**7.1.1 Generalized Additive Models**

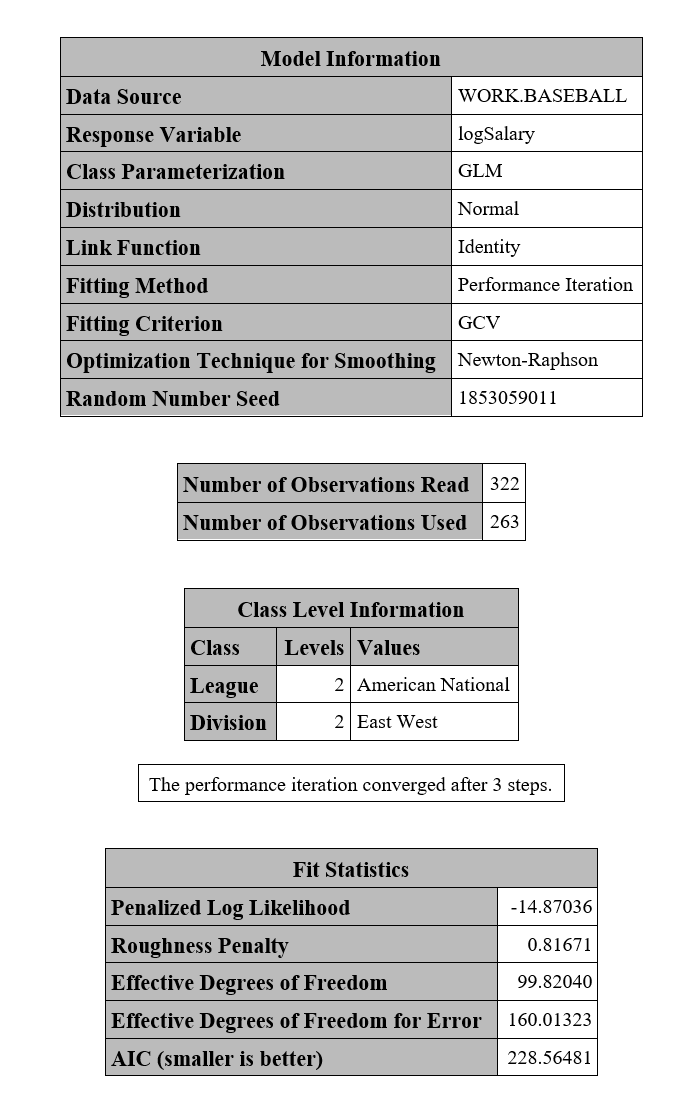
**Stat 5100 – Dr. Bean**

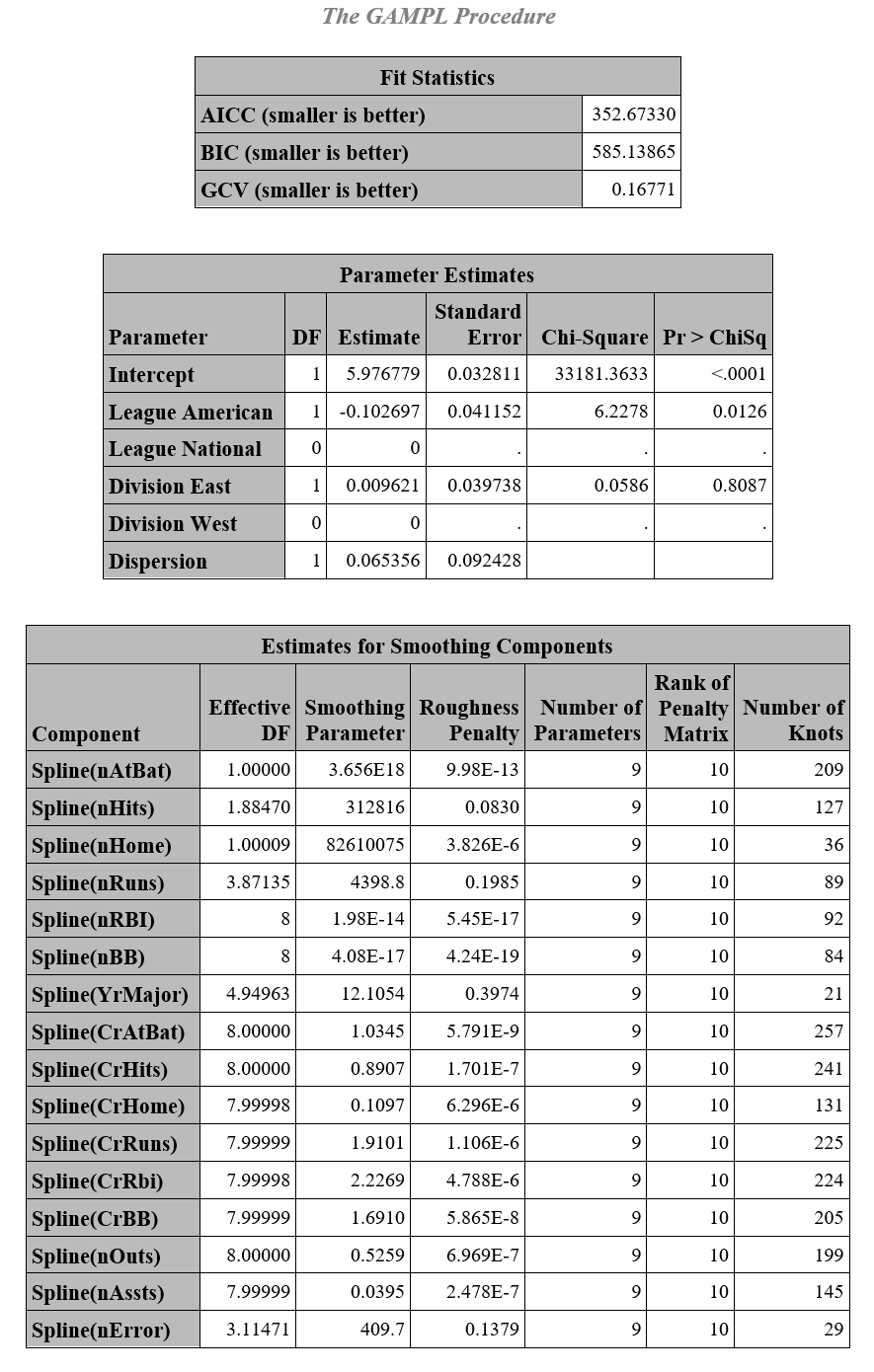
**Baseball Dataset (4.1.1)**

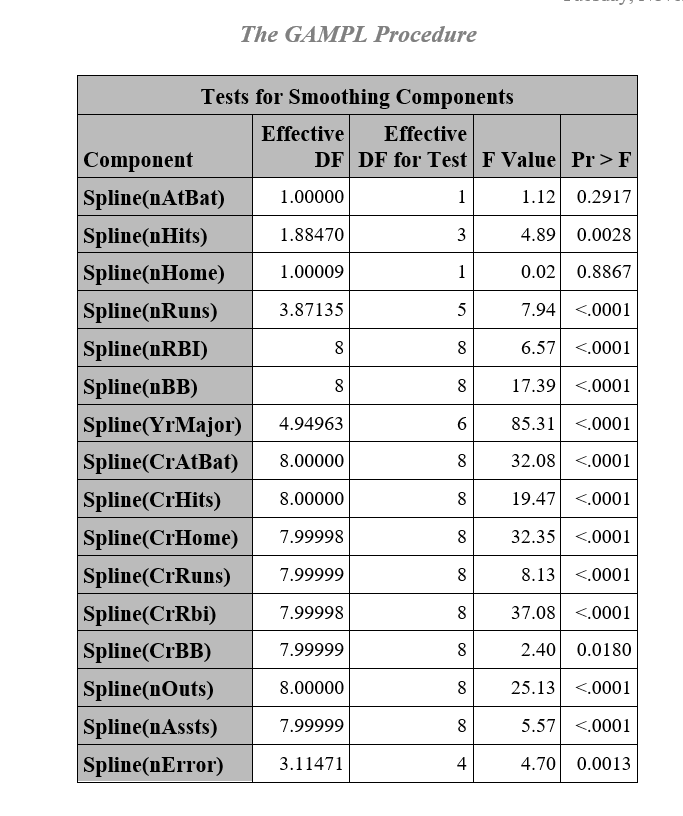
See if we can improve upon the penalized linear regression model to predict the log of salary for professional (non-pitcher) baseball players. Note that answers will differ slightly depending on the seed.

data baseball; set sashelp.baseball;  
AmerLg = (League="American");  
EastDv = (Division="East");  
run;

/\* s() indicates a smoothing spline is fit to the effect \*/  
proc gampl data = baseball seed=12345;   
class league division;  
model logSalary = s(nAtBat) s(nHits) s(nHome) s(nRuns) s(nRBI) s(nBB)  
 s(yrMajor) s(crAtBat) s(crHits) s(crHome) s(crRuns) s(crRbi)  
 s(crBB) s(nOuts) s(nAssts) s(nError) param(league division);  
run;

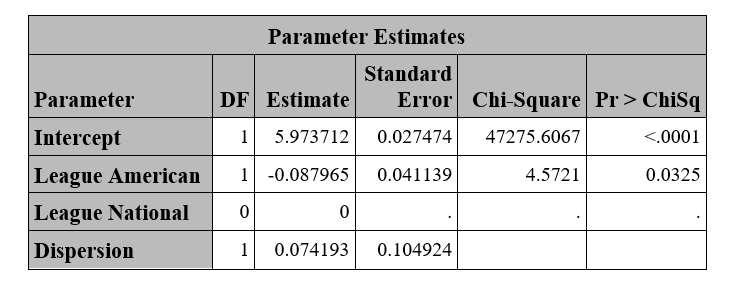


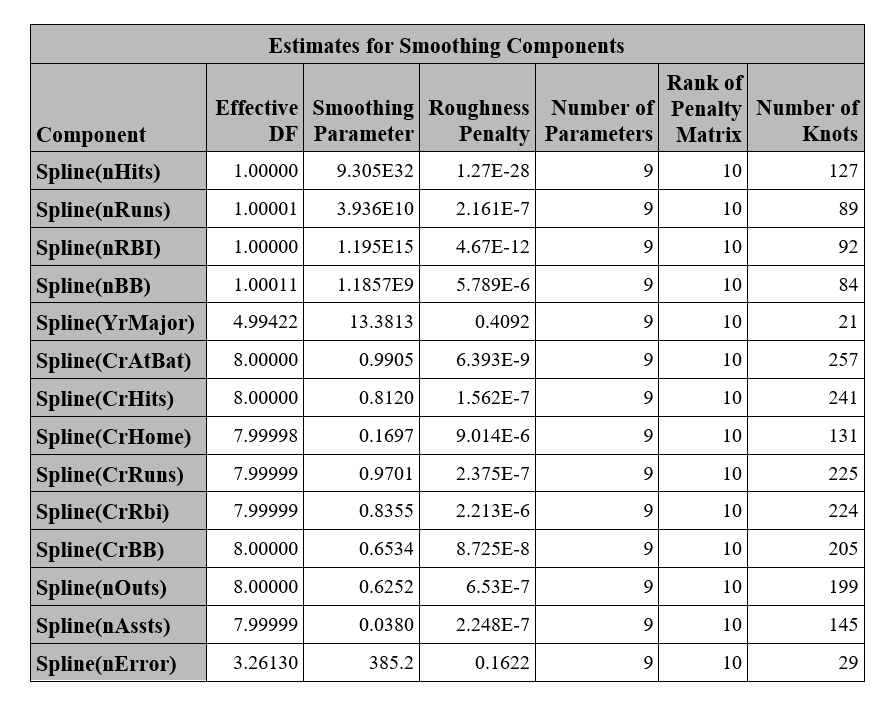


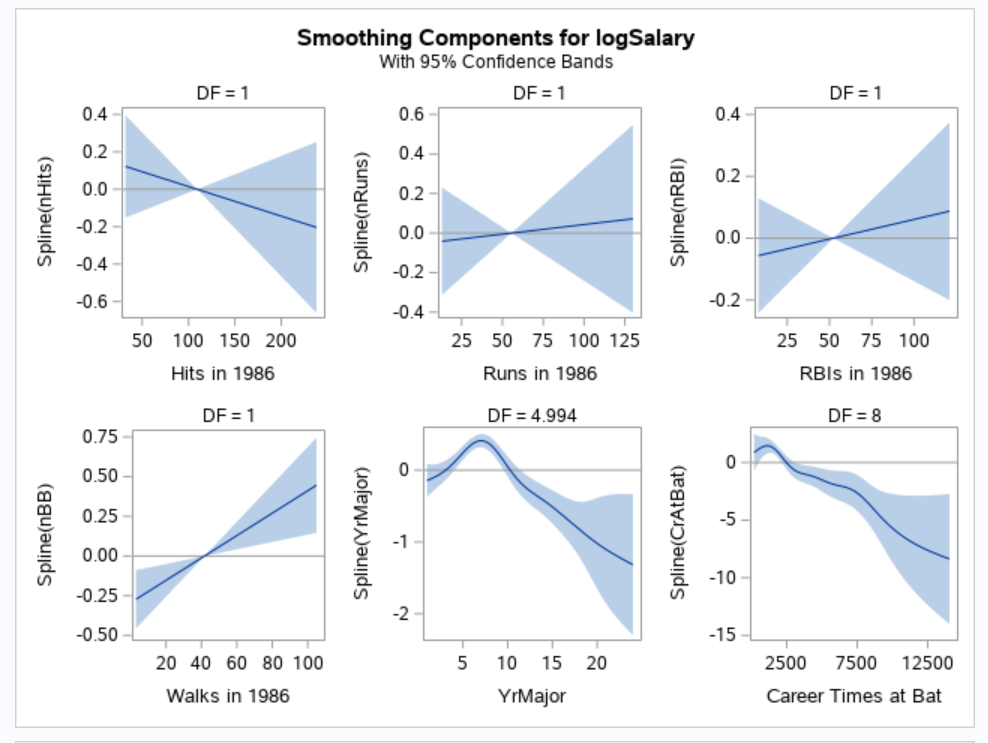


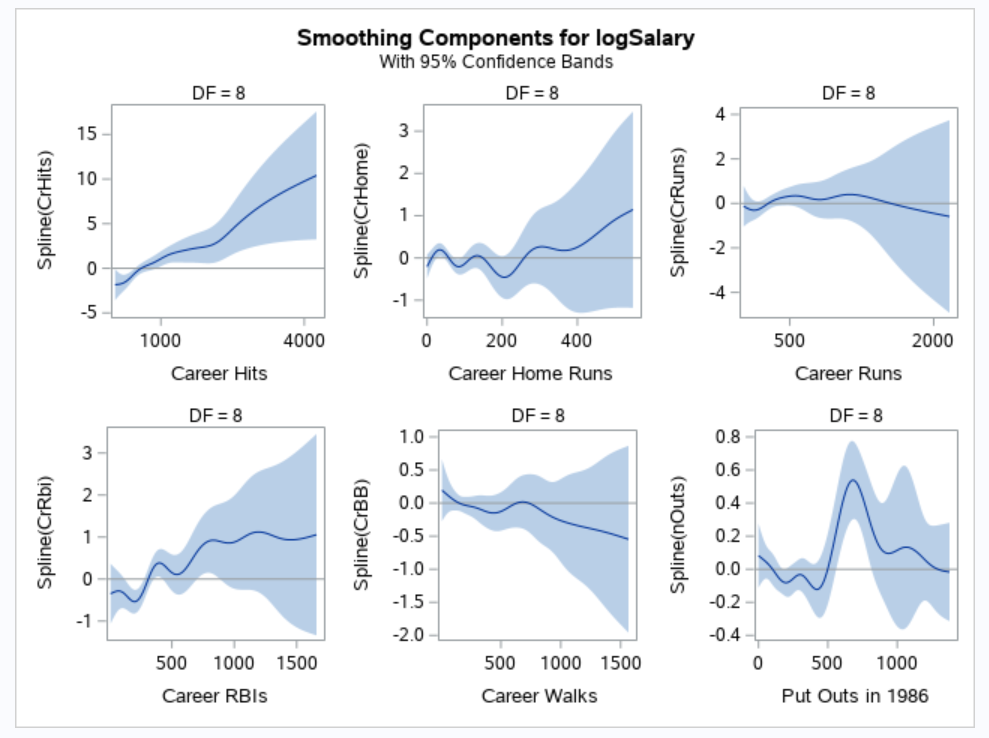
Refit the models using only the significant terms.

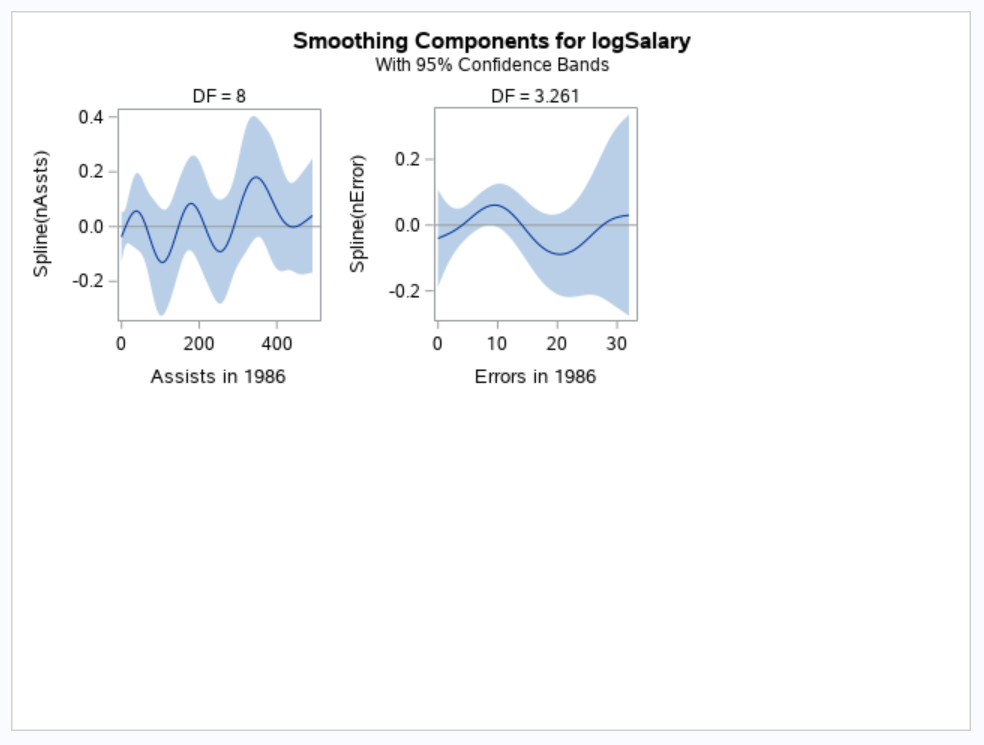
proc gampl data = baseball plots(unpack) = all seed=12345;   
class league;  
model logSalary = s(nHits) s(nRuns) s(nRBI) s(nBB)  
 s(yrMajor) s(crAtBat) s(crHits) s(crHome) s(crRuns) s(crRbi)  
 s(crBB) s(nOuts) s(nAssts) s(nError) param(league);  
run;











/\* Next Example for GAMs for Logistic Regression, taken directly from   
SAS documentation \*/  
  
/\*-----------------------------------------------------------------  
 S A S S A M P L E L I B R A R Y   
   
 NAME: hpgamex2   
 TITLE: Example 2 for PROC GAMPL   
 DESC: Pima Indians Diabetes data set   
 REF: Lim, Loh and Shih (2000)   
 PRODUCT: STAT   
 SYSTEM: ALL   
 KEYS:   
 PROCS: GAMPL   
   
 SUPPORT: Weijie Cai   
-----------------------------------------------------------------\*/  
  
title 'Diabetes Study';  
data DiabetesStudy;  
 input NPreg Glucose Pressure Triceps BMI Pedigree Age Diabetes Test@@;  
 datalines;  
 6 148 72 35 33.6 0.627 50 1 1 1 85 66 29 26.6 0.351 31 0 1  
 1 89 66 23 28.1 0.167 21 0 0 3 78 50 32 31 0.248 26 1 0  
 2 197 70 45 30.5 0.158 53 1 0 5 166 72 19 25.8 0.587 51 1 1  
 0 118 84 47 45.8 0.551 31 1 1 1 103 30 38 43.3 0.183 33 0 1

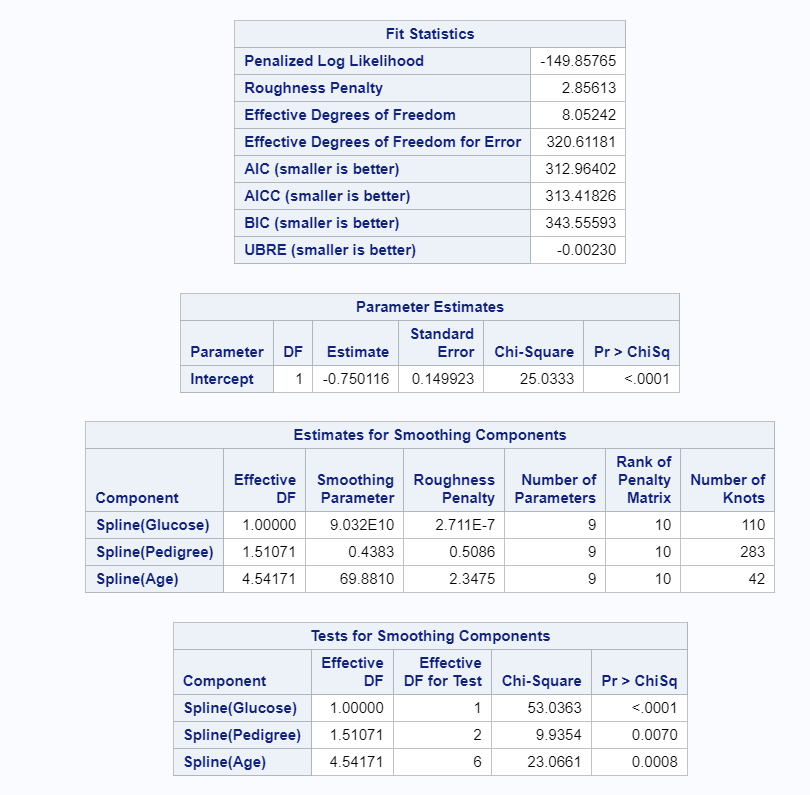
**…**

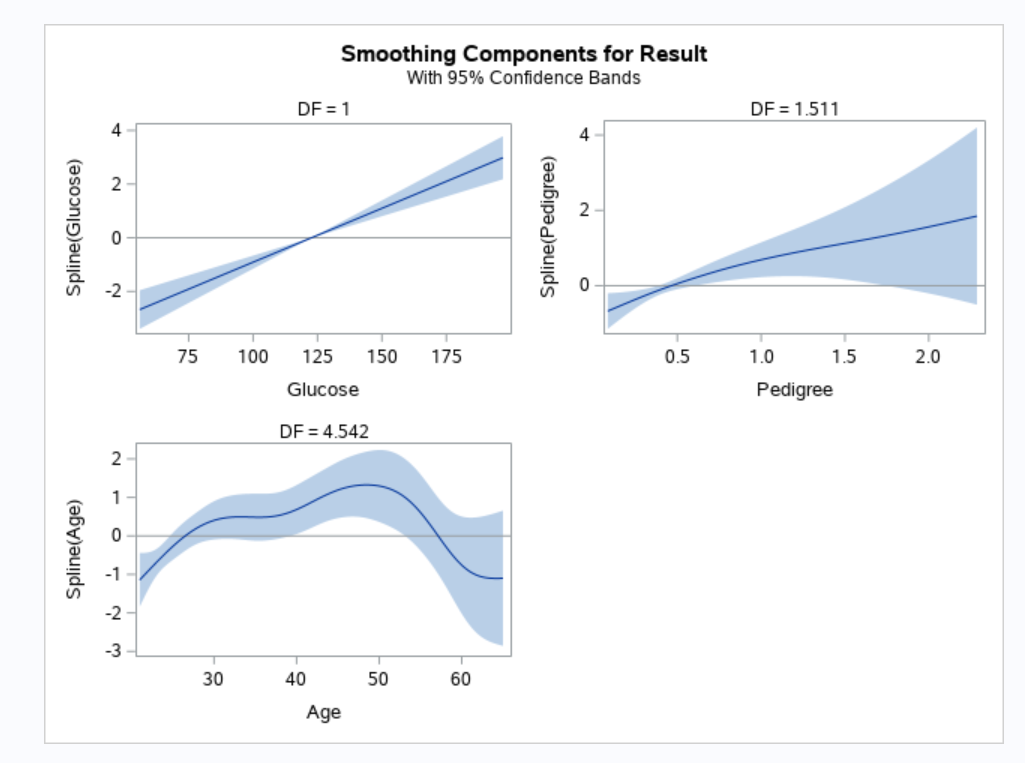
**;**

**Run;**

data DiabetesStudy;  
 set DiabetesStudy;  
 Result = Diabetes;  
 if Test=1 then Result=.;  
run;

ods graphics on;  
proc gampl data=DiabetesStudy plots seed=12345;  
 model Result(event='1') = spline(Glucose)  
 spline(Pedigree) spline(Age) / dist=binary;  
 output out=DiabetesStudyOut;  
 id Diabetes Test;  
run;





**Now, lets see how accurate we are on a test set.**

data test;  
 set DiabetesStudyOut(where=(Test=1));  
 if ((Pred>0.5 & Diabetes=1) | (Pred<0.5 & Diabetes=0))  
 then Error=0;  
 else Error=1;  
run;  
  
proc freq data=test;  
 tables Diabetes\*Error/nocol norow;  
run;

