D	
+	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.

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

**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. 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 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 have become increasingly popular. Some key features of these measures:

- One of their greatest strengths 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.
- 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.
- For AIC and AIC\*n, the smaller the value, the better the fit is.
- For BIC and BIC’, the more negative the value is, the better the fit is.

*AIC.* Akaike’s Information Criterion (AIC) is defined as

$$AIC = \frac{Dev_M + 2P}{N}$$

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

$$AIC = \frac{Dev_M + 2P}{N} = \frac{3073.89 + 4}{500} = \frac{3077.89}{500} = 6.156$$

and in the logistic regression example,

$$AIC = \frac{Dev_M + 2P}{N} = \frac{644.979 + 4}{500} = \frac{648.979}{500} = 1.298$$

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.

Not all authors divide by  $N$ , i.e. they would report the above AIC measures as 3077.89 and 648.979. `fitstat` reports these alternative AICs as `AIC*n`.

**BIC.** The Bayesian Information Criterion (BIC) is another measure that 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 BIC is

$$BIC_M = Dev_M - df_M \ln N$$

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

$$BIC_M = Dev_M - df_M \ln N = 3073.890 - 498 * \ln 500 = 3073.89 - 498 * 6.2146 = -20.985$$

and for the logistic regression

$$BIC_M = Dev_M - df_M \ln N = 644.979 - 498 * \ln 500 = 644.979 - 498 * 6.2146 = -2449.896$$

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

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

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

For the OLS example above,

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

For the logistic regression example,

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

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.

Either BIC or BIC' can be used to compare models, regardless of whether they are nested or not. Furthermore, it really doesn't matter which one you use, since  $BIC_1 - BIC_2 = BIC'_1 - BIC'_2$  (where the subscripts refer to the two models you are comparing).

The model with the smaller BIC or BIC' is preferred, i.e. if  $BIC_1 - BIC_2 < 0$ , model 1 is preferred. If  $BIC_1 - BIC_2 > 0$ , the second model is preferred.

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 black
```

```
. fitstat, diff
```

Measures of Fit for logit of incbinary

	Current	Saved	Difference
Model:	logit	logit	
N:	500	500	0
Log-Lik Intercept Only	-346.574	-346.574	0.000
Log-Lik Full Model	-242.471	-322.489	80.019
D	484.941(496)	644.979(498)	160.038(2)
LR	208.206(3)	48.168(1)	160.038(2)
Prob > LR	0.000	0.000	0.000
McFadden's R2	0.300	0.069	0.231
McFadden's Adj R2	0.289	0.064	0.225
ML (Cox-Snell) R2	0.341	0.092	0.249
Cragg-Uhler(Nagelkerke) R2	0.454	0.122	0.332
McKelvey & Zavoina's R2	0.523	0.122	0.400
Efron's R2	0.334	0.087	0.248
Variance of y*	6.891	3.749	3.142
Variance of error	3.290	3.290	0.000
Count R2	0.840	0.660	0.180
Adj Count R2	0.680	0.320	0.360
AIC	0.986	1.298	-0.312
AIC*n	492.941	648.979	-156.038
BIC	-2597.505	-2449.896	-147.609
BIC'	-189.562	-41.954	-147.609
BIC used by Stata	509.799	657.408	-147.609
AIC used by Stata	492.941	648.979	-156.038

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

Note: p-value for difference in LR is only valid if models are nested.

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 (.986 versus 1.298). The difference in the BICs is much bigger than 10 (-2597.505 versus -2449.896, 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<sup>2</sup> measures go up (of course, most have to when you add variables, but McFadden's Adj R<sup>2</sup> and the Adj Count R<sup>2</sup> go up too).

Compare that with the results from our earlier grade example, when we first enter gpa and psi and then enter tuce:

```
. use http://www.nd.edu/~rwilliam/xsoc73994/statafiles/glm-logit.dta, clear
. quietly logit grade gpa psi
. qui fitstat, save
. quietly logit grade gpa psi tuce
```

```
. fitstat, diff
```

Measures of Fit for logit of grade

	Current	Saved	Difference
Model:	logit	logit	
N:	32	32	0
Log-Lik Intercept Only	-20.592	-20.592	0.000
Log-Lik Full Model	-12.890	-13.127	0.237
D	25.779(28)	26.253(29)	0.474(1)
LR	15.404(3)	14.930(2)	0.474(1)
Prob > LR	0.002	0.001	0.491
McFadden's R2	0.374	0.363	0.012
McFadden's Adj R2	0.180	0.217	-0.037
ML (Cox-Snell) R2	0.382	0.373	0.009
Cragg-Uhler(Nagelkerke) R2	0.528	0.515	0.013
McKelvey & Zavoina's R2	0.544	0.520	0.024
Efron's R2	0.426	0.407	0.019
Variance of y*	7.210	6.856	0.354
Variance of error	3.290	3.290	0.000
Count R2	0.813	0.813	0.000
Adj Count R2	0.455	0.455	0.000
AIC	1.056	1.008	0.048
AIC*n	33.779	32.253	1.526
BIC	-71.261	-74.253	2.992
BIC'	-5.007	-7.999	2.992
BIC used by Stata	39.642	36.650	2.992
AIC used by Stata	33.779	32.253	1.526

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

Note: p-value for difference in LR is only valid if models are nested.

Most of the measures which can go down when variables are added (AIC, BIC, BIC', McFadden's Adj R<sup>2</sup>) do go down, while the other Pseudo R<sup>2</sup> measures go up very little. The LR chi-square contrast between the models is not significant.

*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:

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

```
. fitstat, diff
```

Measures of Fit for logit of incbinary

	Current	Saved	Difference
Model:	logit	logit	
N:	500	500	0
Log-Lik Intercept Only	-346.574	-346.574	0.000
Log-Lik Full Model	-256.251	-301.713	45.462
D	512.503(497)	603.426(498)	90.923(1)
LR	180.644(2)	89.721(1)	90.923(1)
Prob > LR	0.000	0.000	0.000
McFadden's R2	0.261	0.129	0.131
McFadden's Adj R2	0.252	0.124	0.128
ML (Cox-Snell) R2	0.303	0.164	0.139
Cragg-Uhler(Nagelkerke) R2	0.404	0.219	0.185
McKelvey & Zavoina's R2	0.442	0.248	0.194
Efron's R2	0.285	0.160	0.125
Variance of y*	5.896	4.376	1.520
Variance of error	3.290	3.290	0.000
Count R2	0.660	0.660	0.000
Adj Count R2	0.320	0.320	0.000
AIC	1.037	1.215	-0.178
AIC*n	518.503	607.426	-88.923
BIC	-2576.157	-2491.449	-84.709
BIC'	-168.215	-83.507	-84.709
BIC used by Stata	531.147	615.855	-84.709
AIC used by Stata	518.503	607.426	-88.923

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

Note: p-value for difference in LR is only valid if models are nested.

Note that the model with education and job experience fits the data much better than the model with race only. It has a smaller AIC and a more negative value on the BIC measures. 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.

*Alternative link functions: Logit versus Probit.* Incidentally, if you were hopelessly torn between the logit and probit models, you could do something like this:

```
. use http://www.nd.edu/~rwilliam/xsoc73994/statafiles/glm-logit.dta, clear
. quietly logit grade gpa tuce psi
. quietly fitstat, save
. quietly probit grade gpa tuce psi
. fitstat, diff
```

Measures of Fit for probit of grade

Current model estimated by probit, but saved model estimated by logit.  
To make the comparisons anyway, use the force option.

```
. fitstat, dif force
```

Measures of Fit for probit of grade

Warning: Current model estimated by probit, but saved model estimated by logit

	Current	Saved	Difference
Model:	probit	logit	
N:	32	32	0
Log-Lik Intercept Only:	-20.592	-20.592	0.000
Log-Lik Full Model:	-12.819	-12.890	0.071
D:	25.638(28)	25.779(28)	0.142(0)
LR:	15.546(3)	15.404(3)	0.142(0)
Prob > LR:	0.001	0.002	.
McFadden's R2:	0.377	0.374	0.003
McFadden's Adj R2:	0.183	0.180	0.003
Maximum Likelihood R2:	0.385	0.382	0.003
Cragg & Uhler's R2:	0.532	0.528	0.004
McKelvey and Zavoina's R2:	0.570	0.544	0.027
Efron's R2:	0.422	0.426	-0.004
Variance of y*:	2.328	7.210	-4.882
Variance of error:	1.000	3.290	-2.290
Count R2:	0.813	0.813	0.000
Adj Count R2:	0.455	0.455	0.000
AIC:	1.051	1.056	-0.004
AIC*n:	33.638	33.779	-0.142
BIC:	-71.403	-71.261	-0.142
BIC':	-5.149	-5.007	-0.142

Difference of 0.142 in BIC' provides weak support for current model.

Note: p-value for difference in LR is only valid if models are nested.

(Note how it warned me about comparing models estimated in different ways; and I made it do the comparison anyway by adding the force option to fitstat.) As we see, there is very little reason for preferring logit over probit or vice-versa in this case.

**Additional Reading.** 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.