MWSUG 2016 - Paper AA19

Is My Model the Best? Methods for Exploring Model Fit

Deanna N Schreiber-Gregory, National University, Moorhead, MN

ABSTRACT

In the submission/presentation phase of any research or analytics project, it is reasonable to expect the reception of many types of questions aimed at clarifying the reliability and accuracy of the project's results. One of the most common questions to expect would be: "So the model provides a feasible answer to the question, but does it provide the best answer?" One way to answer this question with utmost confidence is to provide a variety of model fit analyses designed to support the conclusion of your final model. This paper provides a variety of techniques aimed at model fit exploration including default procedure settings as well as additive options. This paper will review the theory behind each of these fit procedures and the pros and cons of their use. Optional R-square calculations will also be explored. This paper is intended for any level of SAS® user. This paper is also written to an audience with a background in behavioral science and/or statistics.

INTRODUCTION

A staple of scientific advancement is the ever-evolving mathematical relationship between predictor and outcome. One of the main problems that arise from this evolution is the question as to whether the model that we are using is appropriate and/or an improvement from earlier iterations of the model. There is an easy solution to this and that is the examination of several forms of predictive power and goodness-of-fit measures.

INTRODUCTION TO THE DATA SET

The Youth Risk Behavior Surveillance System (YRBSS) was developed as a tool to help monitor priority risk behaviors that contribute substantially to death, disability, and social issues among American youth and young adults today. The YRBSS has been conducted biennially since 1991 and contains survey data from national, state, and local levels. The national Youth Risk Behavior Survey (YRBS) provides the public with data representative of the United States high school students. On the other hand, the state and local surveys provide data representative of high school students in states and school districts who also receive funding from the CDC through specified cooperative agreements. The YRBSS serves a number of different purposes. The system was originally designed to measure the prevalence of health-risk behaviors among high school students. It was also designed to assess whether these behaviors would increase, decrease, or stay the same over time. An additional purpose for the YRBSS is to have it examine the co-occurrence of different health-risk behaviors. This particular study exams the co-occurrence of suicidal ideation as an indicator of psychological unrest with other health-risk behaviors. The purpose of this study is to serve as an exercise in correlating two different variables across multiple years with large data sets.

MODEL FIT AND POWER MEASURES

In this paper we will examine several available goodness-of-fit and predictive power measures for a few different regression modeling procedures. But first, it is important for us to cover the nuances and differences between these two different model fit statistics.

POWER MEASURES

Measures of predictive power typically have values that fall between 0 and 1, with 0 indicating a complete lack of predictive power and 1 indicating a perfect predictive relationship. As a general rule, the higher the value, the better, but other than that there are rarely any fixed cut-off values that differentiate whether a model is acceptable or not.

MODEL FIT MEASURES

Goodness-of-fit measures are formal tests of the null hypothesis that the fitted model is correct. These measures output a p-value which is used to decide whether or not the indicated model is a good fit. P-values are numbers between 0 and 1 with higher values indicating a better fit. Contrary to the traditional view of p-values, where one would specify a target α level (such as .05) and accept a model with a p-value below this value, goodness-of-fit test p-values that land below the specified alpha level would indicate that a model is not acceptable.

HOW THESE MEASURS ARE DIFFERENT

It is important to note that goodness-of-fit measures and predictive power measures are testing two very different concepts. It should not be surprising for a model that has a very high R-square to also produce an unacceptable goodness-of-fit statistic. The opposite is also true, with the common existence of models with very low R-square and ideal goodness-of-fit scores. One way to look at these two concepts is like this: R-square scores test how much of the variation in response seen in the outcome variable can be explained by the proposed model, whereas goodness-of-fit scores do not tell you how well the outcome variable is predicted by the model, but rather if a specific model can do a better job at explaining the relationship between predictor and outcome than a previously proposed model. Through goodness-of-fit tests, the analyst can qualitatively compare different models while exploring the utilization of more complex concepts such as the addition of non-linearities, interactions, or changing the link function.

FIT AND POWER MEASURES IN SAS

Now that we have covered what predictive power and goodness-of-fit tests are and how they are different, now we will explore the different measures that are available for some of the regression procedures available in SAS and how to implement them.

PROC LOGISTIC & SURVEYLOGISTIC: MODEL FIT AND POWER

Measures of predictive power in PROC LOGISTIC include a Cox-Snell version R-square, the area under the ROC curve, and some rank-order correlations.

The Cox-Snell R-square and a max-rescaled version of the Cox-Snell R-square statistic (utilized in order to adjust for an upper-bound issue with Cox-Snell) can be produced through the addition of the RSQ option in the MODEL statement:

```
proc logistic data = newYRBS_Total;
     class   SI_Cat(ref='No')  SubAbuse_Cat (ref='1 None') Age_Cat (ref='12 or
     younger') Sex_Cat (ref='Female') Race_Cat (ref='White') Depression_Cat
     (ref='No') RecSubAbuse_Cat (ref='No') VictimViol_Cat (ref='None')
     ActiveViol_Cat (ref='None') / param=ref;
     model SI_Cat = SubAbuse_Cat Age_Cat Sex_Cat Race_Cat Depression_Cat
     RecSubAbuse_Cat VictimViol_Cat ActiveViol_Cat / rsq;
     title 'Predictive Power: Cox-Snell';
run;
```

You can also easily calculate alternative predictive power estimates such as Tjur:

```
proc logistic data = newYRBS_Total;
     class   SI_Cat(ref='No')  SubAbuse_Cat (ref='1 None') Age_Cat (ref='12 or
     younger') Sex_Cat (ref='Female') Race_Cat (ref='White') Depression_Cat
     (ref='No') RecSubAbuse_Cat (ref='No') VictimViol_Cat (ref='None')
     ActiveViol_Cat (ref='None') / param=ref;
     model SI_Cat = SubAbuse_Cat Age_Cat Sex_Cat Race_Cat Depression_Cat
     RecSubAbuse_Cat VictimViol_Cat ActiveViol_Cat;
     output out=Tjur pred=yhat;
     title 'Predictive Power: Tjur';
proc ttest data=a;
     class SI_Cat;
     vary hat;
run;
```

In this example, the OUTPUT statement will produce a new data set that we will call “Tjur” with the predicted probabilities stored in a new variable called “yhat”. PROC TTEST can then be used to compute the mean of the predicted probabilities for each category of the identified dependent variable (SI_Cat), and then take their difference. For more information on the statistics and controversy behind the use of each of these different R-square calculations, please see Dr. Paul D. Allison’s 2014 Global Forum paper, highlighted in the reference section of this paper.

The area under the receiver operating characteristic (ROC) curve (calculated when the outcome variable is binary) and three other indices of rank-order correlations are calculated by default in PROC LOGISTIC and outputted as “c”, “Somers’ D (Gini coefficient)”, “Goodman-Kruskal Gamma”, and “Kendall’s Tau-a”.

Measures of goodness-of-fit in PROC LOGISTIC include calculations of deviance, Pearson chi-square, Hosmer-Lemeshow, Akaike Information Criterion, The Bayesian Information Criterion, -2LogL, Stukel’s test, Information Matrix Test, Unweighted Sum of Squares, and Standardized Pearson Test. The latter three of this list are automatically derived through implementation of the PROC LOGISTIC procedure.

We can calculate deviance and Pearson goodness-of-fit by specifying SCALE=NONE in the model statement:

```
proc logistic data = newYRBS_Total;
     class   SI_Cat(ref='No')  SubAbuse_Cat (ref='1 None') Age_Cat (ref='12 or
     younger') Sex_Cat (ref='Female') Race_Cat (ref='White') Depression_Cat
     (ref='No') RecSubAbuse_Cat (ref='No') VictimViol_Cat (ref='None')
     ActiveViol_Cat (ref='None') / param=ref;
     model SI_Cat = SubAbuse_Cat Age_Cat Sex_Cat Race_Cat Depression_Cat
     RecSubAbuse_Cat VictimViol_Cat ActiveViol_Cat / scale=none;
     title 'Goodness-Of-Fit: Deviance & Pearson Goodness-Of-Fit’;
run;
```

If we specify the ODS Graphics statement and the PLOTS= option then we can produce graphical displays of the ROC curve of model fit:

```
ODS graphics ON;
proc logistic data = newYRBS_Total plots=all;
     class   SI_Cat(ref='No')  SubAbuse_Cat (ref='1 None') Age_Cat (ref='12 or
     younger') Sex_Cat (ref='Female') Race_Cat (ref='White') Depression_Cat
     (ref='No') RecSubAbuse_Cat (ref='No') VictimViol_Cat (ref='None')
     ActiveViol_Cat (ref='None') / param=ref;
     model SI_Cat = SubAbuse_Cat Age_Cat Sex_Cat Race_Cat Depression_Cat
     RecSubAbuse_Cat VictimViol_Cat ActiveViol_Cat;
          title 'Goodness-Of-Fit: ROC Plots’;
run;
ODS graphics OFF;
```

The Hosmer-Lemeshow Test can be produced by specifying the LACKFIT option in the model statement:

```
proc logistic data = newYRBS_Total;
     class   SI_Cat(ref='No')  SubAbuse_Cat (ref='1 None') Age_Cat (ref='12 or
     younger') Sex_Cat (ref='Female') Race_Cat (ref='White') Depression_Cat
     (ref='No') RecSubAbuse_Cat (ref='No') VictimViol_Cat (ref='None')
     ActiveViol_Cat (ref='None') / param=ref;
     model SI_Cat = SubAbuse_Cat Age_Cat Sex_Cat Race_Cat Depression_Cat
     RecSubAbuse_Cat VictimViol_Cat ActiveViol_Cat / lackfit;
     title 'Goodness-Of-Fit: Hosmer-Lemeshow’;
run;
```

The Stukel Test can be produced through the following sequence of events:

```
proc logistic data = newYRBS_Total;
     class   SI_Cat(ref='No')  SubAbuse_Cat (ref='1 None') Age_Cat (ref='12 or
     younger') Sex_Cat (ref='Female') Race_Cat (ref='White') Depression_Cat
     (ref='No') RecSubAbuse_Cat (ref='No') VictimViol_Cat (ref='None')
     ActiveViol_Cat (ref='None') / param=ref;
```

```
     model SI_Cat = SubAbuse_Cat Age_Cat Sex_Cat Race_Cat Depression_Cat
     RecSubAbuse_Cat VictimViol_Cat ActiveViol_Cat;
     output out=a xbeta=xb;
data b;
     set a
     za=xb**2*(xb>=0);
     zb=xb**2*(xb<0);
     num=1;
proc logistic data=b;
     model SI_Cat = SubAbuse_Cat Age_Cat Sex_Cat Race_Cat Depression_Cat
     RecSubAbuse_Cat VictimViol_Cat ActiveViol_Cat za zb;
     test za=0,zb=0;
     title 'Goodness-Of-Fit: Stukel's Test';
run;
```

Some of these newer goodness-of-fit tests can also be implemented through utilization of the GOFLOGIT macro developed by Oliver Kruss and presented at SUGI 25 in 2001. The macro can be downloaded at https://github.com/friendly/SAS-macros/blob/master/goflogit.sas. Through this macro you can implement the Standardized Pearson Test, Unweighted Sum of Squares, and the Information Matrix Test.

Going back to the work used in implementing the Stukel's Test, we see that in the OUTPUT statement we produce a new data set A that contains all of the variables in the model plus a new variable named "XB". XB is a linear predictor based on the fitted model. We then implement the DATA step through which the two vaiables needed for the Stukel test are created, in addition to this NUM=1 is identified In order to create a new "variable" that will be needed for the GOFLOGIT macro. The second logistic procedure is then implemented in order to estimate the extended model with the addition of the two new variables, while also testing the null hypothesis that both ZA and ZB have coefficients of 0.

In order to then calculate the other GOF statistics, the GOFLOGIT macro is called through the following statement:

```
%goflogit(data=b, y=SI_Cat, xlist=SubAbuse_Cat Age_Cat Sex_Cat Race_Cat
Depression_Cat RecSubAbuse_Cat VictimViol_Cat ActiveViol_Cat, trials=num)
```

This macro fits the logistic regression model with the outcome variable specified in Y= and the predictor variables specified in XLIST=. TRIALS=NUM is then specified given that the macro is designed to calculate GOF statistics for either grouped or ungrouped data. In the instance the data is ungrouped, the number of trails must be set to 1.

PROC PHREG: MODEL FIT AND POWER

Goodness-of-fit measures in PROC PHREG include -2logL, Akaike Information Criterion, and the SBC statistic. These measures are automatically produced with implementation of the PHREG procedure and are located in the Model Fit Statistics section of the output.

PROC REG: MODEL FIT AND POWER

Measures of predictive power in PROC REG include R-square and Adjusted R-Square. These measures can be requested by indicating EDF or RSQUARE for the R-square statistic or ADJRSQ for the Adjusted R-Square statistic in the MODEL statement:

```
proc reg data = newYRBS_Total;
     model SI_Cat = SubAbuse_Cat Age_Cat Sex_Cat Race_Cat Depression_Cat
     RecSubAbuse_Cat VictimViol_Cat ActiveViol_Cat / edf rsquare adjrsq;
     title 'Predictive Power for PROC REG';
run;
```

Measures of goodness-of-fit in PROC REG include calculations of Akaike Information Criterion, The Bayesian Information Criterion, Mallows Cp, Estimated MSE of prediction assuming multivariate normality, Jp or the final prediction error, Amemiya's prediction criterion, root MSE, SBC statistic, and Sp statistic.

These measures can be requested by indicating AIC for Akaike Information Criterion, BIC for Bayesian Information Criterion, CP for Mallow's Cp, GMSEP for estimated MSE of prediction assuming multivariate normality, JP for Jp, PC for Amemiya's prediction criterion, RMSE for root MSE, SBC for SBC statistic, and SP for Sp statistic in the MODEL statement, similar to the above code for predictive power.

PROC GENMOD: MODEL FIT AND POWER

Measures of goodness-of-fit in PROC GENMOD include calculations of deviance, Pearson chi-square, Akaike Information Criterion, The Bayesian Information Criterion, and -2LogL. All of which are automatically derived through implementation of the PROC GENMOD procedure.

CONCLUSION

This paper covered several different predictive modeling and goodness-of-fit tests and how to implement them in a few different regression procedures. It also covered different resources for further review and exploration of these procedures. For information involving the output of the example coding and models or for full coding examples, please contact the author.

REFERENCES

Allison, P. D. 2014. "Measures of Fit for Logistic Regression." *Proceedings of the SAS Global 2014 Conference.* Washington, DC.

Bellocco, R., and Algeri, S. 2011. "Goodness of Fit Tests for Categorical Data: Comparing Stat, R, and SAS." *Karolinksa Institutet*. Stockholm, Sweden.

Chesher A. (1984) "Testing for neglected heterogeneity." *Econometrica* 52:865–872.

Cragg, J.G. and R.S. Uhler (1970) "The demand for automobiles." *The Canadian Journal of Economics* 3: 386-406.

Copas, J.B. (1989) "Unweighted sum of squares test for proportions." *Applied Statistics* 38:71 –80.

Cox, D.R. and E.J. Snell (1989) *Analysis of Binary Data*. Second Edition. Chapman & Hall.

Farrington, C. P. (1996) "On assessing goodness of fit of generalized linear models to sparse data." *Journal of the Royal Statistical Society, Series B* **58**: 344–366.

Frost, J. "Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?" *The Minitab Blog.* 2013. Available at http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit.

Hosmer, D.W. and N.L. Hjort (2002) "Goodness-of-fit processes for logistic regression: Simulation results." *Statistics in Medicine* 21:2723–2738.

Hosmer, D.W., T. Hosmer, S. Le Cessie and S. Lemeshow (1997). "A comparison of goodness-of-fit tests for the logistic regression model." *Statistics in Medicine* 16: 965–980. 12

Hosmer D.W. and S. Lemeshow (1980) "A goodness-of-fit test for the multiple logistic regression model." *Communications in Statistics* A10:1043-1069.

Hosmer D.W. and S. Lemeshow (2013) *Applied Logistic Regression*, 3rd Edition. New York: Wiley.

Hosmer, D.W., Taber, S., Lemeshow, S. (1991). *The Importance of Assessing the Fit of Logistic Regression Models: A Case Study.* American Jounral of Public Health. 81(12): 1630-1635.

Karen. "Assessing the Fit of Regression Models." *The Analysis Factor.* 2013. Available at http://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/.

Kvalseth, T.O. (1985) “Cautionary note about R2.” *The American Statistician*: 39: 279-285.

Kuss, O. (2001) “A SAS/IML macro for goodness-of-fit testing in logistic regression models with sparse data.” Paper 265-26 presented at the SAS User’s Group International 26.

Kuss, O. (2002) “Global goodness-of-fit tests in logistic regression with sparse data.” *Statistics in Medicine* 21:3789–3801.

Liu, Y., P.I. Nelson and S.S. Yang (2012) “An omnibus lack of fit test in logistic regression with sparse data. ”*Statistical Methods & Applications* 21:437–452.

McFadden, D. (1974) “Conditional logit analysis of qualitative choice behavior.” Pp. 105-142 in P. Zarembka (ed.), *Frontiers in Econometrics*. Academic Press.

Maddala, G.S. (1983) *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.

McCullagh, P. (1985). “On the asymptotic distribution of Pearson’s statistics in linear exponential family models. *International Statistical Review* 53: 61–67.

Menard, S. (2000) “Coefficients of determination for multiple logistic regression analysis.” *The American Statistician* 54: 17-24.

Mittlbock, M. and M. Schemper (1996) “Explained variation in logistic regression.” *Statistics in Medicine* 15: 1987-1997.

Mroz, T.A. (1987) “The sensitiviy of an empirical model of married women's hours of work to economic and statistical assumptions.” *Econometrica* 55: 765-799.

Orme, C. (1988) “The calculation of the information matrix test for binary data models.” *The Manchester School* 54(4):370 –376.

Orme, C. (1990) “The small-sample performance of the information-matrix test.” *Journal of Econometrics* 46: 309- 331.

Osius, G., and Rojek, D. (1992) “Normal goodness-of-fit tests for multinomial models with large degrees-of-freedom.”*Journal of the American Statistical Association* 87: 1145–1152.

Nagelkerke, N.J.D. (1991) “A note on a general definition of the coefficient of determination.” *Biometrika* 78: 691-692.
Pigeon, J. G., and Heyse, J. F. (1999) “An improved goodness of fit test for probability prediction models. *Biometrical Journal* 41: 71–82.

Press, S.J. and S. Wilson (1978) “Choosing between logistic regression and discriminant analysis.” *Journal of the American Statistical Association* 73: 699-705.

Pulkstenis, E., and T. J. Robinson (2002) “Two goodness-of-fit tests for logistic regression models with continuous covariates.” *Statistics in Medicine* 21: 79–93.

Simpson, P., Hamer, R., ChanHee, J., Huang, B. E., Goel, R., Siegel, E., Dennis, R., and Bogle, M. 2004. “Assessing Model Fit and Finding a Fit Model.” *Proceedings of the SAS Global 2004 Conference*, Québec, CANADA.

Stukel, T. A. (1988) “Generalized logistic models.” *Journal of the American Statistical Association* 83: 426–431.

Tjur, T. (2009) “Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination.” *The American Statistician* 63: 366-372.

Tsiatis, A. A. (1980) “A note on a goodness-of-fit test for the logistic regression model.” *Biometrika*, 67: 250–251.

Unknown. “Evaluating Goodness of Fit.” *MathWorks*. 2016. Available at http://www.mathworks.com/help/curvefit/evaluating-goodness-of-fit.html.

Unknown. “STAT 504: Analysis of Discrete Data.” *PennState Eberly College of Science.* 2016. Available at https://onlinecourses.science.psu.edu/stat504/.

White H. (1982) “Maximum likelihood estimation of misspecified models.” *Econometrica* 50:1 –25.

Xie, X.J., J. Pendergast and W. Clarke (2008) “Increasing the power: A practical approach to goodness-of-fit test for logistic regression models with continuous predictors.” *Computational Statistics & Data Analysis* 52: 2703 – 2713.

ACKNOWLEDGMENTS

The author would like to acknowledge and thank Dr. Allison for his efforts and pulling together his resources and presenting a phenomenal presentation at SAS Global Forum 2014 (see references). His work and presentation was a huge help and guide for this paper. The author would also like to acknowledge her university, National University, for allowing the pursuit of individual interests while completing her Master’s degree this past couple of years. The author would also like to thank MWSUG for their acceptance of this abstract and help with polishing up the final paper/presentation.

CONTACT INFORMATION

Your comments, questions, and suggestions are valued and encouraged. Contact the author at:

Deanna Schreiber-Gregory, MS
Masters Graduate
Health and Life Science Analytics
National University
E-mail: d.n.schreibergregory@gmail.com