STATISTICAL GRAPHICS OF PEARSON RESIDUALS IN SURVEY LOGISTIC REGRESSION DIAGNOSIS

Stanley Weng, National Agricultural Statistics Service, U.S. Department of Agriculture
3251 Old Lee Hwy, Fairfax, VA 22030, 703-877-8000, sweng@nass.usda.gov

Abstract

Survey data logistic regression analysis, as computationally available in SAS SURVEY-LOGISTIC procedure, has been widely conducted in survey research practice. A set of diagnostic statistics in the procedure, borrowed from the logistic regression in generalized linear models, is used for model assessment. However, for survey data, the statistical underpinnings of these statistics may need to be reexamined. In practice, we have observed irregular behaviors of these diagnostic statistics, which make their established statistical criterion suspect. Their naïve use can be misleading.

This presentation reports our use of Pearson residuals normality graphs as graphical diagnostic statistics, to assess survey data logistic modeling, as in a recent NASS study of sampling frame coverage. Statistical graphics summarization may provide broader scope and more elaborated information than would be available through analytical summarization. The statistical graphs of Pearson residuals showed their diagnostic ability, and careful reading of the residual graphs may reveal delicate diagnostic information on modeling effects. We illustrate the statistical graphical modeling process with our analysis.

Key Words Pearson Residuals normality graph, Statistical graphical modeling

Introduction

The National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture conducts surveys to measure a variety of characteristics of U.S. agriculture, and conducts survey methodology research to improve its survey practice. One area of interest is a study of its multiple frame survey coverage. There are two sampling frames used by NASS in conducting its quinquennial census of agriculture. One is the Census Mailing List frame (CML), which despite the agency's best efforts to build it is ultimately incomplete. The other one is the Area Frame, which is a complete geo-spatial land frame, covering the entire U.S. Area Frame surveys routinely reveal farms not on the current Mailing List frame, and these are referred to as NML farms. Examining the characteristics of these NML farms is useful in targeting types of operations to be added to the List frame to improve the coverage. General characterization of likely NML farms was of interest. This study was conducted to address these issues. One analysis performed involved modeling the binary NML farm status on sampling frame and demographic and economic classification information via logistic regression.

Behavior of Survey Logistic Regression

Our survey data logistic regression was conducted using SAS Procedure SURVEYLOGISTIC. This procedure is adapted from the SAS procedure LOGISTIC for ordinary logistic regression, to incorporate survey sample design information into the estimation. The model diagnostic measures of SURVEYLOGISTIC are borrowed from the ordinary logistic regression procedure, mainly including the generalized (coefficient of determination) R-

square, the global null hypothesis test statistics, and Wald Chi-square for individual regression coefficients.

However, in practice we observed that survey logistic regression showed irregular behavior in various statistical aspects, which make model assessment complicated. The conventional statistical measures for model fitting, including R-square and Wald Chi-square, appeared no longer to keep their original statistical meaning and criterion. For example, such situations often appeared where the three tests for the global null hypothesis of the model all indicated strong significance, according to the conventional criterion of significance. But, individually, none of the predictors showed significance by the respective Wald Chi-square test. The generalized R-square was often very close to 1. Model estimates could be extremely unstable with changes in the predictor. In the forward modeling process, adding a predictor, seemingly increasing the predictive ability of the model by the conventional notion of modeling, could cause dramatic changes in the probability estimates, even resulting in extreme estimates, meaningless for use. Also, in some cases a promising model might be destroyed by one step of the forward selection process, so the forward/backward modeling process is not a monotone process, but more complicated. Simply conducting survey data logistic regression, according to the conventional regression modeling procedure, may be naïve.

We discussed our observations with SAS Technical Support statistical experts, regarding the statistical behavior of survey logistic regression, and they confirmed that there are limitations to the usefulness of the generalized coefficient of determination with survey data because of the presence of the sampling weights. The log likelihood values can become quite large for weighted data and this will result in the generalized R-square almost always being 1. Analysts should only use it as a means of comparing competing models (with one being a subset of the other) and not as a measure of overall fit of the model. One suggestion was to normalize the weights such that they sum to the sample size. Actually we found that in the SAS SURVEYLOGISTIC examples, the model assessment statistics are interpreted quite arbitrarily regarding their significance, without description of the criteria.

The irregular behavior of survey logistic regression changed our notion of and strategy for survey data statistical modeling considerably, and we began seeking an alternative applied approach for survey data logistic modeling diagnostics. A literature review of survey methodology didn't bring us established ways.

Pearson residuals

The conventional analytical approach for model assessment appeared to lack full capability in working with survey logistic regression. As a result, an alternative approach, a graphical approach, was considered. Graphical statistics have been used more and more in modern applied statistics. Analytical statistical summary is in compact and accurate form for analytical assessment, however, it may also have many information details suppressed when forming the summary. Graphical summary, instead, may provide broader scope and more elaborated information, including revelation of patterns.

Recall the statistical graph, the residuals vs. the predicted plot, as a basic diagnostic tool in linear regression model assessment. We were seeking an analogue in the context of binary data logistic modeling. The normality graph (normal probability plot) of Pearson residuals, assisted by the corresponding histogram, appeared to be such a statistical graph. Pearson residuals, as the components of the Pearson Chi-square statistic, are in the form

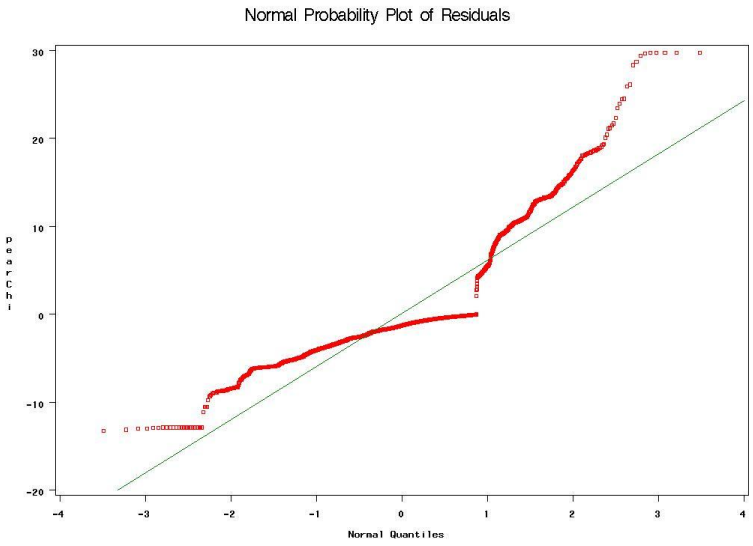
$$\sqrt{w_j}\,\frac{y_j-\hat{p}_j}{\sqrt{\hat{p}_j(1-\hat{p}_j)}},$$

where y_j is the binary response, w_j is the weight, and $\hat{p}_j$ is the model estimate of the probability. The normality graph provides a statistical graphical summary to assess the modeling. Informative reading of the graph reveals diagnostic information. In the following we illustrate a statistical graphical forward modeling process.

Statistical graphical modeling

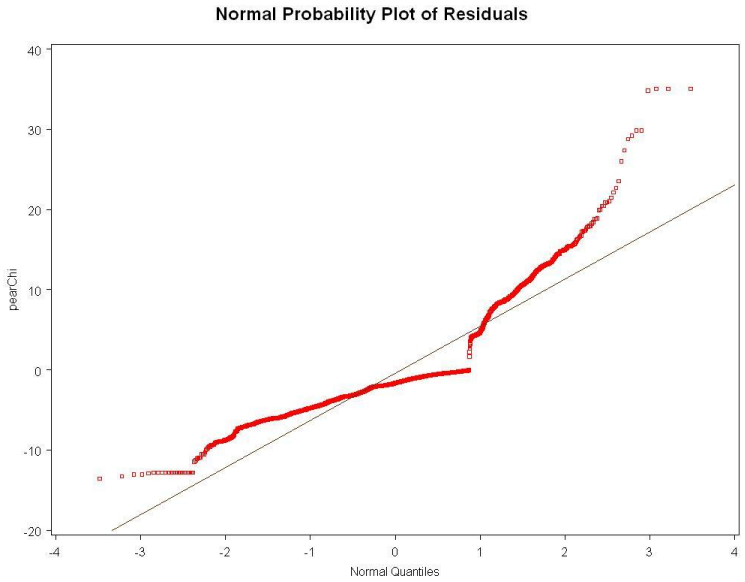
Below, four Pearson residuals normality graphs, from our survey data logistic modeling, illustrate the statistical graphical modeling process. Model 1, a base model, contains three continuous variables, all land measures, selected from all available numerical data. Figure 1 is the Pearson residuals normality graph of model 1.

Figure 1



Model 2 contains, in addition, two key indicators, one economic and one demographic. Figure 2 is the Pearson residuals normality graph of model 2.

Figure 2



Model 3 contains two additional demographic indicators (one for age group and one for occupation category). Figure 3-1 is the normality graph of Pearson residuals of model 3. We also present the corresponding histogram in Figure 3-2.

Figure 3-1

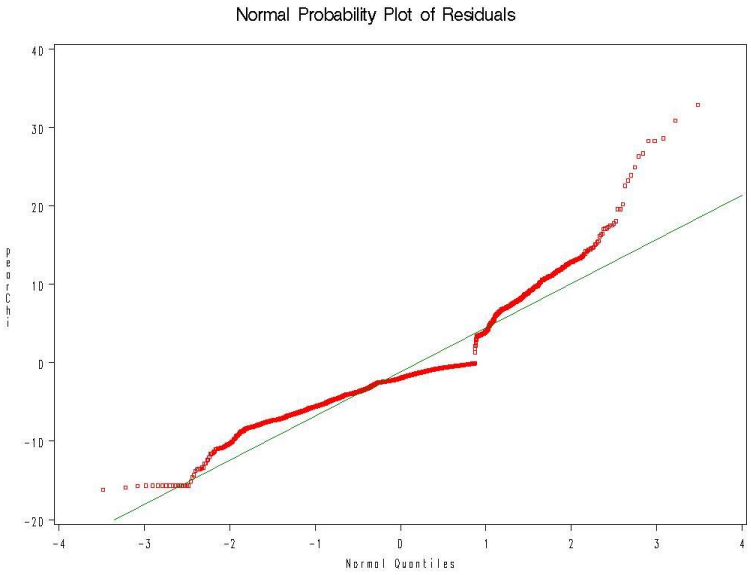
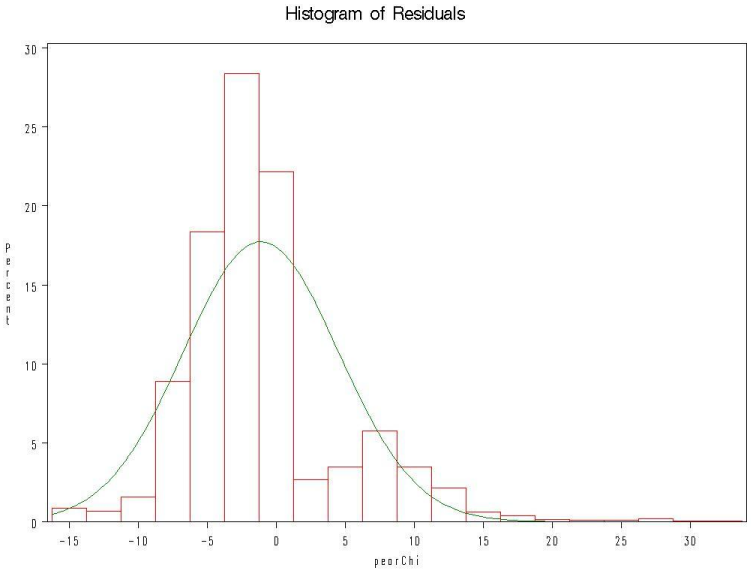


Figure 3-2



One additional demographic indicator for race classification was further identified, which significantly improved the shape of the Pearson residuals normality graph, as shown in Figure 4-1. The corresponding histogram is presented in Figure 4-2. This is model 4. The diagnostic graphs of model 4 appeared in fairly good shape. Visually, the assessment of the model fitting seemed adequate. The probability estimates generated by model 4 also appeared in reasonable shape. Thus our modeling concluded with model 4.

Figure 4-1

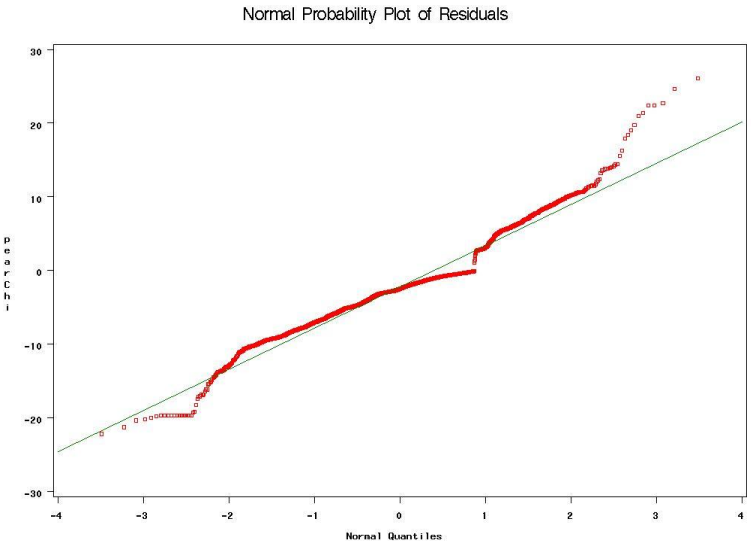
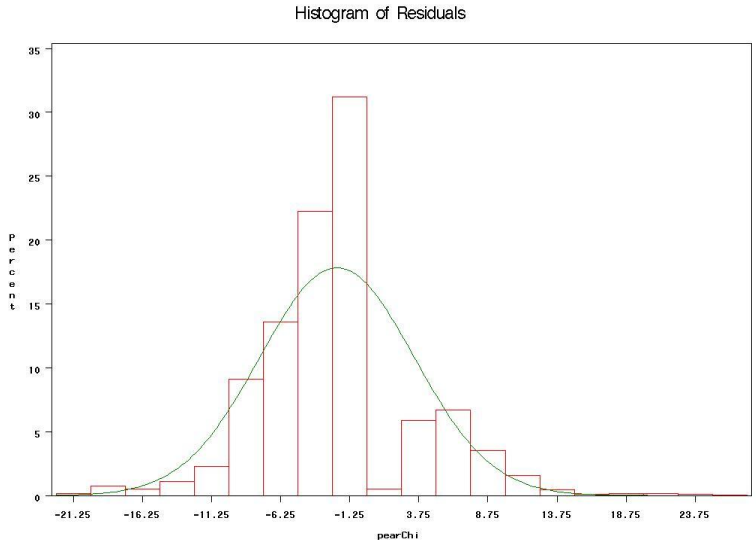


Figure 4-2



In conclusion, graphical modeling, as an applied approach, for survey data logistic modeling, is useful. Careful reading of the statistical graphs may obtain increased diagnostic information for statistical findings. For example, the graph can show the specific contribution to model fitting of an added predictor. Survey data logistic modeling can be very sensitive to change in the predictor. A binary predictor added to a model may considerably improve the fitting, as well as damage a model. In practice, fine tuning with a good sense of reading statistical graphs for model assessment may identify informative predictors and reach workable models.

Acknowledgement

The author wishes to thank Dale Atkinson, Dr. Phil Kott, and Scot Rumburg at NASS for helpful discussions on related issues and suggestions to improve the presentation of this paper.

References

Agresti, A. (2002), *Categorical Data Analysis*, Second Edition, Wiley-Interscience, New York.

Bishop, C.M. (2006), *Pattern Recognition and Machine Learning*, Springer, New York.

Morel, J. G. (1989) "Logistic Regression under Complex Survey Designs," *Survey Methodology*, 15, 203–223.

SAS Institute (2004), SAS/STAT 9.1 User's Guide.

Weisberg, S. (2005), *Applied Linear Regression*, Third Edition, Wiley, New York.