Paper 1477-2017

Analyzing Residuals in a PROC SURVEYLOGISTIC Model

Bogdan Gadidov, Herman E. Ray, Kennesaw State University

ABSTRACT

Data from an extensive survey conducted by the National Center for Education Statistics (NCES) is used for predicting qualified secondary school teachers across public schools in the U.S. The sample data includes socioeconomic data at the county level, which are used as predictors for hiring a qualified teacher. The resultant model is used to score other regions and is presented on a heat map of the U.S. SAS'® survey family of procedures such as SURVEYFREQ and SURVEYLOGISTIC are used in the analyses since the data involves replicate weights to account for the complex survey design. In looking at residuals from a logistic regression, since all the outcomes are binary, the residuals do not necessarily follow the normal distribution that is so often assumed in residual analysis. Furthermore, in dealing with survey data, the weights of the observations must be accounted for, as these affect the variance of the parameter estimates. To adjust for this, rather than looking at the difference in the observed and predicted values, the difference between the expected and actual counts is calculated by using the weights on each observation and the predicted probability from the logistic model for the observation. Two types of residuals are analyzed: Pearson and Deviance, with a slight adjustment to the Pearson residual formula to adjust for the replicate weights in the survey design.

INTRODUCTION

The NCES survey includes responses from secondary school teachers teaching a wide variety of STEM (science, technology, engineering, mathematics) courses across the U.S. Teachers whose primary teaching assignment is math courses represent roughly half of the teachers in the survey, and will be the focus of this paper. The purpose of the logistic model in this paper is to predict whether a given teacher is highly qualified to teach the math course to which they are assigned. To determine whether a teacher is qualified or not, several qualifications of the teacher are considered. The 3 criteria which are used in determining whether a teacher is qualified to teach a given subject are: whether the teacher has a degree in the subject they are teaching, a teaching certification in the subject they are teaching, or at least 5 years of experience in the subject they are teaching. A qualified teacher is defined as one who has all 3 of the previous qualifications. These criteria classify approximately 50% of the total teachers in the survey as being "highly qualified" to teach their respective math courses.

Traditionally, the LOGISTIC procedure can be used to create a logistic regression to model a binary variable (whether the teacher is qualified to teach math). However, in the presence of survey data, each response is given a corresponding weight. This weight reflects how many teachers in the population that the individual survey response represents. Without weights teachers from rural areas in the country may be underrepresented simply due to how the survey is constructed. Weights ensure that each response in the survey is representative of some proportion of the population of teachers. In addition to the weights on each observation, there are also 88 replicate weights. These replicate weights are used to simulate multiple survey samples, which in turn give a better estimate for the standard errors of parameters in the model (IPUMS USA). Without accounting for the replicate weights in the survey design, the standard errors of the parameter estimates for the model will likely be too small, resulting in smaller confidence intervals for estimates. The PROC SURVEYLOGISTIC can be used to account for this survey design, using the WEIGHT and REPWEIGHTS statements to input the final weight and replicate weights of each observation, respectively.

The model selection process included selecting from a wide variety of socioeconomic factors at the county and school district level. One key economic variable of interest is the percentage of students who are on a free or reduced lunch program at the school which the teacher teaches at. This variable is a binary indicator of whether the school has more than 40% of its students on a free or reduced lunch program. The final model contains 3 variables:

- Region – Region in which the school lies in, broken into 4 categories: Northeast, Southeast, Midwest, West
- Schsize – Binned variable for school size with 12 levels in which the first category is for schools with less than 50 students and the last category is for schools with more than 2000 students.
- Needs2 – Binary indicator of whether the school has more than 40% of its students on a free or reduced lunch program

The resultant model includes these 3 variables and the final model with parameter estimates is shown below. The next step in the assessment of the model's performance is to analyze the residuals. We will begin by calculating the Pearson residuals.

CALCULATING THE PEARSON RESIDUALS

In the LOGISTIC and GENMOD procedures, an STDRESCHI option is available on the OUTPUT statement to output the standardized Pearson residuals. As this option is not available in the SURVEYLOGSITIC procedure, we will attempt to calculate it in this paper. A Pearson residual is traditionally calculated by finding the raw difference between the observed and predicted value, and dividing by the standard deviation. In a logistic regression model with y being the binary outcome variable and $\hat{p}$ being the predicted outcome, the Pearson residual can be calculated with the following formula:

$$r = \frac{y - \hat{p}}{\sqrt{\hat{p}(1-\hat{p})}}$$

The denominator is the standard deviation of a binomial random variable with $n = 1$, and this calculation treats each observation as having a binary outcome. To account for the weights in the survey, this formula can be slightly altered to produce a similar result. In this formula, all observations which share the same covariate pattern (or the set of observations which all share the same values for their predictors) will be grouped together (Rodriguez, Germán). For instance, with the 3 variables included in the model, there are 2 levels to the needs2 variable, 4 levels to the region variable, and 12 levels to the schsize variable. In all, there are a possible 96 covariate patterns (attained by multiplying 2 by 4 by 12). These are the unique combinations of values for these variables. In this survey, only 92 of the possible 96 covariate patterns are realized in the data. The predicted probability $\hat{p}$ from the logistic regression is the same among each covariate pattern since each covariate pattern shares the same input values, so the numerator portion of the Pearson residual can be treated as the difference in the sum of the outcomes and the predicted outcomes in each covariate pattern. The denominator remains that of a binomial random variable with n being the sum of the weighted observations in the covariate pattern. The Pearson residual for weighted observations is:

$$r_i = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}}$$

where i represents the i^{th} covariate pattern (Breheny, Patrick). The standard error in the denominator includes the weights, but does not account for the complex survey design. The above formula is implemented in the following SAS code:

```
PROC SQL;
   create table ds2 as select
   sum(weight_final) as n,
   sum(response_variable*weight_final) as y,
   p
   from ds
   group by needs2, schsize, region;
quit;

data ds2;
set ds2;
```

```
    Pearson_Residual = (y - n*p) / ( sqrt(n*p*(1-p)));
run;
```

In the above SAS code, weight_final is the weight of each observation from the survey, response_variable is the binary indicator for whether the teacher was highly qualified, and p is the predicted value from the logistic regression. By summing weight_final and response_variable*weight_final over each covariate pattern (in the GROUP BY statement), the total number of weighted observations (n_i) and total number of outcomes (y_i), respectively, can be calculated for each covariate pattern. Once these are calculated in the SQL procedure, the Pearson residual can be found using the formula for weighted observations. The residuals from this model are shown in Figure 1 below. It is clear to see from this figure that the distribution of the residuals is slightly positively skewed, and the residuals range from -22.5 to 27, far outside the normally acceptable range of ± 3. The standard deviation which is used in the formula for calculating the Pearson residual is the likely cause, at it is not large enough and is causing these huge residuals.

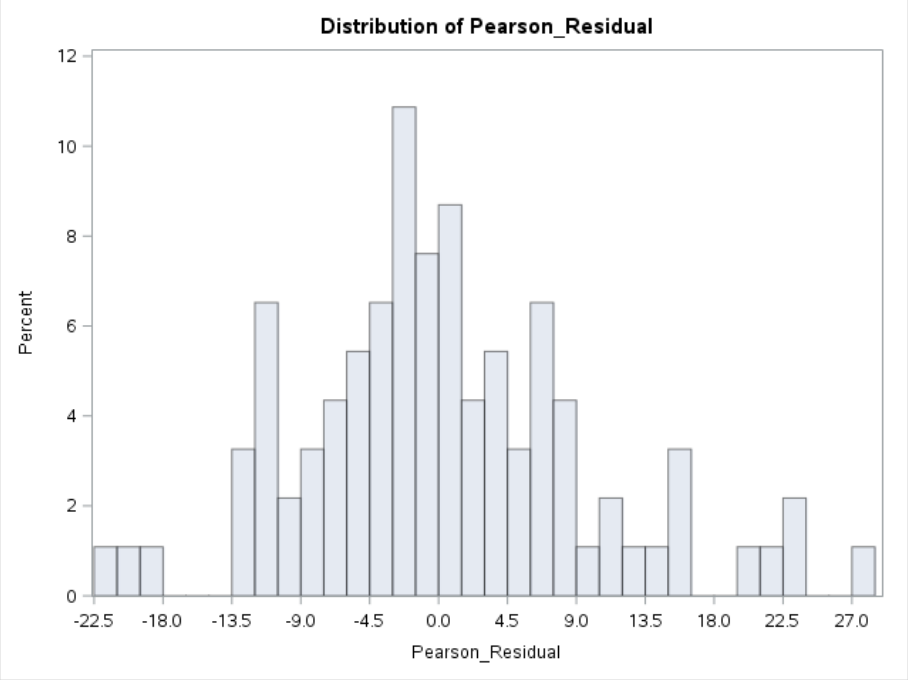

Figure 1. Pearson Residuals from SURVEYLOGISTIC Model

Table 1 shows some summary statistics for the above figure, including the mean and standard deviation of the Pearson residuals. This confirms that the above distribution has some positive skewness with large variations in the residuals.

Analysis Variable : Pearson_Residual			
Minimum	Maximum	Mean	Std Dev
-21.8607616	27.5848253	0.4006356	9.5359382

Table 1. Summary Statistics for Pearson Residuals in Figure 1

COMPARISON WITH PROC GENMOD

To get a better idea of how the residuals should behave, PROC GENMOD is run and the Pearson residuals are outputted from the resulting model. The GENMOD procedure does not include weights as the LOGISTIC and SURVEYLOGISTIC procedures do. Instead, the response variable in the MODEL statement is the proportion of events over the total number of trials n (commonly expressed as r/n). To perform a logistic regression, the DIST=BIN and LINK=LOGIT options can be specified on the MODEL

statement. The Pearson residuals can be output with the RESCHI option on the OUTPUT statement. The residuals from this model are shown below in Figure 2.

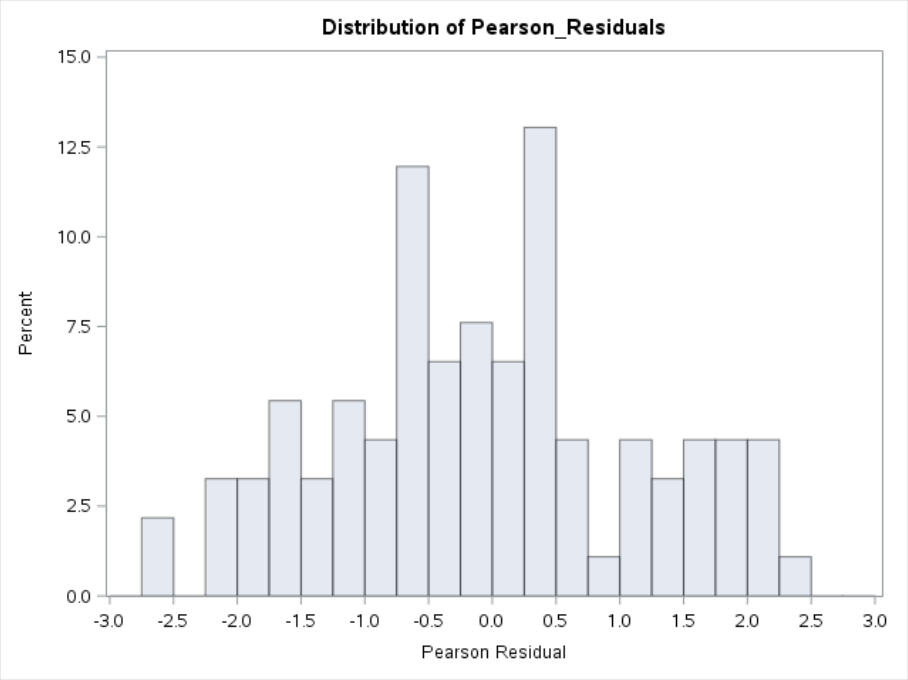

Figure 2. Pearson Residuals from PROC GENMOD Model

The residuals in Figure 2 shed some light on how the residuals should look when not accounting for the weights in the model. This again confirms that there should be an adjustment to the denominator of the traditional Pearson residual formula when working with survey replicate weights. This idea will be tested in the next section.

ADJUSTMENT TO THE PEARSON RESIDUAL FORMULA

To find a better method for calculating the standard error, the 88 replicate weights are used to derive the denominator portion of the Pearson residual equation. Since the replicate weights are intended to simulate multiple samples of this survey, these can be used to estimate the variance of the final weight used in the model. Since the replicate weights are no longer used in the REPWEIGHTS statement of the SURVEYLOGISTIC procedure, a PROC LOGISTIC can iteratively be run, each time using one of the replicate weights as the final weight in the WEIGHT statement of the LOGISTIC procedure. From each of these models, the predicted probabilities are stored to then be used for later calculations. The SAS code below shows a portion of the macro which performs the above procedure:

```
%do i=0 %to 88;
proc logistic data=&dataset;
class &classvars;
model response_variable (desc)= &vars;
weight weight&i;
output out=Predicted p=pred_&i;
run;

data Predicted;
set Predicted;
keep pred_&i;
run;
data &dataset;
merge &dataset predicted;
```

```
run;
%end;
```

PROC LOGISTIC is run once for each of the 88 replicate weights, and additionally once more for the final weight. The DO loop starts at i=0, assuming that the final weight is stored with a suffix of "0", so that in the first iteration the variable weight&0 resolves to the final weight variable (weight0). In each PROC LOGISTIC, the predicted probabilities from the resulting model are stored in a dataset "Predicted", with each column of predicted values numbered with a pred_&i suffix. All of these predicted values from the individual models are then merged together into one dataset containing the results of each LOGISTIC model.

The next step is to find the sum of the weights within each covariate pattern. This must be done for each of the 88 replicate weights, in addition to the final weight. An array total_weight is initialized to store the sum of the weights (weight0-weight88, where weight0 represents the final weight and weight1-weight88 represent the 88 replicate weights). A variable "covariate" is used in the dataset to identify the observations which belong to the same covariate pattern, and will be used in the BY statement to accumulate the weights. The first and last observations within a covariate pattern can be accessed using the first.covariate and last.covariate statements. At each occurrence of a new covariate pattern, the total_weight variable is initialized to 0, and then the corresponding weight is cumulatively summed within that covariate pattern. This is done for each of the 89 weights (again, for the final weight and 88 replicate weights). The weighted_sum variable is calculated in this step, and used in the Pearson residual calculation (weighted_sum is equivalent to y_i of the Pearson residual for weighted observations formula). The following code shows the above described process:

```
data ds2;
set &dataset;
   array total_weight {*} t0-t88;
   array weight {*} weight0-weight88;
   retain t0-t88;
   retain weighted_sum;
         by covariate;
         if first.covariate then do i=1 to 89;
            total_weight[i] = 0;
            weighted_sum = 0;
         end;
         do i=1 to 89;
            total_weight[i] = total_weight[i] + weight[i];
         end;
   weighted_sum = weighted_sum + response_variable*weight0;
   if last.covariate then output;
run;
```

The dataset created above now contains one row for each covariate pattern, and has columns t0-t88 which contain the sum of the weights weight0-weight88. The next step is to calculate the variability in the predicted event counts between the final weight group and those of the 88 replicate weight models. The summation of the squared differences between the final weight group and the 88 replicate weight groups, divided by 88, can be used to approximate the variance of the final weight in the survey. This value can then be used as the denominator of the Pearson residual. The formula below illustrates this, where N_{WF} represents the predicted event counts using the final weight in the LOGISTIC procedure, and N_i represents the predicted event counts using replicate weight i in the LOGISTIC procedure, and r represents the number of replicate weights.

$$\sum_{i=1}^{r} \frac{(N_{WF} - N_i)^2}{r}$$

N_{WF} from above is the product of the sum of the final weights (t0) and the predicted probabilities from its respective model (pred_0). Each of the N_i is the corresponding t_i (sum of replicate weights i by each

covariate pattern) multiplied by the corresponding pred_i from the model. The SAS code below calculates this variance:

```
data ds2;
set ds2;
   array w {88} t1-t88;
   array p {88} pred_1-pred_88;
   array d {88} d1-d88;
   do i = 1 to 88;
      d{i} = (( t0*pred_0 - w{i}*p{i} )**2)/88;
   end;
   variance=sum(of d1-d88);
run;
```

Finally, the Pearson residual is calculated using a slight variant to the original Pearson residuals which were calculated previously. The numerator remains the same, the difference between the observed event counts and expected event counts within each covariate pattern. Keep in mind that weighted_sum resembles y_i from the previous formula given, and t0*pred_0 similarly resembles $n_i\hat{p}_i$, the predicted event counts given by the model. The adjustment is to the denominator, which rather than using $\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}$ uses the variance calculated in the above SAS code. The new calculation can be seen in SAS code below:

```
data ds2;
set ds2;
   Pearson_Residual = ( weighted_sum - t0*pred_0 ) / ( sqrt(variance) );
run;
```

This adjustment yields residuals which are shown in Figure 3. These residuals still exhibit some slight positive skewness, with one clear outlier. However, all the residuals excluding the outlier are now well within the standard range of ± 3, and the distribution appears roughly normally distributed.

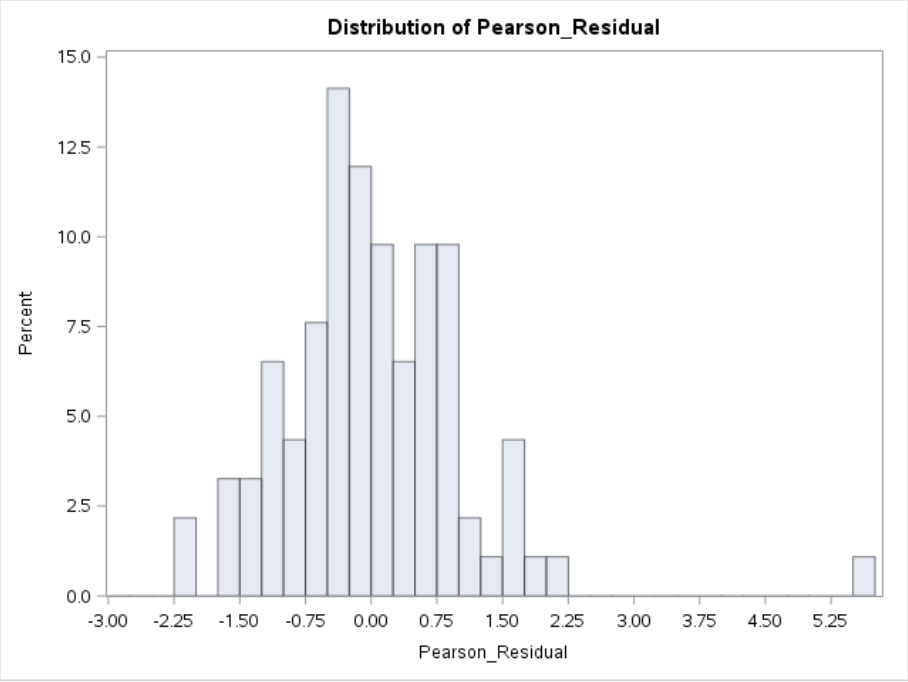

Figure 3. Distribution of New Pearson Residuals

The mean and standard deviation of the Pearson residuals are shown below in Table 2 to confirm what is seen above. The mean is very close to 0, with a standard deviation of approximately 1. This is what we

expect to see in a residual analysis, confirming that the process described in this section has corrected for the complex survey design.

Analysis Variable : Pearson_Residual			
Minimum	Maximum	Mean	Std Dev
-2.1161495	5.6591494	0.0294528	1.0753324

Table 2. Summary Statistics for Pearson Residuals in Figure 3

DEVIANCE RESIDUALS

Another type of residual diagnostic for logistic models is the deviance residual. This residual is computed using the logarithm of the likelihood ratio in both the event and non-event groups of the model. The formula below shows the calculation of the deviance residual, which is calculated at the covariate pattern level:

$$d_i = sign(Pearson\ Residual)\sqrt{2[y_i \log\left(\frac{y_i}{u_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - u_i}\right)]}$$

where y_i is the sum of the event counts in covariate pattern i, u_i is the sum of the predicted event counts in covariate pattern i, and n_i is the sum of the final weight in covariate pattern i. These three values have been calculated in previous SAS code seen in this paper. Additionally, the deviance residual takes on the sign of the Pearson residual (i.e. an observation in which the observed event counts are less than the predicted event counts, then the deviance residual would be negative). These deviance residuals are calculated and shown in Figure 4.

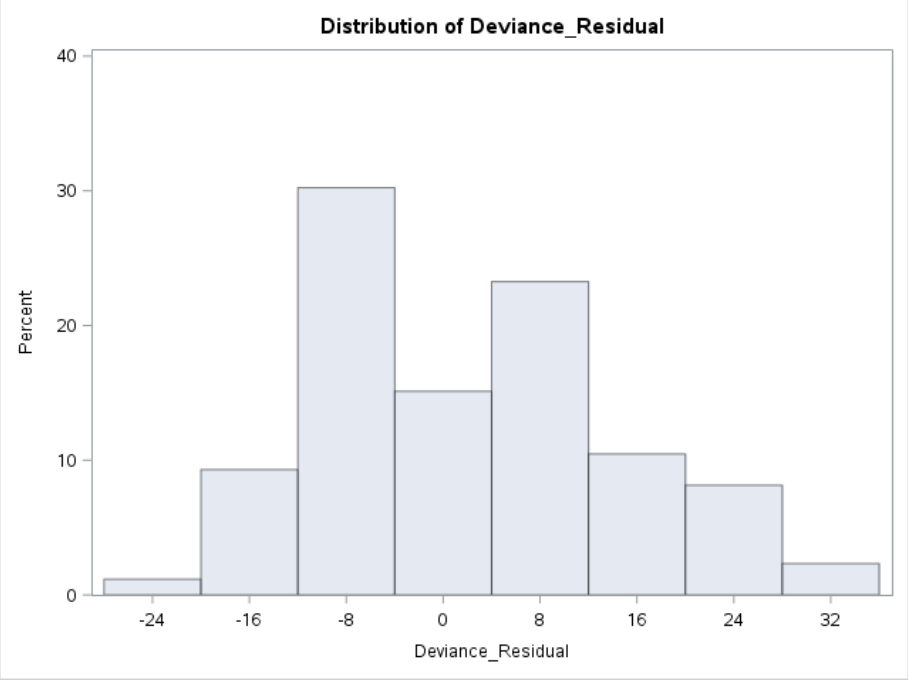

Figure 4. Deviance Residuals

Deviance residuals outside of 2 in magnitude are generally considered to indicate lack of fit. However, much like in the case with the original Pearson residuals which were calculated, there must be an adjustment to the formula of the deviance residual which can account for the complex design of the survey. This piece is still currently being worked on.

CONCLUSION

We began by investigating the Pearson residuals in a PROC SURVEYLOGISTIC model with 88 replicate weights. The standard formula for calculating these residuals led to a distribution which resembled a very heavy tailed normal distribution, as the magnitude of the residuals reached 27. By using the 88 replicate weights to estimate the variance in the final weight, the denominator of the Pearson residual was adjusted to reflect the complex survey design. By doing so, the Pearson residuals appeared to have a normal distribution with only one outlier. Additionally, deviance residuals are explored, but more work needs to be done to account for the complex survey design. Future work includes other residuals and goodness of fit statistics such as the studentized residuals and DFBETAS, which will require the hat (H) matrix.

REFERENCES

Breheny, Patrick. "Logistic regression: Diagnostics." Accessed on January 16, 2017. http://web.as.uky.edu/statistics/users/pbreheny/760/S11/notes/4-12.pdf.

IPUMS USA. "Replicate Weights in the American Community Survey/Puerto Rican Community Survey." Accessed on February 10, 2017. https://usa.ipums.org/usa/repwt.shtml.

Rodriguez, Germán. "Regression Diagnostics for Binary Data." 2017. *Princeton University*. Accessed on January 18, 2017. http://data.princeton.edu/wws509/stata/c3s8.html.

CONTACT INFORMATION

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

Bogdan Gadidov
Kennesaw State University
bgadidov@kennesaw.edu