Scalar Measures of Fit for Regression Models[1]

J. Scott Long and Jeremy Freese
Indiana University and University of Wisconsin-Madison

February 10, 2000

Overview

Many scalar measures have been developed to summarize the overall goodness of fit for regression models of continuous, count, or categorical dependent variables. The post-estimation command `fitstat` calculates a large number of fit statistics for the estimation commands `clogit`, `cnreg`, `cloglog`, `gologit`, `intreg`, `logistic`, `logit`, `mlogit`, `nbreg`, `ologit`, `oprobit`, `omodel`, `poisson`, `probit`, `regress`, `zinb`, and `zip`. With its `saving()` and `using()` options, the command also allows the comparison of fit measures across two models. While `fitstat` duplicates some measures computed by other commands (e.g., the pseudo-R^2 in standard Stata output; `lfit`), `fitstat` adds many more measures and makes it convenient to compare measures across models. Details on the measures that are discussed below can be found in Long (1997) which cites the original sources for each measure and provides further details on their derivation. *Note to the Editor: If you prefer, the original sources can be cited in this article.*

Before proceeding, a word of caution regarding the use of these measures. A scalar measure of fit can be useful in comparing competing models and ultimately in selecting a final model. Within a substantive area of research, measures of fit can provide a *rough* index of whether a model is adequate. However, there is no convincing evidence that selecting a model that maximizes the value of a given measure results in a model that is optimal in any sense other than the model having a larger (or smaller) value of that measure. While measures of fit provide some information, it is only partial information that must be assessed within the context of the theory motivating the analysis, past research, and the estimated parameters of the model being considered.

Syntax

```
fitstat [, saving(name) save using(name) dif bic force]
```

While many measures of fit are based on values returned by the estimation command, for some measures it is necessary to compute additional statistics from the estimation sample. While `fitstat` does not include `if` and `in` options, analysis is based on the sample defined by `e(sample)` from the last model estimated. Accordingly, `fitstat` is appropriate for models estimated using `if` and `in` restrictions in the original model. `fitstat` can also be used when models are estimated with weighted data. Here there are two limitations. First, some measures cannot be computed with some types of weights. Second, with `pweights` we use values of the "pseudo-likelihoods" to compute our measures of fit. Given the heuristic

[1] We thank David M. Drukker, Senior Statistician at Stata Corportion for his helpful suggestions.

nature of the various measures of fit, we see no reason why the resulting measures would be appropriate. `fitstat` ends with an error if the last estimation command does not return a value for the log-likelihood equation with only an intercept (i.e., if `e(ll_0)` returns a missing value). This will occur, for example, if the `noconstant` option is used with the estimation command.

Options

`saving(`*name*`)` saves the computed measures in a matrix for subsequent comparisons. *name* cannot be longer than 4 characters.

`save` is equivalent to `saving(0)` .

`using(`*name*`)` compares the fit measures for the current model with those of the model saved as *name*. *name* cannot be longer than 4 characters.

`dif` is equivalent to `using(0)` .

`bic` presents only BIC and other information measures. In comparing two models, `fitstat` reports Raftery's (1996) guidelines for assessing the strength of one model over another.

`force` allows comparison of two models even when the number of observations or the estimation method varies between the two models.

Models and Measures

Details on the measures of fit are given below. Here we only summarize which measures are computed for which models. ■ indicates a measure is computed; □ indicates the measure is not computed.

							`gologit`
							`nbreg`
		`logistic`				`cnreg`	`poisson`
		`logit`		`ologit`	`clogit`	`intreg`	`zinb`
	`regress`	`probit`	`cloglog`	`oprobit`	`mlogit`	`tobit`	`zip`
Log-likelihood	■	■	■[1]	■	■	■	■[2]
Deviance & LR chi-square	■	■	■	■	■	■	■
AIC, AIC*n, BIC, BIC′	■	■	■	■	■	■	■
R^2 & Adjusted R^2	■	□	□	□	□	□	□
Efron's R^2	□	■	■	□	□	□	□
McFadden's, ML, C&U's R^2	□	■	■	■	■	■	■
Count & Adjusted Count R^2	□	■	■	■	■[3]	□	□
Var(e), Var(y*) and M&Z's R^2	□	■	□	■	□	■	□

1 - For `cloglog` the log-likelihood for the intercept-only model does not correspond to the first step in the iterations.

2 - For `zip` and `zinb`, the log-likelihood for the intercepts-only model is calculated by estimating `zip|zinb` *`lhs-variable`*`, inf(_cons)`.

3 - The adjusted count R^2 is not defined for `clogit`.

Example

To compute fit statistics for a single model:

```
. use mroz, clear

(PSID 1976 from T. Mroz)

. * compute fit statistics for a single model
. logit lfp k5 k618 age wc hc lwg inc

Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -454.32339
Iteration 2:   log likelihood = -452.64187
Iteration 3:   log likelihood = -452.63296
Iteration 4:   log likelihood = -452.63296

Logit estimates                                   Number of obs   =        753
                                                  LR chi2(7)      =     124.48
                                                  Prob > chi2     =     0.0000
Log likelihood = -452.63296                       Pseudo R2       =     0.1209

------------------------------------------------------------------------------
     lfp |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      k5 |  -1.462913   .1970006     -7.426   0.000      -1.849027   -1.076799
    k618 |  -.0645707   .0680008     -0.950   0.342      -.1978499    .0687085
     age |  -.0628706   .0127831     -4.918   0.000      -.0879249   -.0378162
      wc |   .8072738   .2299799      3.510   0.000       .3565215    1.258026
      hc |   .1117336   .2060397      0.542   0.588      -.2920969     .515564
     lwg |   .6046931   .1508176      4.009   0.000       .3090961    .9002901
     inc |  -.0344464   .0082084     -4.196   0.000      -.0505346   -.0183583
   _cons |    3.18214   .6443751      4.938   0.000       1.919188    4.445092
------------------------------------------------------------------------------

. fitstat

Measures of Fit for logit of lfp
Log-Lik Intercept Only:     -514.873    Log-Lik Full Model:         -452.633
D(745):                      905.266    LR(7):                       124.480
                                        Prob > LR:                     0.000
McFadden's R2:                 0.121    McFadden's Adj R2:             0.105
Maximum Likelihood R2:         0.152    Cragg & Uhler's R2:            0.204
McKelvey and Zavoina's R2:     0.217    Efron's R2:                    0.155
Variance of y*:                4.203    Variance of error:             3.290
Count R2:                      0.693    Adj Count R2:                  0.289
AIC:                           1.223    AIC*n:                       921.266
BIC:                       -4029.663    BIC':                        -78.112
```

To compute and save fit measures:

```
. logit lfp k5 k618 age wc hc lwg inc

:::output same as above:::

. fitstat, saving(mod1)

:::output same as above:::

(Indices saved in matrix fs_mod1)
```

To compare saved model to current model:

```
. logit lfp k5 age age2 wc inc

:::output not shown:::

. fitstat, using(mod1)

Measures of Fit for logit of lfp

                             Current            Saved       Difference
Model:                         logit            logit
N:                               753              753                0
Log-Lik Intercept Only:     -514.873         -514.873            0.000
Log-Lik Full Model:         -461.653         -452.633           -9.020
D:                           923.306(747)     905.266(745)      18.040(2)
LR:                          106.441(5)       124.480(7)       -18.040(-2)
Prob > LR:                     0.000            0.000            0.000
McFadden's R2:                 0.103            0.121           -0.018
McFadden's Adj R2:             0.092            0.105           -0.014
Maximum Likelihood R2:         0.132            0.152           -0.021
Cragg & Uhler's R2:            0.177            0.204           -0.028
McKelvey and Zavoina's R2:     0.182            0.217           -0.035
Efron's R2:                    0.135            0.155           -0.020
Variance of y*:                4.023            4.203           -0.180
Variance of error:             3.290            3.290            0.000
Count R2:                      0.677            0.693           -0.016
Adj Count R2:                  0.252            0.289           -0.037
AIC:                           1.242            1.223            0.019
AIC*n:                       935.306          921.266           14.040
BIC:                       -4024.871        -4029.663            4.791
BIC':                        -73.321          -78.112            4.791

Difference of    4.791 in BIC' provides positive support for saved model.
```

Saved Results

`fitstat` saves in `r()` whichever of the following are computed for a particular model:

`r(aic)` - AIC
`r(aic_n)` - AIC×N
`r(bic)` - BIC
`r(bic_p)` - BIC$'$
`r(dev)` - deviance
`r(dev_df)` - degrees of freedom for deviance
`r(ll)` - log-likelihood for full model
`r(ll_0)` - log-likelihood for model with only intercept
`r(lrx2)` - likelihood ratio chi-square
`r(lrx2_df)` - degrees of freedom for likelihood ratio chi-square
`r(lrx2_p)` - probability level of chi-square test
`r(N)` - number of observations
`r(n_parm)` - number of parameters
`r(n_rhs)` - number of right hand side variables
`r(r2)` - R^2 for linear regression model
`r(r2_adj)` - adjusted R^2 for linear regression model
`r(r2_ct)` - count R^2
`r(r2_ctadj)` - adjusted count R^2
`r(r2_cu)` - Cragg & Uhler's R^2
`r(r2_ef)` - Efron's R^2
`r(r2_mz)` - McKelvey and Zavoina's R^2
`r(r2_mf)` - McFadden's R^2
`r(r2_mfadj)` - McFadden's adjusted R^2
`r(r2_ml)` - maximum likelihood R^2
`r(v_error)` - variance of error term
`r(v_ystar)` - variance of y^*

When the `saving(`*`name`*`)` option is specified, computed measures are also saved in matrix `fs_`*`name`*. The column names of the matrix correspond to the names of the measures listed above. The row name is the command used to estimate the saved model. Values of -9999 in the matrix indicate that a measure is not appropriate for the given model. The row name is the name of the estimation procedure.

Extending fitstat to other Models and Measures

`fitstat` can be extended to other models and measures of fit. When doing this, there are several things to keep in mind. First, the program depends on values returned by `eclass` estimation commands (e.g., `e(sample)`). Programs like `ocratio` which do not use eclass returns cannot incorporated into `fitstat` without major changes to the structure of the program. Second, not all measures of fit are appropriate for all models. The programmer must be careful to ensure that `fitstat` does not automatically compute inappropriate fit statistics. Third, the way in which values such as the number of parameters and the number of right-hand-side variables are computed differs across models. Consequently, additional code may be needed for these computations.

Methods and Formulas

This provides brief descriptions of each measure computed by `fitstat`. Full details along with citations to original sources are found in Long (1997). The measures are listed in the same order as the output illustrated above.

Log-Likelihood Based Measures Stata begins maximum likelihood iterations by computing the log-likelihood of the model with all parameters but the intercept(s) constrained to zero, referred to as $L\left(M_{\text{Intercept}}\right)$ below. The log-likelihood upon convergence, referred to as M_{Full} below, is also listed. In Stata this information is usually presented as the first step of the iterations and in the header for the estimation results. Note that in `cloglog`, the value at iteration 0 is not the log-likelihood with only the intercept. For `zip` and `zinb`, the "intercept-only" model can be defined in different ways. These commands return as `e(ll_0)` the value of the log-likelihood with the binary portion of the model unrestricted while only the intercept is free for the Poisson or negative binomial portion of the model. Alternatively, `fitstat` returns the value of the log-likelihood from the model with only an intercept in both the binary and count portion of the model.

Chi-square Test of All Coefficients: *LR* A likelihood ratio test of the hypothesis that all coefficients except the intercept(s) can be computed by comparing the log-likelihoods: $LR = 2\ln L(M_{\text{Full}}) - 2\ln L(M_{\text{Intercept}})$. This statistic is sometimes designated as G^2. *LR* is reported by Stata as: `LR chi2(7) = 124.48` where the degrees of freedom, (7), are the number of constrained parameters. `fitstat` reports this statistic as: `LR(7):   124.48` For `zip` and `zinb`, *LR* tests that the coefficients in the count portion (not the binary portion) of the model are zero.

Deviance: *D* The *deviance* compares a given model to a model that has one parameter for each observation and can reproduce perfectly the observed data. The deviance is defined as $D = -2\ln L(M_{\text{Full}})$, where the degrees of freedom equals N minus the number of parameters. Note that D does not have a chi-square distribution.

$\mathbf{R}^2$ in the LRM For `regress`, `fitstat` reports the standard coefficient of determination which can be defined variously as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2} = \frac{\widehat{Var}(\widehat{y})}{\widehat{Var}(\widehat{y}) + \widehat{Var}(\widehat{\varepsilon})} = 1 - \left[\frac{L(M_{\text{Intercept}})}{L(M_{\text{Full}})}\right]^{2/N} \tag{1}$$

The adjusted R^2 is defined as:

$$\bar{R}^2 = \left(R^2 - \frac{K}{N-1}\right)\left(\frac{N-1}{N-K-1}\right)$$

where K is the number of independent variables.

Pseudo-R^2's While each of the definitions of R^2 in equation 1 give the same numeric value in the LRM, they give different answers and thus provide different measures of fit when applied to the other models evaluated by `fitstat`.

McFadden's R^2 McFadden R^2, also known as "likelihood ratio index," compares a model with just the intercept to a model with all parameters. It is defined as:

$$R^2_{\text{McF}} = 1 - \frac{\ln \widehat{L}(M_{\text{Full}})}{\ln \widehat{L}(M_{\text{Intercept}})}$$

If model $M_{\text{Intercept}} = M_{\text{Full}}$, R^2_{McF} equals 0, but R^2_{McF} can never exactly equal one. This measure, which is computed by Stata as `Pseudo R2 = 0.1209`, is listed in `fitstat` as: `McFadden's R2: 0.121` Since R^2_{McF} always increases as new variables are added, an adjusted version is also available:

$$\bar{R}^2_{\text{McF}} = 1 - \frac{\ln \widehat{L}(M_{\text{Full}}) - K^*}{\ln \widehat{L}(M_{\text{Intercept}})}$$

where K^* is the number of parameters (not independent variables).

Maximum Likelihood R^2 Another analogy to R^2 in the LRM was suggested by Maddala:

$$R^2_{\text{ML}} = 1 - \left[\frac{L(M_{\text{Intercept}})}{L(M_{\text{Full}})}\right]^{2/N} = 1 - \exp(-G^2/N)$$

Cragg & Uhler's R^2 Since R^2_{ML} only reaches a maximum of $1 - L(M_{\text{Intercept}})^{2/N}$, Cragg and Uhler suggested a normed measure:

$$R^2_{\text{C\&U}} = \frac{R^2_{\text{ML}}}{\max R^2_{\text{ML}}} = \frac{1 - \left[\, L(M_{\text{Intercept}}) / L(M_{\text{Full}}) \,\right]^{2/N}}{1 - L(M_{\text{Intercept}})^{2/N}}$$

Efron's R^2 For binary outcomes, Efron's pseudo-R^2 defines $\widehat{y} = \widehat{\pi} = \widehat{\Pr}(y = 1 \mid \mathbf{x})$ and equals:

$$R^2_{\text{Efron}} = 1 - \frac{\sum_{i=1}^{N} (y_i - \widehat{\pi}_i)^2}{\sum_{i=1}^{N} (y_i - \overline{y})^2}$$

$\mathbf{V}(y^*)$, $\mathbf{V}(\varepsilon)$ and McKelvey and Zavoina's R^2 Some models can be defined in terms of a latent variable y^*. This includes the models for binary or ordinal outcomes: `logit`, `probit`, `ologit` and `oprobit`, as well as some models with censoring: `tobit`, `cnreg`, and `intreg`. Each model is defined in terms of a regression on a latent variable y^*:

$$y^* = \mathbf{x}\boldsymbol{\beta} + \varepsilon$$

Using $\widehat{\mathit{Var}}(\widehat{y}^*) = \widehat{\boldsymbol{\beta}}' \, \widehat{\mathit{Var}}(\mathbf{x}) \, \widehat{\boldsymbol{\beta}}$, McKelvey and Zavoina proposed:

$$R^2_{M\&Z} = \frac{\widehat{\mathit{Var}}(\widehat{y}^*)}{\widehat{\mathit{Var}}(y^*)} = \frac{\widehat{\mathit{Var}}(\widehat{y}^*)}{\widehat{\mathit{Var}}(\widehat{y}^*) + \mathit{Var}(\varepsilon)}$$

In models for categorical outcomes, $\mathrm{Var}(\varepsilon)$ is assumed to identify the model; in models with censoring it can be estimated.

Count and Adjusted Count R^2 Observed and predicted values can be used in models with categorical outcomes to compute what is known as the count R^2. Consider the binary case where the observed y is 0 or 1 and $\pi_i = \widehat{\Pr}(y = 1 \mid \mathbf{x}_i)$. Define the expected outcome $\widehat{y}$ as

$$\widehat{y}_i = \begin{cases} 0 & \text{if } \widehat{\pi}_i \leq 0.5 \\ 1 & \text{if } \widehat{\pi}_i > 0.5 \end{cases}$$

This allows us to construct a table of observed and predicted values, such as produced by the Stata command `lstat`:

```
               -------- True --------
Classified |         D            ~D           Total
-----------+--------------------------+-----------
     +     |       342           145  |         487
     -     |        86           180  |         266
-----------+--------------------------+-----------
   Total   |       428           325  |         753
```

A seemingly appealing measure is the proportion of correct predictions, referred to as the *count R^2*:

$$R^2_{\text{Count}} = \frac{1}{N} \sum_j n_{jj}$$

where the n_{jj}'s are the number of correct predictions for outcome j. The count R^2 can give the faulty impression that the model is predicting very well. In a binary model *without* knowledge about the independent variables, it is possible to correctly predict at least 50 percent of the cases by choosing the outcome category with the largest percentage of observed cases. To adjust for the largest row marginal:

$$R^2_{\text{AdjCount}} = \frac{\sum_j n_{jj} - \max\limits_r (n_{r+})}{N - \max\limits_r (n_{r+})}$$

where n_{r+} is the marginal for row r. The *adjusted count R^2* is the proportion of correct guesses beyond the number that would be correctly guessed by choosing the largest marginal.

Information Measures This class of measures can be used to compare models across different samples or to compare non-nested models.

AIC Akaike's (1973) information criteria is defined as AIC= $\dfrac{-2\ln \widehat{L}(M_k)+2P}{N}$ where $\widehat{L}(M_k)$ is the likelihood of the model and P is the number of parameters in the model (e.g., $K+1$ in the binary regression model where K is the number of regressors). All else being equal, the model with the smaller AIC is considered the better fitting model. Some authors define AIC as being N times the value we report. This is done in `mlfit` (Tobias and Campbell, STB-45). We report this quantity as `AIC*n` .

BIC and BIC′ The Bayesian information criterion has been proposed by Raftery (1996 and the literature cited therein) as a measure of overall fit and a means to compare nested and non-nested models. Consider the model M_k with deviance $D(M_k)$. BIC is defined as:

$$\text{BIC}_k = D(M_k) - df_k \ln N$$

where df_k is the degrees of freedom associated with the deviance. The more negative the BIC_k the better the fit. A second version of BIC is based on the LR chi-square with df'_k equal to the number of regressors (not parameters) in the model. Then:

$$\text{BIC}'_k = -G^2(M_k) + df'_k \ln N$$

The more negative the BIC'_k the better the fit. The difference in the BICs from two models indicates which model is more likely to have generated the observed data. Since $\text{BIC}_1 - \text{BIC}_2 = \text{BIC}'_1 - \text{BIC}'_2$ the choice of which BIC measure to use is a matter of convenience. If $\text{BIC}_1 - \text{BIC}_2 < 0$, then the first model is preferred. If $\text{BIC}_1 - \text{BIC}_2 > 0$, then the second model is preferred. Raftery suggested guidelines for the strength of evidence favoring M_2 against M_1 based on a difference in BIC or BIC′:

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

References

Long, J. Scott. (1996). *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA: Sage.

Raftery, A. E. (1996). Bayesian Model Selection in Social Research. In P. V. Marsden (Ed.), *Sociological Methodology,* (Vol. 25, pp. 111-163). Oxford: Basil Blackwell.