

A SAS® Macro for Covariate Specification in Linear, Logistic, or Survival Regression

Sai Liu and Margaret R. Stedman, Stanford University;

ABSTRACT

Specifying the functional form of a covariate is a fundamental part of developing a regression model. The choice to include a variable as continuous, categorical, or as a spline can be determined by model fit. This paper offers an efficient and user-friendly SAS® macro (**%SPECI**) to help analysts determine how best to specify the appropriate functional form of a covariate in a linear, logistic, and survival analysis models. For each model, our macro provides a graphical and statistical single page comparison report of the covariate as a continuous, categorical, and restricted cubic spline variable so that users can easily compare and contrast results. The report includes the residual plot and distribution of the covariate. You can also include other covariates in the model for multivariable adjustment. The output displays the likelihood ratio statistic, the Akaike Information Criterion (AIC), as well as other model-specific statistics. The %SPECI macro is demonstrated using an example data set. The macro includes PROC REG, PROC LOGISTIC, PROC PHREG, PROC REPORT, and PROC SGPLOT procedures in SAS® 9.4.

INTRODUCTION

Many covariates we use in regression models are continuous variables (e.g. age, height, weight), but how we choose to include them in the model is at the discretion of the user. Other functional forms of the covariate (e.g. categorical, or spline) could be specified to improve model fit and have implications for the interpretation of the parameter estimated. Therefore, how to specify the appropriate functional form of a continuous variable is a fundamental consideration and involves a balance between model simplicity and goodness of model fit. Although there are many SAS® procedures available to check data distribution, outliers, and model fit statistics, we are unaware of an existing SAS® procedure that combines the above described outputs together into a one page summary report so that users can quickly compare results from different functional forms of a single covariate.

This paper will introduce a customizable user-friendly SAS® macro **%SPECI** to quickly produce a one page report that organizes multiple commonly-used statistics to help you compare and select the appropriate functional form from continuous, categorical, and spline terms in linear regression, logistic regression, and survival analysis models.

The statistics in the final report include:

- Plot showing an overlay of predicted values from the three functional forms.
- Summary table of model statistics. (See complete list and descriptions for each model in **Appendix A**)
- Panel plot of the residual values from the model where the covariate is continuous, categorical and spline forms.
- Plot of the observed values of the covariate and the outcome variable (linear and logistic regression models only)
- Kaplan Meier plot (survival model only).

INSTRUCTIONS FOR USING MACRO %SPECI

There are two SAS® editor programs: the main macro (**SPECI.sas**) and the program to call the macro (**CALL SPECI.sas**). The call program is provided in the **Appendix B** and both the main macro and the call program are available upon request from the author (Sai Liu) and are posted to the GitHub website (<https://github.com/SaiLMainpage/ModelSpecification>).

First, save the **CALL SPECI.sas** and **SPECI.sas** programs to your computer. Next, open “**CALL SPECI.sas**” and update the include statement to the directory where the “**SPECI.sas**” macro stored

```
%include "Directory/speci.sas";
```

Next, specify the parameters for the macro program (for example %let dataset= mydata) see **Table 1**.

Macro variable	Description	Note
datain	Location your permanent SAS® dataset is saved.	<ul style="list-style-type: none"> • Leave blank if your dataset is already in the work library (Default is SAS® work library). • When specified, include quotations, e.g., “C:\myfiles”.
dataout	Location where one-page report will be saved	<ul style="list-style-type: none"> • This option is required. • Include quotations, e.g., “C:\myfiles”.
dataset	Name of dataset	<ul style="list-style-type: none"> • This option is required.
reportname	Name of one-page report	<ul style="list-style-type: none"> • Default name will be “Model Diagnostic Report”, if left blank.
model	Specify which regression model will be used in this analysis.	<ul style="list-style-type: none"> • This option is required. • Choose one of the following (1-3) 1=linear regression (proc reg) 2=logistic regression (proc logistic) 3=survival model (proc phreg)
yvar	outcome variable	<ul style="list-style-type: none"> • This option is required in linear and logistic models, e.g., %let yvar = stroke • This variable should be coded as 1 for event and 0 for no event for logistic regression. • Leave it blank in survival model.
event	Outcome variable – survival event or status	<ul style="list-style-type: none"> • This option is required in survival model, e.g., %let event = death; • This variable should be coded as 1 for an event and 0 for censored • Leave it blank for linear and logistic models.
time2event	Outcome variable– survival time	<ul style="list-style-type: none"> • This option is required in survival model, e.g., %let time2event = time_to_death; • Survival times should be greater than 0. • Leave it blank in linear and logistic models.
xvar_cont	covariate of interest (continuous)	<ul style="list-style-type: none"> • This option is required, e.g., %let xvar_cont=BMI;
xvar_cat	covariate of interest (categorical)	<ul style="list-style-type: none"> • This option is required, e.g., %let xvar_cat= BMI_CAT;
num_cat	Number of categories for covariate of interest	<ul style="list-style-type: none"> • This option is required, e.g., if BMI_CAT has 4 categories, then %let num_cat= 4; • Must enter a number greater than 1
ref_xvar_cat	the reference category for macro variable “xvar_cat”	<ul style="list-style-type: none"> • Default option will be the alphabetically last or numerically biggest category if left blank

covarlist_cont	List of additional <u>continuous</u> variables for multivariable models	<ul style="list-style-type: none"> List each covariate separated by a single space, e.g. <code>%let xvar_cont = age LOS height;</code> Leave blank if model is not adjusted.
covarlist_cat	List of additional categorical variables for multivariable models	<ul style="list-style-type: none"> List each covariate separated by a single space, e.g. <code>%let xvar_cat = gender race cause_death;</code> Leave blank if model is not adjusted
knot	Number of knots for Spline terms	<ul style="list-style-type: none"> Default is 4 knots, if left blank Otherwise number between 3-10, e.g., <code>%let knot = 5;</code>
norm	Normalization method	<ul style="list-style-type: none"> 0=no normalization 1=normalization (unitless) 2=normalization (original units, default option). see “CALCULATING RESTRICTED CUBIC SPLINES” section for more details
knot1 knot2 knot3 knot4 knot5 knot6 knot7 knot8 knot9 knot10	The percentiles of the data where the 1 st -10 th knots are placed	<ul style="list-style-type: none"> The default assumes 4 knots so if left blank, the default percentiles are: knot1=P5 knot2=P35 knot3=P65 knot4=P95 knot5=blank knot6=blank knot7=blank knot8=blank knot9=blank knot10=blank The number of knots MUST match the number of percentiles for example to specify 70% <code>%let knot1=P70</code>

Table 1: List of macro variable to be specified in the “CALL SPECI” SAS program.

ADDITIONAL NOTES

1. If your working dataset is already in the work library, then only the name of the dataset (“dataset”) is needed. The directory “datain” should be left blank. The program will automatically read the dataset from the current work library. If the dataset is permanent, give the location of your dataset in “datain”, so that the program will find the dataset in the assigned directory.

2. If the covariate of interest has already been categorized in a separate variable, “xvar_cat” should be set equal to that variable. If the covariate of interest has not been categorized in a separate variable, a new variable will need to be created. The categorical variable can be character or numeric. The new dataset with the new variable should be called in the macro.

3. When additional knots are not needed, the rest of the percentile fields should be kept but left blank. For example, if you choose 4 knots in this model, and fill percentiles “knot1”=P5, “knot2”=P35, “knot3”=P65 and “knot4”=P95, then leave “knot5” through “knot10” blank. Do not delete the blank percentiles, otherwise, the program will produce an error.

4. This macro program only allows for a minimum of 3 and a maximum of 10 knots to be included (specified in the %RCSPLINE macro).

CALCULATING RESTRICTED CUBIC SPLINES

A number of SAS® macros are available to perform restricted cubic spline analysis. In this macro we applied %RCSPLINE (Harrell, F.E. 2004) to create the spline terms in the model. This program computes k-2 components of a cubic spline function restricted to be linear before the first knot and after the last knot, where k is the number of knots (Croxford, R. 2016). In addition, the %RCSPLINE program provides three methods to normalize the constructed variables, where normalization means to rescale the values to the normal distribution:

- norm=0: no normalization of constructed variables.
- norm=1: divide by the cube of the difference in the last 2 knots. This normalizes the constructed variables but makes all variables unitless.
- norm=2: divide by square of the difference in the outer knots. This normalizes the constructed variables, but returns all the variables to their original units. (This is the default).

APPLICATIONS OF %SPECI MACRO WITH SAMPLE DATA

In this paper, we apply a logistic regression model to the sample data as an example to illustrate the steps of how to use the %SPECI macro and resulting output. The application of the model to linear regression and survival models will be summarized later highlighting the differences from logistic regression.

SAMPLE DATA

In this paper, we analyzed data from 500 subjects in the Worcester Heart Attack Study (WHAS500, published in Hosmer & Lemeshow, 2008). These data were collected from 1975 to 2001 on all myocardial infarction (MI) patients admitted to hospitals in the Worcester, Massachusetts Standard Metropolitan Statistical Area. The WHAS500 data may be obtained from <http://stats.idre.ucla.edu/wp-content/uploads/2016/02/whas500.sas7bdat>.

Using this data, supposed that we are interested in whether body mass index (BMI) is associated with cardiovascular disease (CVD) and how to best model the association, while adjusting for age (continuous variable) and gender (binary variable). The outcome variable (CVD) is binary (0/1) representing a CVD event occurred (CVD=1) or not (CVD=0). The covariate of interest, BMI, is continuous. Age and gender are additional covariates used to adjust the model. Using the example data, we will examine how to specify the functional form of BMI in a logistic regression model of CVD. We list the selected variables from the WHAS500 dataset in **Table 2**.

Variable Name	Description	Codes / Values
CVD	History of Cardiovascular Disease. Outcome variable.	0=No, 1=Yes
BMI	Body mass index. Independent variable of interest (continuous).	kg/m ²
BMI_CAT	Body mass index. Created from DATA STEP. Independent variable of interest (categorical).	kg/m ²
Age	Age at hospital admission. Covariate.	Years
Gender	Gender. Covariate.	0=Male, 1=Female

Table 2: Description of variables used in the example analysis.

Table 3 shows how to run the %CALL SPECI program using the example data. Since there is not a categorical variable for BMI in the WHAS500 dataset, we first create a new dataset with a categorical variable for BMI. In mydata, BMI is grouped into four categories and named BMI_CAT. We include the code from the main macro program “SPECI.sas” in the %include statement. Next we specify the macro variables in the %let statements. The macro variable “datain” is left blank, because the working dataset “mydata” is created in the work library. If your dataset is a permanent SAS® data set, you will need to let datain equal the path of the dataset here. Xvar_cont and xvar_cat are set equal to the continuous and categorical variables for BMI. In this case there are 4 categories of BMI so num_cat is set equal to 4. We assign the second category of BMI_cat as the reference (%let ref_xvar_cat=2) and adjust for age and gender (%let covarlist_cont=age; %let covarlist_cat=gender). Lastly, we decide to have 4 knots in the spline with cutoff percentiles at 5, 35, 65 and 95 percent, respectively (%let knot=4; %let knot1=P5; %let knot2=P35; %let knot3=P65; %let knot4=P95;). We apply the normalization method that keeps the original units (%let norm=2). After executing %SPECI, the program will automatically generate 4 figures to compare the fit of the continuous, categorical and spline forms of BMI. The 4 figures are combined into a one page report in PDF format, called “Model Diagnostic Report”, and saved in the folder: "C:\Users\sliu\Desktop\Sai Liu\Logistic\report".

```

/* Read in whas500 dataset and grouping BMI into BMI_CAT */
libname lib "C:\Users\sliu\Desktop\Sai Liu\Logistic\Data";
data mydata;
    set lib.whas500;

    if bmi <=18.5 then bmi_cat=1;
    else if 18.5< bmi <=24.9 then bmi_cat=2;
    else if 24.9< bmi <=29.9 then bmi_cat=3;
    else if 29.9< bmi then bmi_cat=4;

run;

%include " C:\Users\sliu\Desktop\Sai Liu\Logistic\Data\speci.sas";

%let datain=""; /*leave blank because the "data" is already in work.library*/
%let dataout="C:\Users\sliu\Desktop\Sai Liu\Logistic\report";
%let dataset=mydata;
%let reportname=Model Diagnostic Report;
%let model=2;
%let yvar=cvd;
%let event=;
%let time2event=;
%let xvar_cont=bmi;
%let xvar_cat=bmi_cat;
%let num_cat=4;
%let ref_xvar_cat=2; /* assign the second group as the reference*/
%let covarlist_cont=age;
%let covarlist_cat=gender;

%let knot=4;
%let norm=2;
%let knot1=P5;
%let knot2=P35;
%let knot3=P65;
%let knot4=P95;
%let knot5=;
%let knot6=;
%let knot7=;
%let knot8=;
%let knot9=;
%let knot10=;
%SPECI; quit;

```

Table 3. Example code for “CALL SPECI.sas”.

Figure A – Predicted Plot Overlay of Continuous, Categorical and Spline Forms

The macro uses the PROC LOGISTIC procedure to estimate model parameters (e.g. $\beta_0, \beta_1, \beta_2, \beta_3, \dots$) for the association between CVD and BMI. The parameter estimates are then used to predict the log odds of (P), where P is the probability of having a CVD event, for a given BMI adjusting for age and gender. The following are the logistic regression models with the continuous, categorical, and spline forms of the covariate and respective SAS code in Table 4.

BMI is continuous:

$$\text{Log} \left(\frac{p}{1-p} \right) = \alpha_0 + \alpha_1 * \text{BMI} + \alpha_2 * \text{AGE} + \alpha_3 * \text{GENDER}$$

Four category BMI (the second category is the reference group):

$$\text{Log} \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 * \text{BMI_cat1} + \beta_2 * \text{BMI_cat3} + \beta_3 * \text{BMI_cat4} + \beta_4 * \text{AGE} + \beta_5 * \text{GENDER}$$

where BMI_cat1=I (BMI<=18.5), BMI_cat3=I (25.0<=BMI<=29.9), BMI_cat4=I (BMI>=30)

BMI in spline form (BMI₁ and BMI₂ are spline terms created from %RCSPINE (Harrell, F.E. 2004);

$$\text{Log} \left(\frac{p}{1-p} \right) = \gamma_0 + \gamma_1 * \text{BMI} + \gamma_2 * \text{BMI}_1 + \gamma_3 * \text{BMI}_2 + \gamma_4 * \text{AGE} + \gamma_5 * \text{GENDER}$$

where $\text{BMI}_i = (\text{BMI} - t_i)_+^3 - (\text{BMI} - t_3)_+^3 - \frac{t_k - t_i}{t_k - t_{k-i}} + (\text{BMI} - t_4)_+^3 + \frac{t_{k-1} - t_i}{t_k - t_{k-i}}$, $i = 1, \dots, k - 2, k = 4$

t_1, \dots, t_4 are the location of the 4 knots, and $u_+ = u$ if $u > 0$, $u_+ = 0$ if $u \leq 0$.

```
* Run separate logistic regression models with exposure variable as
continuous, categorical and spline;
proc logistic data=Data_prep descending outest=est1Line;
  class &covarlist_cat. / param=glm;
  model &yvar.= &xvar_cont. &covarlist_cat. &covarlist_cont. /
    rsquare influence;
run; quit;

data data_prep;
  set data_prep;
  CALL SYMPUT('newref_xvar_cat', PUT(newref, 3.));
run;
%let newcat_macro=newcat;
proc logistic data=data_prep descending outest=est1cat;
class &newcat_macro.(ref="&newref_xvar_cat.")
  &covarlist_cat. / param=glm;
model &yvar.= &newcat_macro. &covarlist_cat. &covarlist_cont./
  rsquare influence; run; quit;

proc logistic data = Data_prep descending outest=est1Spline;
  class &covarlist_cat. / param=glm;
  model &yvar. = &xvar_cont.
    &xvar_cont.1 -- &xvar_cont.%eval(&knot-2)
    &covarlist_cat. &covarlist_cont./
    rsquare influence; run; quit;
```

Table 4: Model code in %SPECI macro.

The output below (Example Output 1) contains a subset of the results from the models. The first table contains parameter estimates where BMI is kept as a continuous variable. The second table contains the parameter estimates for BMI categories and the third table contains parameter estimate from the spline model.

Using these estimates we can predict, for example, the log odds of CVD for a specific BMI, age, and gender, as $-2.8661 + 0.0801 * \text{bmi} + -0.5499 * \text{GENDER} + 0.0321 * \text{AGE}$. From the second table we predict the log odds of CVD, as $-0.9936 + 0.6407 * \text{bmi_cat1} + 0.6066 * \text{bmi_cat3} + 0.8160 * \text{bmi_cat4} + -0.6263 * \text{GENDER} + 0.0309 * \text{AGE}$. From the spline model, we can predict the log odds of CVD as $-5.5472 + 0.2117 * \text{BMI} + -0.4495 * \text{bmi1} + 1.4298 * \text{bmi2} + -0.6474 * \text{GENDER} + 0.0323 * \text{AGE}$.

Example Output 1: Analysis of Maximum Likelihood Estimates with exposure variable in continuous, categories and spline form.

Parameter Estimates with BMI in continuous term

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.8661	1.0008	8.2018	0.0042
BMI		1	0.0801	0.0235	11.6646	0.0006
GENDER	0	1	-0.5499	0.2381	5.3349	0.0209
GENDER	1	0	0	.	.	.
AGE		1	0.0321	0.00820	15.3815	<.0001

Parameter Estimates with BMI in Categorical Form

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.9936	0.6649	2.2327	0.1351
bmi_cat	1	1	-0.6407	0.5041	1.6155	0.2037
bmi_cat	3	1	0.6066	0.2593	5.4713	0.0193
bmi_cat	4	1	0.8160	0.3127	6.8078	0.0091
bmi_cat	2	0	0	.	.	.
GENDER	0	1	-0.6263	0.2454	6.5148	0.0107
GENDER	1	0	0	.	.	.
AGE		1	0.0309	0.00824	14.0437	0.0002

Parameter Estimates with BMI in spline form

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.5472	1.7828	9.6818	0.0019
BMI		1	0.2117	0.0772	7.5236	0.0061
bmi1		1	-0.4495	0.3063	2.1534	0.1423
bmi2		1	1.4298	1.1526	1.5388	0.2148
GENDER	0	1	-0.6474	0.2491	6.7566	0.0093
GENDER	1	0	0	.	.	.
AGE		1	0.0323	0.00825	15.2992	<.0001

To ensure that the plots were aligned with one another, we centralized the predictions (continuous, categorical and spline form) at the midpoint of the exposure of interest so that all lines would cross in the middle of the figure (**Figure 1**). The algorithm for centralization included the following:

- Find the midpoint of the covariate of interest (in this example, the midpoint of BMI is 28.9);
- Predict the value of log odds (Logit (P)) for the continuous, categorical and spline form at middle point (in this example, the predicted values of log odds are -0.546, -0.387, and -0.516 respectively).
- Calculate the centralized value by subtracting each midpoint value of Logit (P) from each predicted value of logit(p).
- Plot centralized log odds of CVD and BMI.

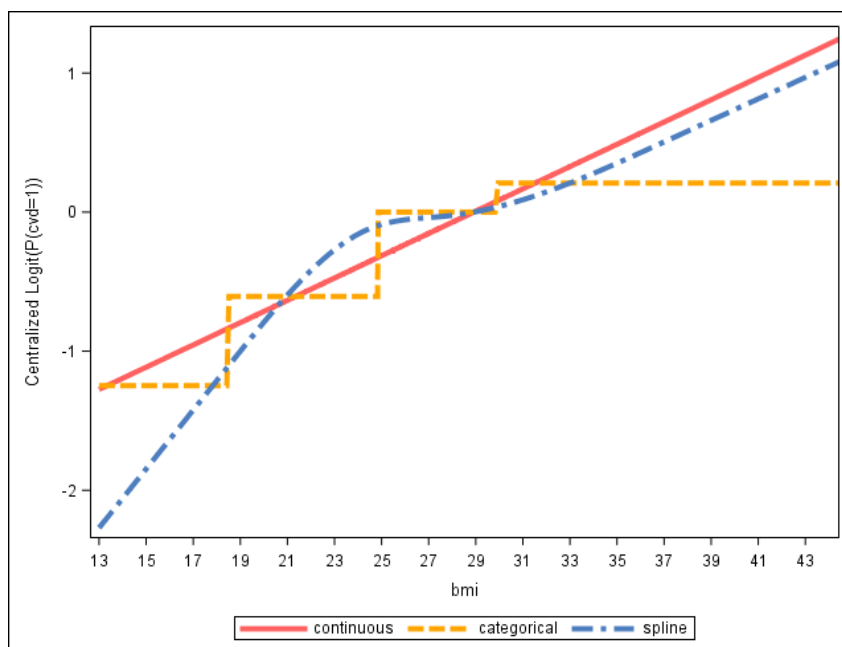


Figure 1. Predicted log odds of CVD with BMI in continuous, categorical and spline form.

Figure B – Summary Table of Statistics

In the logistic regression model, the selected statistics include R-squared, Max-rescaled R-squared (adjusted R-squared), C-statistic, AIC, -2LogL, Likelihood Test, Wald Test, and Model convergence status (see detailed definitions of each statistic in the **Appendix A-1**)

The ODS OUTPUT statement is used to store the desired statistics:

```
* Run Logistic regression model with covariate in continuous form;
ods output Rsquare=estlline_sq FitStatistics=estlline_fit
GlobalTests=estlline_glob convergencestatus=estlline_con
association=estlline_c;

* Run Logistic regression model with covariate in categorical form;
ods output Rsquare=estlcat_sq FitStatistics=estlcat_fit
GlobalTests=estlcat_glob convergencestatus=estlcat_con
association=estlcat_c;

* Run Logistic regression model with exposure variable in spline form;
ods output Rsquare=estlspline_sq FitStatistics=estlspline_fit
GlobalTests=estlspline_glob convergencestatus=estlspline_con
association=estlspline_c;
```

The PROC REPORT procedure is used to create the summary table (**Figure 2**)

```
* Report Summary Table;
options printerpath=png nodate papersize=('8in', '5in');
ods printer file="%dataout.\FigB - &xvar_cont..png";
Proc report data=table nowd ;
column _name_ col1 col2 col3 col4 col5 col6 col7 col8;
define _name_ /"Diagnostic Statistics" group order=data ;
define col1/ "R-Squared" analysis format=10.5 ;
define col2/ "Max-rescaled R-Squared" analysis format=10.5 ;
define col3/ "C-Statistics (bigger is better)" analysis format=10.5 ;
define col4/ "AIC (smaller is better)" analysis format=10.5 ;
define col5/ "-2LogL (bigger is better)" analysis format=10.5 ;
define col6/ "Likelihood Test (P-value)" analysis format=10.4 ;
define col7/ "Wald Test (P-value)" analysis format=10.4 ;
define col8/ "Model Convergence(0=Yes, 1=No)" analysis format=1.0 ;
run;
ods printer close;
ods listing;
```

Diagnostic Statistics	R-Squared (bigger is better)	Max-rescaled R-Squared (bigger is better)	C-Statistics (bigger is better)	AIC (smaller is better)	-2LogL (bigger is better)	Likelihood Test (P-value)	Wald Test (P-value)	Model Convergence(0=Yes, 1=No)
Continuous	0.05754	0.08521	0.66578	540.70592	532.70592	0.0000	0.0000	0
Categorical	0.05577	0.08259	0.66617	545.64311	533.64311	0.0000	0.0001	0
Spline	0.06374	0.09440	0.67342	541.40336	529.40336	0.0000	0.0000	0

Figure 2. Summary table of statistics from models with BMI in continuous, categorical and spline Form.

Figure C – Pearson Chi-Square Residual Plot

The Pearson Residual is one of the most commonly used methods for logistic regression diagnostics. Obvious patterns (e.g. U shaped) in the distribution of the residuals are a likely indicator that the continuous variable has a poor fit. The PROC SGPLOT procedure was used to plot the Pearson (Chi-square) residual value and observed values for the continuous, categorical and spline forms of BMI (see **Figure 3**). The Pearson chi-square residuals measure the relative deviations of the observed values from the fitted values. It is calculated from the differences between the observed and fitted values and divided by the standard deviation. A residual greater than 3 or less than -3 shows areas where there is poor fit or an outlier. The Pearson residual is calculated as

$$p_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1 - \hat{\mu}_i)}}$$

y_i is the observed value of the outcome for the i th observation (0/1) and $\hat{\mu}_i$ is the predicted probability of the event for the i th observation (SAS institute,2009).

```
* output Pearson residual with model of covariate in continuous,
categorical and spline terms;
proc logistic data=Data_prep descending outest=est1Line;
  class &covarlist_cat. / param=glm;
  model &yvar.= &xvar_cont. &covarlist_cat. &covarlist_cont.;
  output out=rplot_c prob=p reschi=pr;run;
proc logistic data=Data_prep descending outest=est1Line;
  class &xvar_cat. &covarlist_cat. / param=glm;
  model &yvar.= &xvar_cat. &covarlist_cat. &covarlist_cont.;
  output out=rplot_cat prob=p reschi=pr;run;
proc logistic data=Data_prep descending outest=est1Line;
  class &covarlist_cat. / param=glm;
  model &yvar.= &xvar_cont. &xvar_cont.1 -- &xvar_cont.%eval(&knot-2)
&covarlist_cat. &covarlist_cont.;
  output out=rplot_sp prob=p reschi=pr;run;

data rplot_cat;length Form $12.;set rplot_cat(keep=&xvar_cont.
pr);Form='Categorical';run;
data rplot_c;length Form $12.;set rplot_c(keep=&xvar_cont.
pr);Form='Continuous';run;
data rplot_sp;length Form $12.;set rplot_sp(keep=&xvar_cont.
pr);Form='Spline';run;

proc sort data=rplot_cat;by &xvar_cont.;run;
proc sort data=rplot_c;by &xvar_cont.;run;
proc sort data=rplot_sp;by &xvar_cont.;run;
data rplot_all;set rplot_c rplot_cat rplot_sp;by &xvar_cont.;run;

* Panel plot of Pearson Residual with observed data by functional forms;
ods listing gpath="&dataout.";
ods graphics on /reset=index imagefmt=png imagename="FigC - &xvar_cont." ;
proc sgpanel data =rplot_all;
panelby Form/columns=3;
scatter x=&xvar_cont. y=pr;
colAXIS values=(&minvalue. to &maxvalue.) LABEL="&xvar_cont." ;

refline 0 /transparency=0.2 axis=y;
refline 3 -3 /transparency=0.6 axis=y;run;
ods graphics off;
```

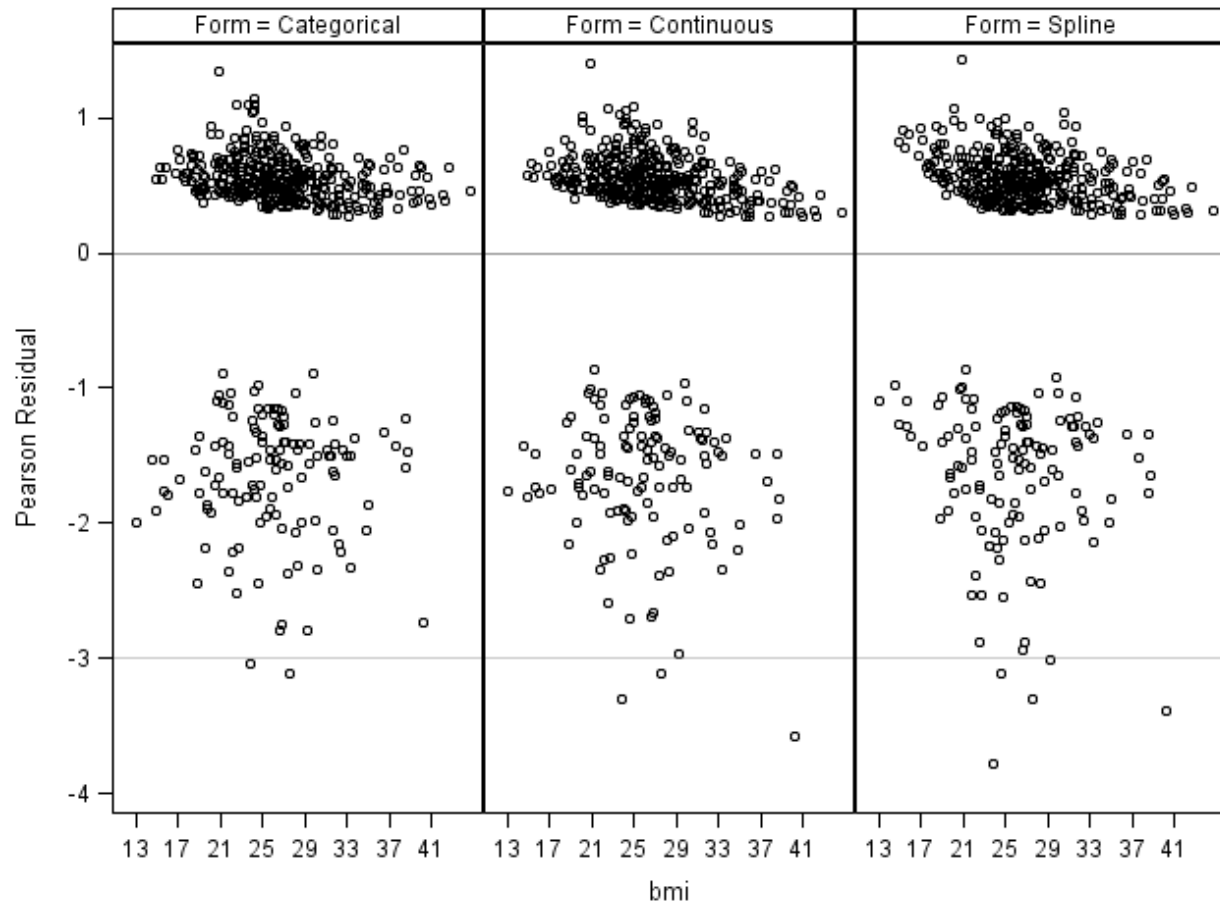


Figure 3. Pearson chi-square residual plot with BMI in continuous, categorical and spline forms

Figure D – Distribution of Observations

We use the SGPLOT procedure to show a scatter plot of the distribution of BMI with CVD (0/1) (see **Figure 4**). This figure can be used to check the variability in the data and identify outliers.

```

Plot data distribution;
ods listing gpath="&dataout.";
ods graphics on /reset=index imagefmt=png imagename="FigD - &xvar_cont." ;
proc sgplot data = Data_fig;
scatter x = &xvar_cont. y=&yvar.;
yaxis values=(0 to 1 by 1);
XAXIS values=(&minvalue. to &maxvalue.) LABEL="&xvar_cont." ;
run;
ods graphics off;
ods printer close;

```

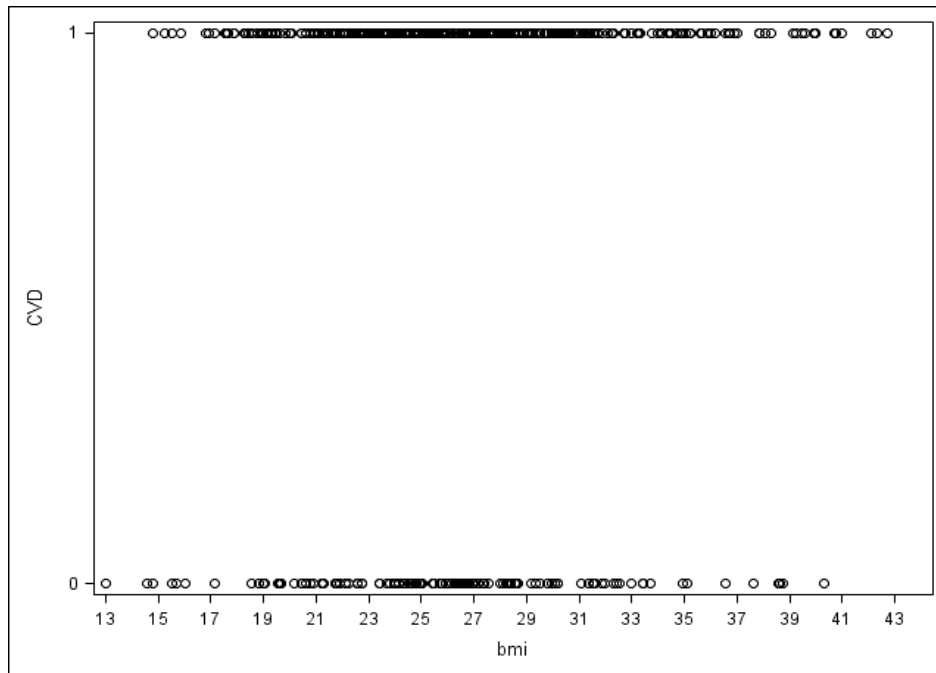


Figure 4. Correlation between BMI and CVD

FINAL REPORT AND INTERPRETAION OF RESULTS

The four figures described above are combined into a one page PDF report (**Figure 5**). The top left figure shows the predicted centered log odds of CVD by BMI for the 3 functional forms: continuous, categorical, and spline. The plots have been realigned to overlay at the midpoint of BMI (28.9). From this figure we see that slope of the continuous result aligns well with the categorical and spline results above the midpoint. For low values of BMI (below 18), the categorical and spline results are below the continuous result. For BMIs between 24 and 28, the spline and categorical variables show a flat association between BMI and CVD, which is not captured by the continuous variable. In the residual plot (bottom left), most of data fall between -3 and +3. There are few points below -3, in all three forms, however, the spline has the most points below -3, which may indicates a worse fit. The top right figure summarizes results from the diagnostic statistics to help compare how well each model fits. In this case the spline model had the best r-square and c statistic, but the AIC and -2 log Likelihood statistics favored the continuous and categorical forms, so there is no obvious winner among the forms examined. The bottom right figure of CVD and BMI shows that the data do not have extreme high or low observations, however there is some sparse data at the tails which may be contributing to the discordance between the results in the low BMI range.

Based on these results we recommend investigating the points with residuals below -3 and possibly excluding them from the analysis or selecting the categorical form to improve model fit. Also we may want to limit the analysis to only those BMI's greater than 18. Given that the differences between the models are not extreme the user may prefer the simplicity of interpreting a continuous or categorical form of BMI rather than the spline form.

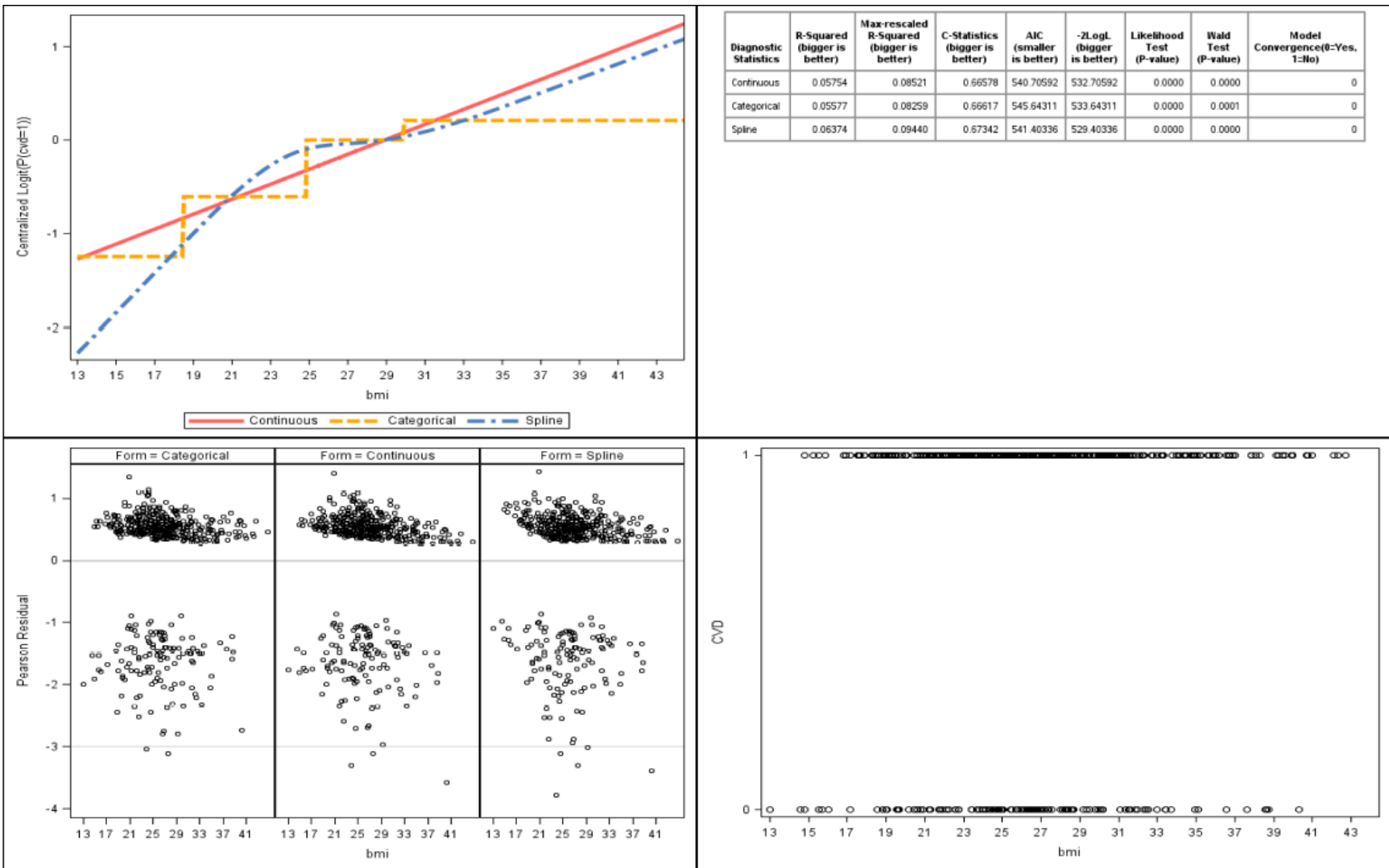


Figure 5. Model Specification Report

LINEAR REGRESSION

In the linear regression analysis, we use the PROC REG procedure to model the data. Since the REG procedure does not support categorical predictors directly, the categorical variable is recoded into a series of dummy variables prior to including them in the model. The program also provides an option to specify the reference group (see “ref_xvar_cat” in the “Parameter Setting” section). For example, consider a four category BMI variable. The program would automatically create four indicator variables as bmi_catdum1 to bmi_catdum4, three of which would be included in the model. The appendix contains a list of statistics included in the report (**Appendix A-2**). We present standard Pearson residuals with observed data for the residual plot.

SURVIVAL ANALYSIS

In the survival model, we apply the PROC PHREG procedure to perform the Cox proportional hazards model. In this case we plot the predicted of Log (HR) with the covariate of interest in continuous, categorical and spline form as well as provide a summary table of statistics (see **Appendix A-3**). Unlike the linear and logistical models we plot deviance residuals and include a Kaplan-Meier plot. If the Deviance residuals are above 3 or below -3, then there the model fits poorly in that area. (Paul D.Allison, 1995).

CONCLUSION

The %SPECI macro is a user-friendly tool to support modelers with determining the best functional form for a continuous predictor variable for linear, binary, and survival models. The macro creates a summary report with visual and statistical diagnostics to describe model fit for 3 different functional forms of the variable of interest: continuous, categorical, and spline. This paper illustrates the features of this macro using a real life example where CVD is modeled from BMI. Mathematical modeling is a challenging problem requiring topic expertise as well as mathematical and computational skill. This macro offers a tool to support model development and results should be considered carefully in the context of existing knowledge about the topic.

REFERENCE

Croxford, R. (2016). “Restricted Cubic Spline Regression: A Brief Introduction.” paper 5621-2016. *Proceedings of the SAS Global 2016 Conference, Las Vegas, NT*. Available at <http://support.sas.com/resources/papers/proceedings16/5621-2016.pdf>

Harrell, F.E. (2004) SAS Macros for Assisting with Survival and Risk Analysis, and Some SAS Procedures Useful for Multivariable Modeling. Available at <http://biostat.mc.vanderbilt.edu/wiki/Main/SasMacros>.

Allison, Paul D., *Survival Analysis Using the SAS® System: A Practical Guide*, Cary, NC: SAS Institute Inc., 1995. 292 pp.

WHAS500, published in Hosmer & Lemeshow (2008). Available downloading at <http://stats.idre.ucla.edu/wp-content/uploads/2016/02/whas500.sas7bdat>.

SAS/STAT(R) 9.2 User's Guide, Second Edition, Regression Diagnostics, Cary, NC: SAS Institute Inc., https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_logistic_sec1042.htm

ACKNOWLEDGMENTS

I would like to thank my colleagues Dr. Maria M. Rath, Dr. Jin Long and Yuanchao Zheng for their insightful comments and review.

CONTACT INFORMATION

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

Sai Liu

Division of Nephrology, Department of Medicine

Stanford University School of Medicine

1070 Arastradero Rd., Suite 100

Palo Alto, CA. 94304

Phone: 213-793-1055

Email: sailiu.tian@gmail.com or

sailiu@stanford.edu

APPENDIX

Appendix A

Diagnostic Statistics	Explanation	Status
Error Degree of Freedom	Error Degree of Freedom = Degree of Freedom Total - Degree of Freedom Model	larger values indicate better models
R-squared	The proportion of variance in outcome variable explained by covariates.	larger values indicate better models
Adjusted R-squared	Adjusted R-squared is the r-squared adjusted for the number of covariates in the model.	larger values indicate better models
MSE (Mean Square Error)	The average squared difference between the observed outcomes and the predict outcomes.	smaller values indicate better models
AIC (Akaike Information Criterion)	It is calculated as $AIC = 2p - 2 * \ln \hat{L}$ where \hat{L} = the maximized value of the likelihood function of the model, p = the number of estimated parameters in the model.	smaller values indicate better models
BIC (Sawa's Bayesian Information Criterion)	It is calculated as $BIC = p * \ln(n) - 2 * \ln \hat{L}$, where \hat{L} = the maximized value of the likelihood function of the model, n= sample size, p= the number of estimated parameters in the model.	smaller values indicate better models

Table A-1 Statistics for Linear Regression Model

Diagnostic Statistics	Explanation	Status
R-squared	The proportion of variance in the outcome explained by the covariates.	larger values indicate better models
Adjusted R-squared	Adjusted R-squared is the r-squared adjusted for the number of covariates in the model. also called "Max-rescaled R-Squared".	larger values indicate better models
C-statistics	Equivalent to the area under the receiver operating characteristic (ROC) curve. A value below 0.5 indicates a very poor model. A value of 0.5 means that the model is no better than predicting the outcome than random chance.	larger values indicate better models
AIC (Akaike Information Criterion)	It is calculated as $AIC = -2 \text{Log } L + 2((p-1) + s)$, where p is the number of levels of the dependent variable and s is the number of predictors in the model.	smaller values indicate better models
-2LogL	-2 Log L is negative two times the log-likelihood, which used in hypothesis tests for nested models.	larger values indicate better models
Likelihood Test (P-value)	This is the Likelihood Ratio (LR) Chi-Square test. Test is significant if at least one of the predictors' regression coefficient is significant in the model.	Significant if $P < 0.05$
Wald Test (P-value)	Wald Chi-Square. Tests that at least one of the predictors' regression coefficient is not equal to zero in the model.	Significant if $P < 0.05$
Model Convergence	This describes whether the maximum-likelihood algorithm has converged or not.	0=Converged 1=Not Converged

Table A-2 Statistics for Logistic Regression Model

Diagnostic Statistics	Explanation	Status
-2LogL	-2 Log L is negative two times the log-likelihood, which is used	larger values

	in hypothesis tests for nested models.	indicate better models
AIC (Akaike Information Criterion)	It is calculated as $AIC = -2 \log L + 2(p-1) + s$, where p is the number of levels of the dependent variable and s is the number of predictors in the model.	smaller values indicate better models
BIC (Sawa's Bayesian Information Criterion)	It is calculated as $BIC = p * \ln(n) - 2 * \ln \hat{L}$, where \hat{L} = the maximized value of the likelihood function of the model, n= sample size, p= the number of estimated parameters in the model.	smaller values indicate better models
Likelihood Test (P-value)	This is the Likelihood Ratio (LR) Chi-Square test that at least one of the predictors' regression coefficient is not equal to zero in the model.	Significant if $P < 0.05$
Wald Test (P-value)	Wald Chi-Square Test that at least one of the predictors' regression coefficient is not equal to zero in the model.	Significant if $P < 0.05$
Model Convergence	This describes whether the maximum-likelihood algorithm has converged or not.	0=Converged 1=Not Converged

Table A-3 Statistics for Survival Model

APPENDIX B - FULL CODES OF "CALL SPECI.SAS" MACRO PROGRAM

```
/******  
/*      NAME: CALL SPECI.SAS                               */  
/*      TITLE: Functional form Specification for Linear, Logistic, and Survival      */  
/*              Models                                     */  
/*      AUTHOR: Sai Liu, MPH, Stanford University          */  
/*      OS: Windows 7 Ultimate 64-bit                    */  
/*      Software: SAS 9.4                                 */  
/*      DATE: 29 DEC 2016                                 */  
/* DESCRIPTION: This program shows how to call the SPECI.sas macro                  */  
/*      DOWNLOAD: Both CALL SPECI.SAS AND SPECI.SAS macro programs could be      */  
/*              download at the following site:          */  
/*              https://github.com/SaiMainpage/ModelSpecification */  
/******  
  
%let datain=;                /*Location of permanent SAS dataset. Leave it blank, if  
                             your dataset is in the work library*/  
  
%let dataout=;              /*Location of one-pager report will be saved*/  
  
%let dataset=;             /*Name of the dataset*/  
  
%let reportname=;         /*Name of the one-pager report If you leave it blank, this  
                           program will give a name as "Model Diagnostic Report" */  
  
%let model=;              /*1=linear, 2=logistic, 3=survival*/  
  
%let yvar=;               /*Dependent variable of interest for linear and logistic  
                           regression model only, otherwise leave it blank. e.g. %let  
                           yvar= heartfail; or %let yvar= ;(if this is not for linear  
                           nor logistic regression model)*/  
  
%let event=;             /*Dependent variable of interest for survival model only,  
                           otherwise leave it blank e.g. %let event= death; or %let  
                           event= ;(if this is not for survival model)*/  
  
%let time2event=;        /*Dependent variable time component for survival models  
                           only, otherwise leave it blank. e.g. %let time2event=  
                           time2death; or %let time2event= ;(if this is not for  
                           survival model)*/  
  
%let xvar_cont=;         /*Independent variable of interest (continuous). e.g. %let  
                           xvar_cont= age; */  
  
%let xvar_cat=;          /*Independent variable of interest (categorical).  
                           If you don't have an exist categorical variable in the  
                           dataset, please create a categorical variable and entry the  
                           name of created variable here and leave datain= blank,  
                           because the main dataset is already in work library, this  
                           program won't read a permanent SAS dataset.e.g. %let  
                           xvar_cat= bmi_cat; */  
  
%let num_cat=;          /*# of categories of above categorical variable. MUST enter  
                           a numeric number, can't leave it blank. e.g. bmi_cat has 4  
                           categories, then %let num_cat= 4;*/  
  
%let ref_xvar_cat=;      /*Specify the reference group of the xvar_cat variable.  
                           e.g. the 2nd category of bmi_cat is the reference,  
                           then %let ref_xvar_cat= 2; If you leave it blank, this  
                           program will set 1st category as the reference */  
  
%let covarlist_cont=;    /*continuous covariates for model adjustment. OPTIONS: 1)  
                           Leave it blank if no continuous covariate to be included in  
                           the model or 2) add continuous covariates as needed and add  
                           one space between covariates. e.g. %let covarlist_cont= age  
                           height weight; */  
  
%let covarlist_cat=;     /*categorical covariates for model adjustment. Need one  
                           space between covariates. OPTIONS: 1) Leave it blank if no  
                           categorical covariate to be included in the model or 2) add
```

```

categorical covariates as needed and add one space between
covariates. e.g. %let covarlist_cat= race year; */

%let knot=; /*# of Knots for Spline. MUST enter a number from 4 to
10.Because NO SPLINE VARIABLES CREATED if number of knots
<=3 in this program Default is 4*/

%let norm=; /*Normalization method. Options: 0, 1, or 2. Default is 2*/

%let knot1=; /*Cutoff percentile at 1st knot e.g. %let knot1= p5 (p5 for
5th percentile)*/

%let knot2=; /*Cutoff percentile at 2nd knot e.g. %let knot1= p35 (p35
for 35th percentile)*/

%let knot3=; /*Cutoff percentile at 3rd knot e.g. %let knot1= p65 (p65
for 65th percentile)*/

%let knot4=; /*Cutoff percentile at 4th knot e.g. %let knot1= p95 (p95
for 95th percentile)*/

%let knot5=; /*Cutoff percentile at 5th knot. Leave it blank if you
don't specify*/

%let knot6=; /*Cutoff percentile at 6th knot. Leave it blank if you
don't specify*/

%let knot7=; /*Cutoff percentile at 7th knot. Leave it blank if you
don't specify*/

%let knot8=; /*Cutoff percentile at 8th knot. Leave it blank if you
don't specify*/

%let knot9=; /*Cutoff percentile at 9th knot. Leave it blank if you
don't specify*/

%let knot10=; /*Cutoff percentile at 10th knot. Leave it blank if you
don't specify*/
/*****
/* please download SPECI.sas first then call it from the directory */
%include "Directory/speci.sas";

/* In case you need DATA STEP, please add them below */

/* libname lib "";
data mydata;
set lib.&dataset.;

* creating categorical variable; run;*/

* Call %SPECI.sas macro program;
%SPECI;
Quit;

```