

The Thorn in My Side!! Logistic Regression Continuous Variables that Violate the Assumption of Linearity on the Log-odds (Logit) Scale: How to Identify and What to Do?

Janet Grubber, Durham Veterans Affairs Health Care System; Cynthia Coffman, Durham Veterans Affairs Health Care System, Duke University Medical Center

ABSTRACT

I find myself crossing my fingers when I check continuous variables that are to be included in binary logistic regression models for linearity on the log-odds scale"Please, please...let them be linear"! Recently, my pleading did not work. Though using the Box-Tidwell method for checking for linearity seems to have fallen out of favor, I used it anyway. Based on Box-Tidwell results, one of the independent continuous variables of great interest to my study team appeared to be linearly associated with the log-odds of the dependent variable; however, it was unclear whether this classification was accurate. Many other independent continuous variables to be used in the model were, based on Box-Tidwell results, not linearly associated with the log-odds of the dependent variable – but - what good was the knowledge of lack of linearity when I needed to understand the shape of the non-linearity to appropriately use the variables in a model and understand their associations with the dependent variable?

This paper works through the sequence of steps that I ultimately used via SAS® software to understand the nature of (what turned out to be!) the non-linear relationship of the independent variable of interest with a binary dependent variable in a logistic regression model:

1. Identifying lack of linearity on the log-odds scale (Box-Tidwell, use of %PSPLINET macro (Frank Harrell) to plot the association, use of restricted cubic splines in SAS PROC LOGISTIC procedure)
2. Handling a continuous variable in a logistic regression model once a lack of linearity is detected (creation of multiple "dummy-like" continuous variables to represent the independent variable of interest, use of restricted cubic splines in PROC LOGISTIC procedure)

INTRODUCTION

I have used SAS to run logistic regression models for over twenty-five years. During this time, I have always cringed when one or more of the variables to be included in a model is continuous because I have never truly felt comfortable 1) identifying whether the continuous variable(s) are linearly related to the log-odds of the dependent variable and 2) figuring out what to do with them if they are not.

Most of the time, my solution has been to categorize these types of variables so that I no longer have to worry about the assumption of linearity on the log-odds scale. Recently, however, there were too many independent variables relative to the number of events in the dependent variable (Peduzzi) to allow me to categorize the continuous variables into as many categories as needed to well-describe the continuous variables in a logistic regression model...and so began my journey to finally figure out what to do with continuous variables. In particular, there was one continuous variable, a comorbidity score called Gagne (Gagne, 2011) that my study team wanted to understand in terms of the increased odds of 90 day hospitalization for a one unit change in the Gagne Score.

This paper describes the steps that I took to identify and understand the shape of the association between the log-odds of the 90 day hospitalization rate and the Gagne Score as well as the method that I used to incorporate this variable into the model. My hope is that this information will help others who find themselves cringing when they need to figure out how to handle continuous variables in logistic regression models, especially in situations in which the association between the independent continuous variable and the log-odds of the dependent variable is not linear, and when there is a need or desire to be able to interpret that association in a way that is understandable to the lay person.

This paper assumes that the audience has a working knowledge of logistic regression.

WHAT ARE STANDARD ASSUMPTIONS FOR RUNNING (BINARY) LOGISTIC REGRESSIONS?

The standard assumptions for running binary logistic regression analyses are:

1. Patients are randomly selected from the population.
2. Individual observations (i.e. rows of data representing, for example, people) are independent of one another.
3. There is little or no multicollinearity between independent variables. (Note that in some situations multicollinearity between independent variables can be safely ignored (Allison)).
4. There is a linear association between the log of the odds of the outcome/dependent variable and each continuous independent variable of interest.

This paper will focus solely on assumption #4, above. In this paper, ordinal variables, such as the Gagne Score, which ranges from -2 to 13 with only integer values, will be treated as a continuous variables.

WHAT DOES IT MEAN FOR AN ASSOCIATION BETWEEN AN INDEPENDENT AND DEPENDENT VARIABLE TO BE “LINEAR ON THE LOG-ODDS SCALE”?

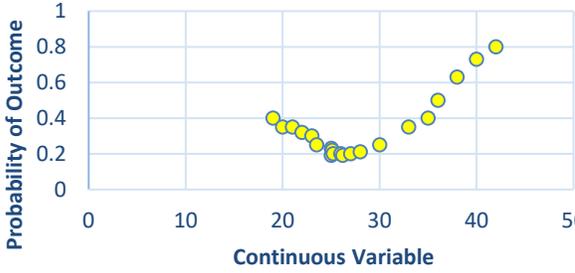
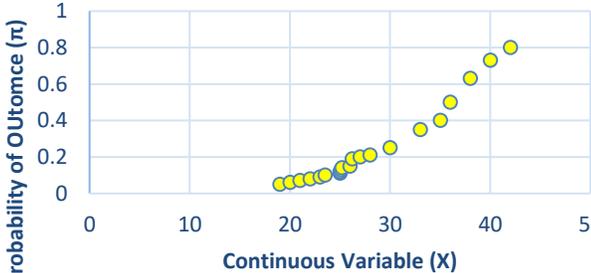
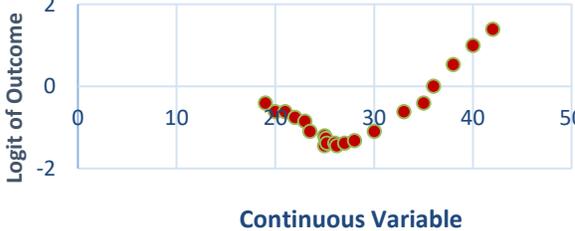
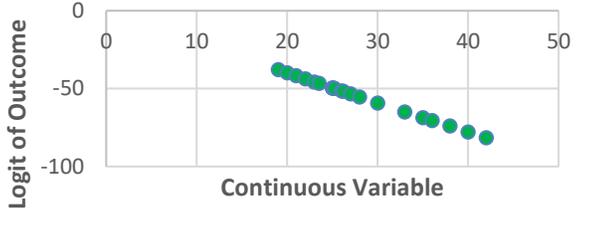
A continuous independent variable is “linear on the log-odds scale” if the relationship between it and the natural log of the odds (also known as the “logit”) of the dependent variable is linear. In our example, this would mean that the natural log of the odds of being hospitalized in the upcoming 90 days has a linear association with the independent variable, Gagne Score.

The log-odds (or logit) is represented by the logistic regression model below where π =probability of the outcome of interest, X is the continuous independent variable, β_0 is the intercept of the line, and β_1 is the slope of the line.

$$\text{Logit}(\pi) = \log_e(\pi / (1 - \pi)) = \beta_0 + \beta_1 X$$

This logistic regression model works well when the assumptions listed above are met. If the assumptions are not met, however, the logit of the predicted probability may not accurately reflect the reality of the data. In this paper we focus solely on the assumption of the linear relationship.

Below are plots of the probability of the binary outcome, and the log-odds of the outcome vs. each of 2 different hypothetical continuous variables. Figures 1a and 1b plot the association between the continuous variables and the *probability* of the outcome. Neither are linearly associated – and this is not a problem because they are not required to be. However, the plot of the *log-odds* of the outcome must look at least somewhat like a straight line for the assumption of linearity on the log-odds scale to be met. Figure 1c is an example of a non-linear association between the log-odds of the outcome, while Figure 1d is an example of a linear association.

Example of Continuous Variable That <i>Is Not Linearly Associated</i> with the Log-odds of the Dependent Outcome Variable	Example of Continuous Variable That <i>Is Linearly Associated</i> with the Log-odds of the Dependent Outcome Variable
<p>Figures 1a and 1b. The association between the probability (π) of the binary outcome and the continuous independent variable is not, and does NOT have to be, linear for the assumption of a linear relationship between the log-odds of the binary outcome variable and the continuous independent variable to be met. The colors of the dots are yellow, as in “CAUTION: THIS GRAPH IS NOT ENOUGH”...it still needs to be determined whether the assumption of linearity is met by graphing the log-odds of the dependent variable against the continuous variable.</p>	
<p style="text-align: center;">1a. Probability of Outcome (π) vs. Continuous Variable (X)</p> 	<p style="text-align: center;">1b. Probability of Outcome (π) vs. Continuous Variable (X)</p> 
<p>Figures 1c. and 1d. The association between the natural log-odds of the outcome (Logit (π)) and the continuous independent variable should be linear for the assumption of linearity to be met. Figure 1c is an example in which the log-odds of the dependent variable is not linearly associated with the continuous variable (i.e. red dots for “STOP”...don’t use this variable as a continuous variable in a logistic model. Figure 1d is an example in which the assumption of a linear relationship between the log-odds of the binary dependent variable and the continuous independent variable is met (green dots for “Things are good; the assumption appears to be met. This continuous variable may be included in the logistic model as a continuous variable.”)</p>	
<p style="text-align: center;">1c. Loge-Odds (Logit) of Outcome vs. Continuous Variable</p> 	<p style="text-align: center;">1d. Loge-Odds (Logit) of Outcome vs. Continuous Variable (X)</p> 

HOW DO I CHECK FOR LINEARITY ON THE LOG-ODDS SCALE?

There are various methods for checking to see if continuous variables are linearly associated with the logit of the dependent variable; however, not all methods yield the same result. This section describes several methods for checking for linearity on the log-odds scale. In the Conclusion section of this paper, we provide recommendations that may be helpful in situations similar to ours.

1. BOX-TIDWELL? – NOT RECOMMENDED!!

In the past I have used Box-Tidwell to examine linearity, so I decided to start there. The heart of this analysis is to include the continuous variable of interest (e.g. *gagne* in the example below) in a logistic regression model as an independent variable and the outcome variable of interest as the binary dependent variable. In addition, one adds, as a second independent variable, an interaction term between the continuous variable and the log of the continuous variable (e.g. *gagne*log_gagne* in the example below). A p-value of ≤ 0.05 for the interaction term indicates that the association between the log-odds of the dependent variable and the continuous variable is not linear.

Here is source code for creating the “logged” variable that is used in the interaction term and the logistic code for checking for linearity using the Box-Tidwell method:

```
data sesug;
  set imputed (keep=gagne hosp_90d);
  gagne=gagne+2.000000001; *shifting gagne score because you can't
                           take the log of a negative number or 0;
  log_gagne=log(gagne);
run;

*box-tidwell;
proc logistic data=sesug descending;
  model hosp_90d=gagne gagne*log_gagne;
run;
```

Hopefully Helpful Hints: In SAS, you must create the log variable in a data step prior to using it in the interaction term (e.g. $\log_gagne=\log(gagne)$).

Below is an excerpt from the Box-Tidwell SAS output used to check for a linear association between the logit of the binary outcome (90 day hospitalization) and the continuous independent variable (Gagne Score).

Output 1.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.3314	0.1222	743.3441	<.0001
gagne	1	0.0656	0.0690	0.9036	0.3418
gagne*log_gagne	1	0.0413	0.0262	2.4806	0.1153

The p-value for the interaction term (highlighted in yellow) is not ≤ 0.05 so, based on this analysis, we initially concluded that the the log-odds of hospitalization in 90 days could be assumed to be linearly associated with the *gagne* variable and used in the logistic regression as a continuous variable. HOWEVER, this is not the end of the story...please read on!

2. RESTRICTED CUBIC SPLINES

We had many continuous variables, not just the Gagne Score, to examine for a linear association with the log-odds of the outcome variable; therefore we examined them all and found that many, based on the Box-Tidwell method, did not meet the linearity assumption. As a result, we took a next step to figure out the shape of the relationship between the log-odds of the outcome variable and the continuous variables. (We had not set a priori cutpoints for categorizing the continuous variables in case the relationship was not linear and, in addition, we did not have enough degrees of freedom in our logistic model to create lots of additional categories for the many “continuous” variables (Peduzzi)).

In this section we discuss our next step, which was using restricted cubic splines. In particular, we’ll describe the %PSPLINET macro (Harrell) and the “native” implementation of restricted cubic splines in SAS.

Restricted cubic splines, as explained very understandably by Ruth Croxford (2016), are transformations of continuous predictors. Her article includes a great picture of a draftman’s spline and describes splines in the following terms: “The range of values of the predictor is subdivided using a set of knots. Separate regression lines or curves are fit between the knots.” She defines a spline function as “a set of smoothly joined piecewise polynomials” and reports that cubic splines are the most often used splines in regression models because they’re the form of polynomial that takes up the fewest degrees of freedom in a model while still allowing for an inflection in a curve. Furthermore, Croxford states that the tails prior to the first and after the last knot aren’t “well behaved” and that the use of cubic splines avoids this problem by constraining the tails to be straight lines.

A. %PSPLINET MACRO (HARRELL)

Frank Harrell’s SAS macro, %PSPLINET (Plot Spline Transformation) came to the rescue! It plots restricted cubic spline transformations along with their 95% confidence intervals for single predictors in binary logistic models. This macro calls a second macro, %DASPLINE, which automatically computes knots and creates the individual spline effects. %PSPLINET is based on the Stone and Koo (1985) additive spline transformation of continuous independent variables.

When we ran this macro (for all our continuous variables), we discovered that the Gagne Score was actually NOT linearly associated with the log-odds of 90 day hospitalization. Here is the code we used to run the %PSPLINET macro (yellow-highlights indicate our specifications; most other specifications are defaults as described in the box below):

```
*psplinet;  
%PSPLINET(x=gagne, y=hosp_90d, model=LOGISTIC, range=, event=, k=1, nk=4, knot=,  
plot=2, outer=, second=, adj=, saxis=, xaxis=, testlin=1, plotprob=0, groups=,  
paxis=, short=0, nopst=0, print=0, PLOTDATA=PUNCH, COMMTYPE=1, FILEMODE=MOD,  
data=sesug);
```

Hopefully Helpful Hints: Please refer to Frank Harrell's website (Harrell) for SAS code for his %PSPLINET macro as well as his detailed description of each element of the macro. Here is an excerpt from the documentation of his macro:

Usage:

```
%PSPLINET(X,Y,MODEL=LOGISTIC(default) or COX (default if EVENT given),
          RANGE= low TO high BY increment range for evaluating X
              (default=range in data, default increment=1 if RANGE
              given without increment),
          EVENT=event indicator for PROC PHGLM if COX model used,
          K=max Y value for PROC LOGIST if ordinal logistic model used,
          NK=number of knots to use if KNOT omitted (3,4,5, default=4),
          OUTER=outer percentile for 1st knot if omit KNOT (DASPLINE),
          SECOND=percentile for 2nd knot if omit KNOT (DASPLINE),
          KNOT= knot points (computed by DASPLINE by default),
          PLOT=1 (PROC PLOT) 2 (PROC GPLOT, default)
              3 (PROC PLOT and PROC GPLOT),
              4 (no graphics but produce text file with coordinates
                  suitable for Harvard Graphics or similar software.
                  Uses ^ as field delimiter.),
          PLOTPROB to plot probability est. for logistic if ADJ omitted,
          TESTLIN=1 to fit linear and non-linear logistic model to allow
                  computation or LR statistic for linearity (for Cox, always
                  computes Wald test of linearity),
          GROUPS= number of quantile groups (e.g. 10 for deciles),
          PRINT to print estimates from EMPTREND if GROUPS is used,
          ADJ=list of adjustment variables
          DATA=input dataset (default=_LAST_),
          SAXIS=low to high by inc y-axis specification for plotting
                  spline transformation,
          PAXIS=low to high by inc y-axis spec. for plotting
                  probabilities if PLOTPROB is used,
          XAXIS=x-axis specification,
          SHORT=to suppress confidence intervals and knots on
                  SAS/GRAPH for PLOTPROB output if PLOT>1,
          NOPST=to suppress plotting of spline transformation on
                  graphics device if PLOTPROB is given and PLOT>1,
          PLOTDATA=file name for PLOT=4. Default is PUNCH.
          FILEMODE=output mode for PLOTDATA file. Default is MOD.
                  May also be NEW - a new file is started.
          COMMTYPE=type of software being used to download PLOTDATA
                  1=standard text file (default)
                  2=Barr/Hasp var. length records with | as rec. delim.);
```

Below is output from the %PSPLINET macro along with associated descriptions.

Output 2a.

Analysis of Maximum Likelihood Estimates

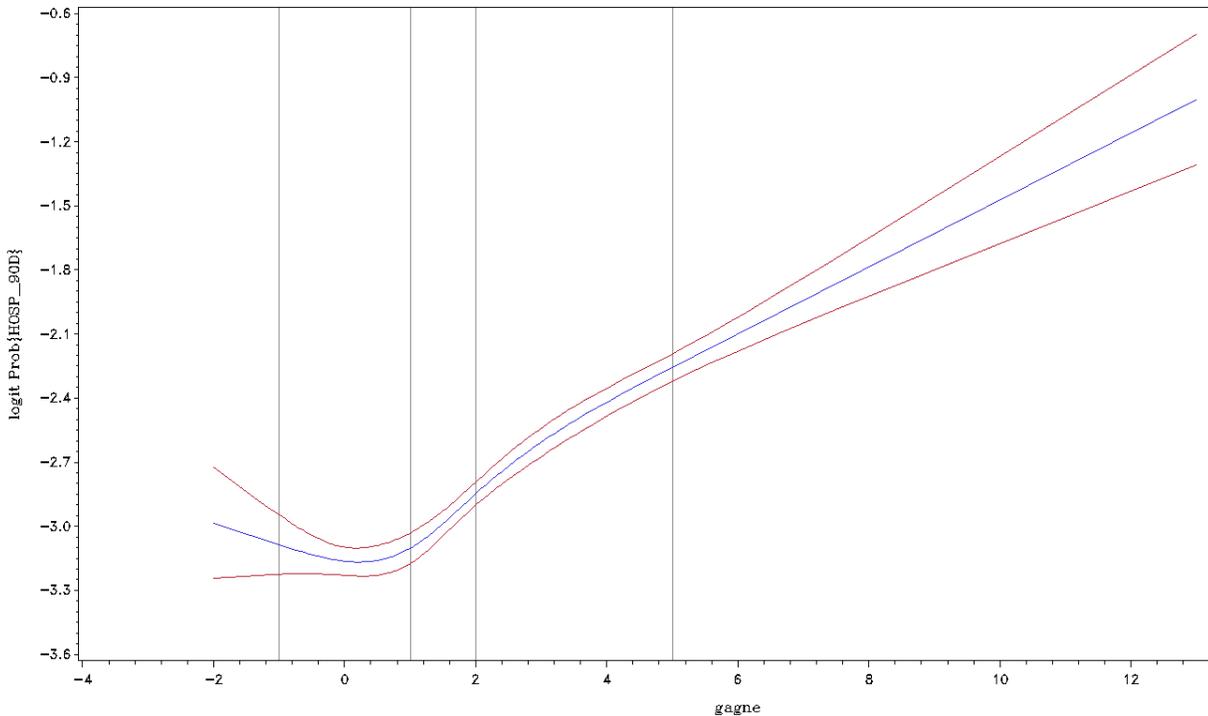
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.1872	0.0351	8234.6903	<.0001
gagne	1	-0.1017	0.0654	2.4190	0.1199
GAGNE1	1	0.8344	0.2056	16.4686	<.0001
GAGNE2	1	-2.9791	0.7693	14.9966	0.0001

GAGNE1 and GAGNE2 are spline effect variables automatically created by the macro. A p-value of ≤ 0.05 for a spline effect indicates that the log-odds of the outcome is not linearly associated with the continuous independent variable. In this case, the log-odds of the probability of hospitalization in 90 days is not linearly associated with the Gagne Score.

The plot shown in Output 2b, below, is also generated by the macro and describes the shape of the association between the Gagne Score and the log-odds of the probability of hospitalization in 90 days.

Output 2b.

Estimated Spline Transformation and 95% C.I.



This plot visualizes the lack of linearity which occurs around Gagne ~ 0 where the slope of the line changes from negative to positive.

“NATIVE” IMPLEMENTATION OF RESTRICTED CUBIC SPLINES IN SAS

Another option for testing for lack of linearity between the log-odds of the dependent variables and a continuous independent variable is by including an EFFECT statement within PROC LOGISTIC and specifying that spline effect variables should be created. There are various options for specifying the way in which splines are created (Wicklin). Below is the code that we used:

```
*effect statement using spline in proc logistic;
proc logistic data=sesug descending;
  effect gagnespl=spline(gagne/details naturalcubic
  basis=tpf(noint) knotmethod=percentiles(4));
  model hosp_90d=gagnespl;
run;
```

Hopefully Helpful Hints: This native implementation of restricted cubic splines within SAS does not automatically create graphs for visualizing the shape of the relationship between the independent variable and the log-odds of the dependent variable as does the %PSPLINET macro; however, predicted probabilities can be output from PROC LOGISTIC, transformed to logits, and graphed on the Y-axis with the independent variable graphed on the X axis to get a similar plot.

The below output is produced by using the SAS code above to run a logistic regression model using restricted cubic splines to test for a linear association between Gagne Score and the logit of 90 day hospitalization.

Output 3.

Testing Global Null Hypothesis: BETA=0					
Test		Chi-Square	DF	Pr > ChiSq	
Likelihood Ratio		360.8436	3	<.0001	
Score		432.8312	3	<.0001	
Wald		401.5970	3	<.0001	
Type 3 Analysis of Effects					
Effect	DF	Mald Chi-Square		Pr > ChiSq	
gagnespl	3	401.5970		<.0001	
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Mald Chi-Square	Pr > ChiSq
Intercept	1	-3.1481	0.0339	8619.2156	<.0001
gagnespl 1	1	-0.0256	0.0540	0.2240	0.6360
gagnespl 2	1	0.2059	0.0623	10.9357	0.0009
gagnespl 3	1	-0.3464	0.1111	9.7228	0.0018

Gagnespl 2 and gagnespl 3 are spline effect variables created by SAS as a result of the inclusion of the EFFECT statement in the logistic regression code above. Similarly to the results from %PSPLINET, the yellow-highlighted p-values of ≤ 0.05 for these spline effects indicate that the log-odds of the outcome is not linearly associated with the continuous variable. More specifically, these analyses indicate the lack of a linear relationship between the Gagne Score and the log-odds of hospitalization within 90 days.

WHAT DO I DO IF THE RELATIONSHIP BETWEEN THE CONTINUOUS INDEPENDENT VARIABLE AND THE LOG-ODDS OF THE DEPENDENT VARIABLE IS NOT LINEAR?

There were several options we considered for addressing the non-linear relationship between the Gagne Score and logit of the 90 day hospitalization, including: categorizing the Gagne score, creating “discontinuous continuous” (i.e. piecewise linear) variables, and restricted cubic splines. These methods are described below.

1. CATEGORIZE THE CONTINUOUS VARIABLE

The continuous variable can be categorized based on a priori cutpoints informed by relevant literature or by clinician input. Another option is to use as cutpoints the values of the continuous variable from the points at which the slope of the relationship between the log-odds of the dependent variable and the independent variable changes. These cutpoints can be visualized by graphing the association using the %PSPLINET macro described in the previous section. We chose not to categorize the Gagne Score because we wanted to calculate the association between the odds of hospitalization in 90 days and a one unit increase in the Gagne Score; categorizing did not allow for this estimate.

2. CREATE “DISCONTINUOUS CONTINUOUS” VARIABLES

Methods for modeling discontinuous or nonlinear changes described by Singer and Willett (2003; pgs. 189-242) may also be used. They describe methods for modeling discontinuities in elevation, slope, or both in the relationship between dependent and independent variables. We used the following code based on their guidance to model the discontinuity in the slope at Gagne =0 (see Output 2b) in our study:

```
data sesug;
  set imputed (keep=gagne hosp_90d);

  *making 2 variables to include gagne in model as two separate continuous
  variables to account for lack of linearity;
  if gagne=. then do; gagne_shifted=.; gagne_post_shift=.; end;
  else
  do;
    gagne_shifted=gagne+2;
    gagne_post_shift=gagne_shifted-2;
    if gagne_shifted le 1 then gagne_post_shift=0;
  end;
run;

*checking coding;
proc freq data=sesug;
  tables gagne*gagne_shifted*gagne_post_shift/missing list; run;

*running proc logistic with the 2 new continuous variables for gagne score
and estimating appropriate odds ratios for each part of the curve;

proc logistic data=sesug desc;
  ods output parameterestimates=parmsdsn covb=covbdsn
             estimates=gagne_est;
  model hosp_90d=gagne_shifted gagne_post_shift/covb cl;
  estimate "gagne_le0" gagne_shifted 1/exp cl;
  estimate "gagne_gt0" gagne_shifted 1 gagne_post_shift 1/exp cl;
run;
```

A crosstabulation of the original *gagne* variable and the two newly created *gagne_shifted* and *gagne_post_shift* variables are presented in Output 4 below. The *gagne_shifted* variable simply adds a constant (2) to the original *gagne* variable to simplify the recoding of the original variable into 2 variables. This new *gagne_shifted* variable represents the slope of the original *gagne* variable (i.e. the negatively sloped part of the curve displayed in Output 2b). The intercept of this line is now shifted by the constant of 2; however, the location of the intercept is not important in that we want to estimate an odds ratio for a 1 unit change in the slope, not an odds ratio for one particular response level.

The *gagne_post_shift* variable represents the “turning on” of the positive slope at the original Gagne Score of 1 (i.e. just past the location of the slope discontinuity at original *gagne*=0). At the original Gagne Score of 1 (i.e. *gagne*=1 in Output 4 below), the *gagne_shifted* variable has a value of 3, and the *gagne_post_shift* variable has the value of 1.

Output 4.

The FREQ Procedure

<i>gagne</i>	<i>gagne_shifted</i>	<i>gagne_post_shift</i>	Frequency	Percent
-2	0	0	10	0.02
-1	1	0	3710	7.92
0	2	0	10970	23.42
1	3	1	11530	24.61
2	4	2	8130	17.35
3	5	3	5270	11.25
4	6	4	3310	7.07
5	7	5	1740	3.71
6	8	6	1150	2.45
7	9	7	600	1.28
8	10	8	180	0.38
9	11	9	90	0.19
10	12	10	80	0.17
11	13	11	30	0.06
12	14	12	40	0.09
13	15	13	10	0.02

Frequency of original Gagne Score variable (*gagne*) and its two associated recoded variables which reflect the negative (*gagne_shifted*) and positive (*gagne_post_shift*) slopes of the curve in Output 2b.

The associated logistic regression equation and formulas for calculating the odds ratios for a one unit change in Gagne Score on the positively and negatively sloped sides of the curve, respectively, are:

Logistic Regression Equation

$$\text{Logit } (P(\text{hospitalization in 90 days}=1)) = \beta_0 + \beta_1 \text{gagne_shifted} + \beta_2 \text{gagne_post_shift}$$

One unit change in *gagne* on positively sloped side (i.e. original *gagne* > 0):

At original *gagne*=2: $\beta_0 + \beta_1(4) + \beta_2(2)$

At original *gagne*=1: $\beta_0 + \beta_1(3) + \beta_2(1)$

Difference: $\beta_1(1) + \beta_2(1) = \beta_1 + \beta_2$

Odds ratio: $e^{(\beta_1 + \beta_2)} = 2.71828^{(\beta_1 + \beta_2)}$ [Note: 2.71828 is the value of the natural log]

This formula is applied through the use of the below ESTIMATE statement excerpted from the complete SAS code further above. It calculates the difference in the odds of 90 day hospitalization associated with a one unit increase in Gagne Score when Gagne Score is greater than 0:

```
estimate "gagne_gt0" gagne_shifted 1 gagne_post_shift 1/exp cl;
```

One unit change in gagne on negatively sloped side (i.e. original $gagne \leq 0$):

At original $gagne = 0$: $\beta_0 + \beta_1(2) + \beta_2(0)$

At original $gagne = -1$: $\beta_0 + \beta_1(1) + \beta_2(0)$

Difference: $\beta_1(1) = \beta_1$

Odds ratio: $e^{\beta_1} = 2.71828^{(\beta_1)}$ [Note: 2.71828 is the value of the natural log]

This formula is applied through the use of the below ESTIMATE statement, excerpted from the more complete SAS code further above. It calculates this difference in odds of 90 day hospitalization for a one unit increase in Gagne Score when Gagne Score is less than or equal to 0:

```
estimate "gagne_le0" gagne_shifted 1/exp cl;
```

The default odds ratio estimates and 95% confidence intervals are presented below in Output 5a. The odds ratios and 95% confidence intervals generated by the ESTIMATE statements follow in Output 5b.

Output 5a.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.7386	0.1555	310.3016	<.0001
gagne_shifted	1	-0.2447	0.0824	8.8203	0.0030
gagne_post_shift	1	0.4341	0.0855	25.7591	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
gagne_shifted	0.783	0.666	0.920
gagne_post_shift	1.544	1.305	1.825

The *gagne_shifted* odds ratio and 95% confidence interval accurately represent the negative slope of the Gagne curve as verified by the output (5b), below, generated by the ESTIMATE statement for "gagne_le0" (aka $gagne \leq 0$). The odds ratios are the same (0.78; 95% CI 0.67, 0.92) for both the *gagne_shifted* and the estimated *gagne_le0* terms. HOWEVER – the *gagne_post_shift* odds ratio and confidence interval must be further "massaged".

The odds ratio and 95% confidence interval for the negative slope of the Gagne curve as calculated by the ESTIMATE statement are in Output 5b below and are identical to the default odds ratio estimates from Output 5a. However, to accurately represent the positive slope, the negative slope contributed by the *gagne_shifted* variable must be “subtracted out”. This “subtracting out” is done via the ESTIMATE statement for “*gagne_gt0*” (aka *gagne* > 0) with results presented below (Output 5b).

Output 5b.

Estimate										
Label	Estimate	Standard Error	z Value	Pr > z	Alpha	Lower	Upper	Exponentiated	Exponentiated Lower	Exponentiated Upper
<i>gagne_le0</i>	-0.2447	0.08238	-2.97	0.0030	0.05	-0.4061	-0.08320	0.7830	0.6662	0.9202
Estimate										
Label	Estimate	Standard Error	z Value	Pr > z	Alpha	Lower	Upper	Exponentiated	Exponentiated Lower	Exponentiated Upper
<i>gagne_gt0</i>	0.1894	0.009365	20.23	<.0001	0.05	0.1711	0.2078	1.2085	1.1866	1.2309

The appropriate odds ratio and 95% confidence interval for the positive side of the Gagne curve are represented by the exponentiated values for “*gagne_gt0*” in the above output generated by the ESTIMATE statement in PROC LOGISTIC: OR=1.21, 95% CI (1.19, 1.23). This odds ratio, as expected, is less than the odds ratio for the *gagne_post_shift* variable once the contribution of the negative slope part of the equation has been subtracted out.

3. USE RESTRICTED CUBIC SPLINES

Please see Sections 2A and 2B above for a description of how to use restricted splines. The interpretation of spline effect variables is not straightforward (Wicklin), so we did not use this method to present logistic regression estimates for our study.

CONCLUSION

This paper has described several methods for identifying lack of linearity between independent continuous variables and the log-odds of binary dependent variables. The association between these variables can take on many forms. Our paper describes the methods we explored for our specific situation (binary dependent variable, fairly simple J-shaped relationship with only one change in slope). Based on our exploration, we offer a few recommendations below – however, please keep in mind that they reflect our experience and will not apply in all situations.

Of the 3 methods explored for identifying lack of linearity, we recommend the use of the %PSPLINET in order to be able to easily visualize the relationship between the independent variable and the logit of the dependent variable. Box-Tidwell can fail to identify the non-linear relationship and, though it is easy to use, the native SAS implementation of restricted cubic splines through the use of ESTIMATE statements in logistic regression does not automatically produce plots and does not seem to have the capability of sifting through a variety of knots to automatically figure out the best cutpoints for creating spline effects as does %PSPLINET.

This paper also describes several methods for handling continuous variables in logistic regression models when the association between the continuous variables and the logit of the dependent variable do not meet the linearity assumption. Our primary recommendations are:

1. Develop an a priori analysis plan for the handling of continuous variables whose association with the log-odds of the dependent variable might be found to be non-linear. This plan could involve defining a priori cut-points for categorization of the continuous variable into buckets based on relevant literature or clinician judgement; creating “discontinuous” continuous variables using the method of Singer and Willett (2003) to be used in the logistic model with appropriate estimation of the associated odds ratios; or a host of other possibilities – but it is ideal to have a plan before entering the continuous variable fray!
2. Visualize, visualize, visualize...the relationship between the independent continuous variable and the log-odds of the dependent variable. There is no substitute for understanding the form of this relationship!

REFERENCES

- Allison, P. “When Can You Safely Ignore Multicollinearity?”. Accessed September 24, 2019. <https://statisticalhorizons.com/multicollinearity>
- Croxford R. “Restricted Cubic Spline Regression: A Brief Introduction”. Accessed September 17, 2019. <http://support.sas.com/resources/papers/proceedings16/5621-2016.pdf>
- Gagne JJ, Glynn RJ et al. 2011. “A combined comorbidity score predicted mortality in elderly patients better than existing scores”. J Clin Epidemiol 64(7):749-759.
- Harrell, F. “SAS macros and data step programs useful in survival analysis and logistic regression. Includes several macros for fitting spline functions”. Accessed August 26, 2019. <http://biostat.mc.vanderbilt.edu/wiki/pub/Main/SasMacros/survrisk.txt>
- Peduzzi P, Concato J, et al. 1996. “A simulation study of the number of events per variable in logistic regression analysis”. J Clin Epidemiol 49(12):1373-1379.
- Singer, J and Willett JG. 2003. Applied Longitudinal Data Analysis. Modeling Change and Event Occurrence. 1st ed. New York, NY: Oxford University Press.
- Stone CJ and Koo C. “Additive Splines in Statistics” (reprinted from the 1985 Statistical Computing Section, Proceedings of the American Statistical Association. Original pagination is p. 45-58). Accessed September 4, 2019. <http://www.markdiamond.com.au/download/Stone-and-Koo-1986-A4.pdf>
- Wicklin, R. “Regression with restricted cubic splines in SAS”. Accessed August 27, 2019. <https://blogs.sas.com/content/iml/2017/04/19/restricted-cubic-splines-sas.html>

ACKNOWLEDGMENTS

Thanks to the project leads, Matt Maciejewski and Donna Zulman, and team members of the Interconnected Factors That influence Health, Experiences and Needs (IF-THEN) Study for providing both the impetus and the data to explore methods for handling real life situations in which independent continuous variables are not linearly associated with the log-odds of dependent binary variables in logistic regression models. In particular, thanks to team members Liberty Greene, for her extremely helpful review of and suggestions for this paper, and Valerie Smith, for her help implementing the method suggested by Singer and Willett (2003) for creating “discontinuous continuous” variables.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please feel free to contact the author at:

Janet Grubber
Durham Veterans Affairs Medical Center, Health Services Research and Development
919-286-0411 x174080
Janet.grubber@va.gov