SAS® GLOBAL FORUM 2020 #SASGF

Paper 4925-2020

Use SAS® to Model Non-Linearity

Xiaoting Wu, Department of Cardiac Surgery, Michigan Medicine, Ann Arbor, MI;

Donald S. Likosky, Department of Cardiac Surgery, Michigan Medicine, Ann Arbor, MI; The Michigan Society of Thoracic and Cardiovascular Surgeons Quality Collaborative

ABSTRACT

The relationship between an outcome and a continuous predictor is often not linear. However, linearity is one of the assumptions for a multivariable regression model. Many options, such as transformation and restricted cubic splines, are available to handle non-linear relationships; however, these models are often hard to interpret. Linear spline is a simple approach to account for non-linearity and can provide interpretable results. This paper illustrates the use of linear splines to describe the relationship between a continuous variable and a binary outcome in a regression model. We used the relationship between hematocrit and blood transfusion as an example. Low hematocrit, a continuous measurement of volume percentage of red blood cells in whole blood, is known to be associated with blood transfusion in a clinical setting. We first assessed the non-linearity using the LOESS, GAM and SGPLOT procedures. We then constructed the linear splines using BASE SAS programming. Lastly SAS Logistic procedure was used to estimate the linear splines. Emphasis is given to model interpretation to demonstrate the value of linear splines.

Key words: splines, regression models, non-linearity

INTRODUCTION

Regression model is a common analytic approach in epidemiology study. These models require linearity assumption between independent continuous variables and dependent variables. However, nonlinear relationship between risk factors and outcome is common. Failure to correctly specify the functional forms of a continuous risk variable could lead to poor model fit and inaccurate estimate of a relationship [1]. There are many strategies to handle non-linearity including categorization, transformation, cubic splines or nonparametric models [2, 3] . Categorizing the continuous variable is simple but usually cause loss of information and induce bias [4, 5]. Other techniques such as transformations and high polynomials functional form provide flexibility in describing non-linearity, but are difficult to interpret particularly when outcome is binary [6].

Using linear splines is a simple way to handle non-linearity and can provide interpretable results. When using linear spline, knots are introduced to the model and slopes change at the data knots. In a logistic model, odds ratio can be calculated within intervals between the knots.

A logistic model using linear splines can be defined as below with k nots at a_1 ….a_k as following:

$$Logit\ E\ (Y) = \beta_0 + \beta_1 X + \beta_2 (X - a_1)_+ + \beta_3 (X - a_2)_+ + \cdots + \beta_{k+1} (X - a_k)_+$$

The terms of $(X - a_k)_+$ has the value of $X - a_k$ if it is positive, and 0 otherwise.

The goal of this paper is to compare methods to model non-linearity and demonstrate the process of using linear spline to model non-linear relationship between a continuous variable and a binary outcome. Odds ratio and confidence intervals were derived from a linear spline model to provide interpretation. This paper will showcase the implementation in SAS.

DATA EXAMPLE

In cardiac surgery, blood transfusion risk is closely related to a patient's hematocrit (hct) level. Hematocrit is a continuous measurement of patients' red blood cells volume ratio to the whole blood. The thresholds of hematocrit that impact operative blood transfusion can provide guidance for blood reservation strategies prior to the surgery. These thresholds however remain obscure. We utilized surgical data which comprised more than 20,000 coronary artery bypass grafts procedures from multiple hospitals [7]. The transfusion rate was 36.8%. The aim of this analysis is to identify potential thresholds of preoperative hematocrit that associates with increased blood transfusion risk.

DIAGNOSTIC OF NON-LINEARITY

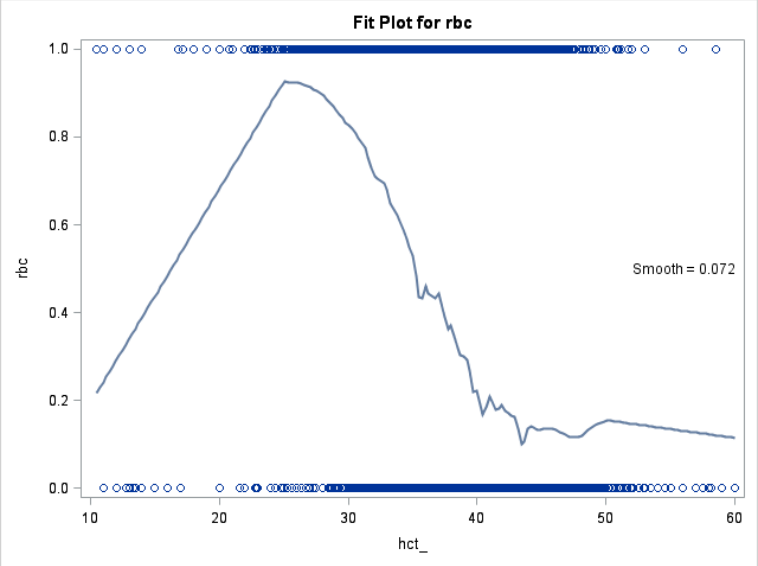
Local regression: LOESS procedure

Locally estimated scatterplot smoothing (LOESS) method implements nonparametric local regression to the data. LOESS curves can be used to reveal trends in data. In SAS LOESS procedure, a plot function- fitplot can be used to obtain a smoothing plot from local regression. Here, transfusion (variable: rbc) is a binary outcome with the scattered observed data points at 1 and 0 (Figure 1). The

relationship between rbc and hematocrit with a smooth parameter of 0.072 is shown as a non-linear curve below. Note that a logistic model has a logit link function for the outcome. Thus this LOESS plot cannot directly provide a visualization for relationship between covariates and logit form of outcome. However, LOESS procedure can be very useful for checking relationship between two continuous variables.

```
proc loess data=spline_ex plots (MAXPOINTS=NONE)=fitplot;
  model rbc = hct_;
run;
```

Figure 1. Fit Plot from LOESS



Generalized Additive Model (GAM)

In order to adjust other covariates and use a logit link for the dependent variable, a generalized additive model (GAM) is implemented with SAS GAM procedure [8]. Below PROC GAM fits a logistic additive model with the binary outcome (transfusion) against hematocrit using smoothing splines while adjusting for other parameters including female and admission acuity status.

$$\text{logit (rbc)} = \beta_0 + \beta_1 female + \beta_2 status3c + \beta_3\, hct + \text{spline (hct)}$$

```
proc gam data=spline_ex descending plots=all;
   class female status3c;
   model rbc (event='1')= param (female   status3c) spline(hct_, df=4)
/dist=binomial;
   output out=predictgam p UCLM LCLM;
run;
```

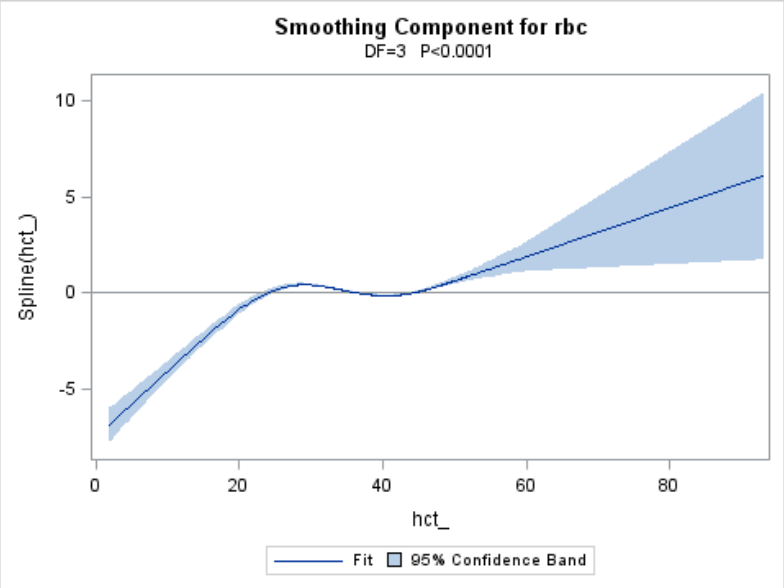
PROC GAM by default produces a panel of plots of partial prediction curves of smoothing components. The partial prediction for a predictor such as hct is its

nonparametric contribution to the model such as spline (hct_). Figure 2 shows the shape of spline function for hct. The slopes change at around hct of 28 and 43 for a degree of freedom=3. The analysis of deviance (Table 1) provides a test for non-linearity. This test shows a significant contribution to the model (p<.0001) from the spline terms of hematocrit, which indicates a non-linear relationship.

Table 1. GAM smoothing model analysis output

Smoothing Model Analysis Analysis of Deviance				
Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(hct_)	3.00000	419.281286	419.2813	<.0001

Figure 2. Plot of smoothing component in GAM



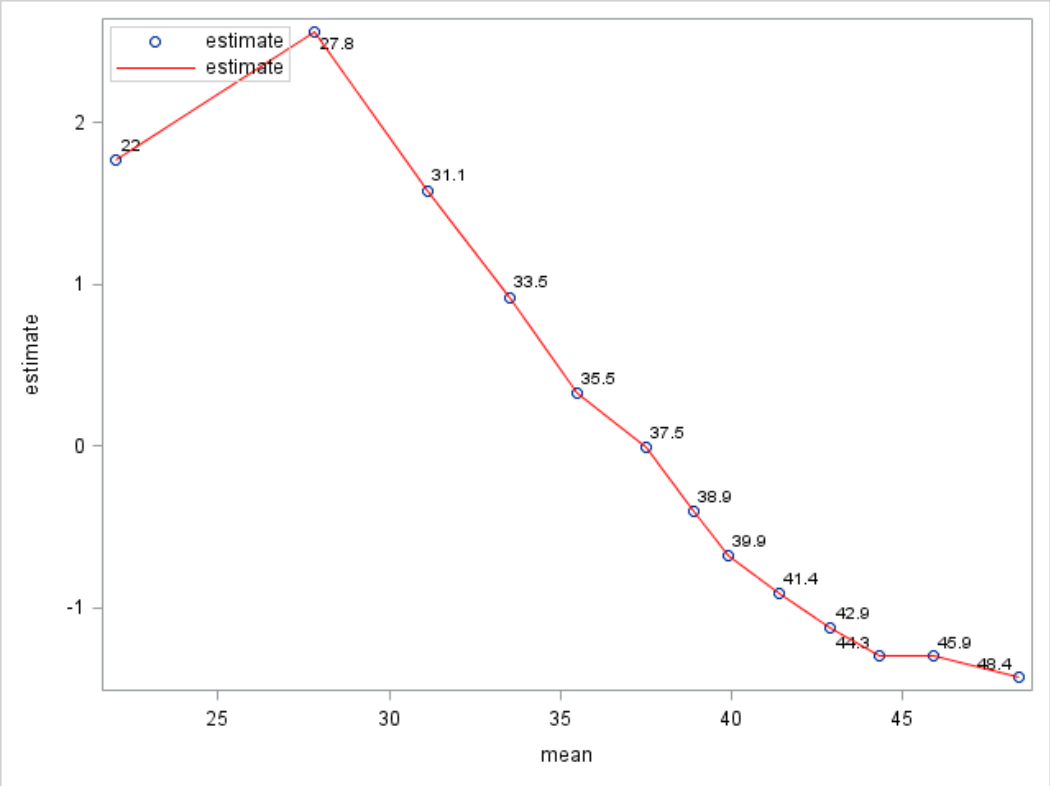
Grouped data plot

In order to visualize the crude relationship between hematocrit and the transfusion outcome, we grouped the continuous variable into selected percentiles. We used the groups as a categorical variable and fit a logistic model to obtain estimates for each group. We plotted the mean of the continuous variable within each group as x axis, and logistic model estimates as the y axis.

Specifically, we used a macro to create groups variable based on some selected percentiles of hematocrit. As the lower and higher ends of hct are likely to have different effect on transfusion risk, we put in more categories at both ends. So we use percentile at 1, 5, 10, 20, …, 80, 90, 95, 99 for hematocrit. We then insert the categorized variable quinthct into the model and obtain model estimates.

We next plot the estimate from the model (log odds of transfusion) as the y-axis and the mean of hct at each category as the x-axis (Figure 3). The codes of this process is provided in the supplemental materials.

Figure 3. Log odds of transfusion against hematocirt percentiles



COMPARISONS OF METHODS FOR MODELING NON-LINEARITY

We next compared several strategies to handle the non-linear relationship between hct and transfusion risk using SAS (Table 2). SAS procedures LOGISTIC, GLIMMIX and GLMSELECT have EFFECT statement to generate splines terms, such as cubic splines, B-splines. Note that GLMSELECT cannot be used for logistic models. Besides manually creating splines with DATA STEP, OUTDESIGN option in LOGISTIC, GLIMMIX procedures can be used to output spline terms into a dataset.

SAS GAM and GAMPL procedures enable more flexible splines models than LOGISTIC, and allow logit link function for a binary response. GAM constructs splines using partial residuals against individual smoothing terms, while GAMPL procedures construct splines using global model evaluation criterial. These two procedures also have different algorithms to estimate spline parameters. Both procedures do not provide model diagnostic such as AIC and c-statistics. We thus output the prediction from these models and used ROC options in LOGISTIC procedure to obtain c-statistics.

In Table 2, we summarize SAS procedures for modeling non-linearity with a logistic model. In our data example, all models have very similar model

performance in terms of c-statistics. Linear splines and GAM models have the smaller AIC which indicates better models.

Table 2. Different strategies to model non-linearity in SAS

Method ID	Strategy	SAS procedures	df	AIC	c-statistic
0	Categories	Logistic	4	21865.52	0.777
1	Assume linearity	Logistic	1	21832.41	0.782
2	RCS	Logistic, effect	4	21727.6	0.783
3	Truncated power function spline	Logistic, effect	6	21505.578	0.784
4	GAM, GCV	GAM, method=GCV	4	-	0.783
5	GAM	GAMPL	7.1	21500	0.784
6	Linear splines	Knots at 18,25,36,43	5	21503.0	0.784
7	Linear splines2	Splines knots of 25, 36, 43	4	21502.4	0.784

In our example, we used a DATA step to generate linear spline terms of hematocrit at 25, 36 and 43.

```
*linear splines;
data spline_ex;
  set spline_ex;
    hct1 = hct_;
    hct2 = max(hct_ - 25, 0);
    hct3 = max(hct_ - 36, 0);
    hct4 = max(hct_ - 43, 0);
run;

proc logistic data=spline_ex ;
        class female  status3c ;
        model rbc (event='1')=   female  status3c  hct1 hct2  hct3 hct4 /lackfit covb;
        output out=estimates p=est_response xbeta=linear_pred; *output predicted
from the original data;
        score data=score_dat out=score_est; *output predicted with new data that
fixed the covariable values;
run;
```

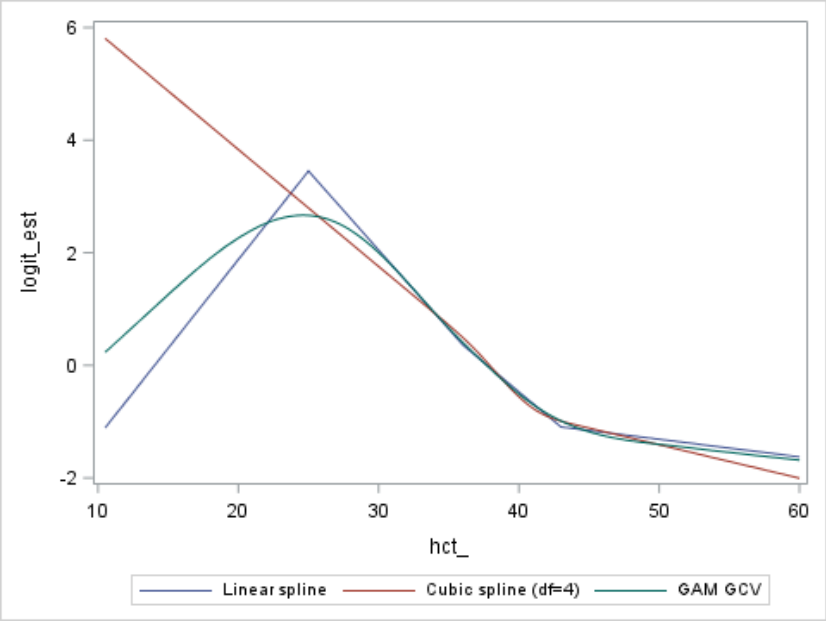
To visualize the relationship between hct and transfusion, we used SCORE statement to obtain predicted probability from models. We then combined the prediction from each model, and used SGPLOT to plot logit form of transfusion probability against hct. From our results (Figure 4), GAM gcv model is the most flexible model, but it's complicated and hard to interpret. Cubic spline models with degree freedom of 4 failed to capture the relationship at the lower end of hct. Linear spline models maintained the distribution shape similar to the ones from

GAM GCV model and the grouped data plot (Figure 3), as well as provide simple interpretation.

Data Merge_score was constructed to combine prediction in logit form from different models. Here is an example to visualize the estimated distribution from different models.

```
PROC SGPLOT DATA = merge_score;
 SERIES X = hct_ Y = logit_est /LEGENDLABEL="Linear spline";
 SERIES X = hct_ Y = logit_cubic /LEGENDLABEL="Cubic spline (df=4)";
 SERIES X = hct_ Y = LINP_rbc /LEGENDLABEL="GAM GCV";
RUN;
```

Figure 4. The relationship of hematocrit and logit form of transfusion from different models



CHOICE OF SPLINE KNOTS

Previous study has found that the location of knots in a spline model is not very crucial in most situations [1]. It is common to place knots at fixed quantiles (percentiles) of a predictor's marginal Distribution, or place them at cutoff based on prior experience [1]. Knots placements may be chosen with KNOTMETHOD=option in the EFFECT statements. These knots options include percentiles, equal, rangefractions, list, percentilelist.

Based on observations from our data and clinical perspective, possible knots for hematocrit locate roughly at 18, 25, 36, 43. We plugged in different choices of knots from those candidates and tested the model fit. Choosing splines knots of 25, 36, 43 yields a simpler and better model based on the smaller AIC.

LINEAR SPLINE INTEPRETATION

Finally, we identified our model with linear spline terms of hematocrit with knots at 25, 36 and 43. Here, we showcase the process to estimate odds ratio in each interval of hematocrit, and the corresponding confidence interval.

Table 3A. The linear spline model output.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.4782	0.8118	30.4279	<.0001
female	0	1	-0.3689	0.0189	380.5256	<.0001
status3c	Elective	1	-0.3063	0.0357	73.5909	<.0001
status3c	Emergent	1	0.6976	0.0607	132.0481	<.0001
hct1		1	0.3147	0.0335	88.4773	<.0001
hct2		1	-0.5957	0.0380	245.6057	<.0001
hct3		1	0.0728	0.0163	19.8520	<.0001
hct4		1	0.1770	0.0223	63.1901	<.0001

Table 3B. The covariance matrix of spline terms

Estimated Covariance Matrix				
Parameter	hct1	hct2	hct3	hct4
hct1	0.00112	-0.00123	0.00013	-0.00004
hct2	-0.00123	0.001445	-0.00029	0.000102
hct3	0.00013	-0.00029	0.000267	-0.0002
hct4	-0.00004	0.000102	-0.0002	0.000496

The above results (Table 3A) will be interpreted as the equation below:
Log (odds)= --4.47 – 0.368(female)-0.306 (Elective admission) + 0.697 (Emergent admission) + 0.314 (hct) -0.59 $(\text{hct-25})_+$ +0.072 $(\text{hct-36})_+$ +0.177 $(\text{hct-43})_+$
Odds ratio for each interval between knots can be calculated (Table 4). For example, for hct less than 25, the odds ratio = $e^{0.3147}$ = 1.37. This is interpreted as the following: for patients with hct less than 28, increasing one unit of hct is associated with 1.37-fold increase in odds of blood transfusion. For hct between 28 and 36, odds ratio= $e^{0.3147-0.5957}$= 0.75. This is interpreted as the following: for patients with hct between 28 and 36, increasing one unit of hct is associated with 25% decrease of odds of blood transfusion.

To calculate the confidence interval for each spline term, we need to calculate the variance in each interval between the knots (Table 3B). COVOUT and OUTEST in PROC Logistic output the covariance between the variables as above. Or option covb in the MODEL statement can be used as well.

We then used Delta method to calculate the standard error (sigma) in each interval (less than 25, 25-36, 36-43, greater than 43) from variance and covariance of each interval estimate.

For the Wald confidence interval (CI) for each odds ratio, we could then calculate using the following equation. For example, upper CI=exp (beta+1.96*sigma), lower CI =exp (beta-1.96*sigma).

Table 4. Odds ratios and confidence interval in each hematocrit range.

Hematocrit range	beta	sigma	Odds ratio	Lower 95 %CL	Upper 95 %CL
HCT <25	0.315	0.033	1.370	1.284	1.461
HCT (25-36)	-0.281	0.010	0.755	0.740	0.770
HCT (36-43)	-0.208	0.007	0.812	0.801	0.824
HCT (>43)	-0.031	0.016	0.969	0.940	1.000

SUMMARY

Our results show a non-linear relationship between hematocrit and log odds of transfusion risk in cardiac surgery. This study compares different strategies for handling non-linearity using SAS. We demonstrated the use of linear splines and its interpretation in a logistic regression model.

REFERENCES

1. Harrell, F.E., *Regression modeling strategies : with applications to linear models, logistic regression, and survival analysis*. Springer series in statistics. 2001, New York: Springer. xxii, 568 p.
2. Steyerberg, E.W., *Clinical Prediction Models*. 2009.
3. Croxford, R., *Restricted Cubic Spline Regression : A Brief Introduction.* SAS Paper 5621-2016, 2016.
4. Austin, P.C. and L.J. Brunner, *Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses.* Statistics in Medicine, 2004. **23**(7): p. 1159-1178.
5. Altman, D.G. and P. Royston, *The cost of dichotomising continuous variables.* BMJ, 2006. **332**(7549): p. 1080.
6. Jonas V. Bilenas, N.H., *Using Regression Splines in SAS® STAT Procedures.* Paper BF-140 SAS.
7. Likosky, D.S., et al., *Prediction of Transfusions After Isolated Coronary Artery Bypass Grafting Surgical Procedures.* Annals of Thoracic Surgery, 2017. **103**(3): p. 764-772.

8. Cai, W., *Fitting Generalized Additive Models with the GAM Procedure in SAS 9.2.* SAS Global Forum 2008, 2008.

ACKNOWLEDGMENTS

Thanks to the support from The Michigan Society of Thoracic and Cardiovascular Surgeons Quality Collaborative.

CONTACT INFORMATION

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

Xiaoting Wu (Ting), PhD, MS
Department of Cardiac Surgery
1500 E Medical Center Drive
Ann Arbor, MI 48109
734.936.7731
xiaotinw@med.umich.edu

Supplemental Codes:

```
**********part 1.1 non-linear diagnostic with LOESS**************;
ods html;
ods graphics on;
proc loess data=spline_ex plots (MAXPOINTS=NONE)= fitplot ;
  model rbc = hct_;
run;
ods graphics off;
ods html close;

**********part 1.2 non-linear diagnostic with GAM **************;
ods graphics on;
proc gam data=spline_ex descending plots=all ;
class female   race3c   status3c;
model rbc (event='1')=    param (female   status3c) spline(hct_, df=4)
/dist=binomial;
output out=predictgam p UCLM LCLM;
run;
ods graphics off;
```

```
**********part 1.3 group data plot **************;

%macro pct(dsn,var,pctvar);
/* calculate the cutpoints for the percentiles */
proc univariate noprint data=&dsn;
  var &var;
  output out=pct pctlpts=1 5 10 20 30 40 50 60 70 80 90 95 99 pctlpre=pct;
run;
/* write the quintiles to macro variables */
data _null_;
set pct;
call symput('q0',pct1) ;
call symput('q1',pct5) ;
call symput('q2',pct10) ;
call symput('q3',pct20) ;
call symput('q4',pct30) ;
call symput('q5',pct40) ;
call symput('q6',pct50) ;
call symput('q7',pct60) ;
call symput('q8',pct70) ;
call symput('q9',pct80) ;
call symput('q10',pct90) ;
call symput('q11',pct95) ;
call symput('q12',pct99) ;
run;

/* create the new variable in the main dataset */
data &dsn;
set &dsn;
      if &var =. then &pctvar = .;
  else if &var le &q0 then &pctvar =0;
  else if &var le &q1 then &pctvar =1;
  else if &var le &q2 then &pctvar =2;
  else if &var le &q3 then &pctvar =3;
  else if &var le &q4 then &pctvar =4;
  else if &var le &q5 then &pctvar =5;
  else if &var le &q6 then &pctvar =6;
  else if &var le &q7 then &pctvar =7;
  else if &var le &q8 then &pctvar =8;
  else if &var le &q9 then &pctvar =9;
    else if &var le &q10 then &pctvar =10;
           else if &var le &q11 then &pctvar =11;
  else &pctvar =12;
run;
%mend pct;

%quint (spline_ex, hct_, pcthct);
%pct (spline_ex, hct_, pcthct);
```

```
proc logistic data=spline_ex outmodel=par;
class female      status3c pcthct (ref='12');
model rbc (event='1')=   female    status3c  pcthct/lackfit;
run;

proc means data=spline_ex mean maxdec=1 ; var hct_; class pcthct;run;

*The estimates were obtained from the logistic model above, and plot against the
means in each percentile of HCT;
data estimate_plot;
  input hct_cat mean estimate ;
   datalines;
0  22.0  1.77
1  27.8  2.56
2  31.1  1.58
3  33.5  0.92
4  35.5  0.33
5  37.5  -0.0031
6  38.9  -0.40
7  39.9  -0.68
8  41.4  -0.91
9  42.9  -1.12
10 44.3  -1.30
11 45.9  -1.30
12 48.4  -1.43
;


proc sgplot data=estimate_plot;
 scatter x=mean y= estimate /datalabel=mean;
 series x=mean y=estimate/ lineattrs=(color=red thickness=1);
 keylegend / across = 1 location=inside position=topleft;
run;
quit;
```