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;
```

Model Information	
Data Source	WORK.BASEBALL
Response Variable	logSalary
Class Parameterization	GLM
Distribution	Normal
Link Function	Identity
Fitting Method	Performance Iteration
Fitting Criterion	GCV
Optimization Technique for Smoothing	Newton-Raphson
Random Number Seed	1853059011

Number of Observations Read	322
Number of Observations Used	263

Class Level Information		
Class	Levels	Values
League	2	American National
Division	2	East West

The performance iteration converged after 3 steps.

Fit Statistics	
Penalized Log Likelihood	-14.87036
Roughness Penalty	0.81671
Effective Degrees of Freedom	99.82040
Effective Degrees of Freedom for Error	160.01323
AIC (smaller is better)	228.56481

The GAMPL Procedure

Fit Statistics	
AICC (smaller is better)	352.67330
BIC (smaller is better)	585.13865
GCV (smaller is better)	0.16771

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	5.976779	0.032811	33181.3633	<.0001
League American	1	-0.102697	0.041152	6.2278	0.0126
League National	0	0	.	.	.
Division East	1	0.009621	0.039738	0.0586	0.8087
Division West	0	0	.	.	.
Dispersion	1	0.065356	0.092428		

Estimates for Smoothing Components						
Component	Effective DF	Smoothing Parameter	Roughness Penalty	Number of Parameters	Rank of Penalty Matrix	Number of Knots
Spline(nAtBat)	1.00000	3.656E18	9.98E-13	9	10	209
Spline(nHits)	1.88470	312816	0.0830	9	10	127
Spline(nHome)	1.00009	82610075	3.826E-6	9	10	36
Spline(nRuns)	3.87135	4398.8	0.1985	9	10	89
Spline(nRBI)	8	1.98E-14	5.45E-17	9	10	92
Spline(nBB)	8	4.08E-17	4.24E-19	9	10	84
Spline(YrMajor)	4.94963	12.1054	0.3974	9	10	21
Spline(CrAtBat)	8.00000	1.0345	5.791E-9	9	10	257
Spline(CrHits)	8.00000	0.8907	1.701E-7	9	10	241
Spline(CrHome)	7.99998	0.1097	6.296E-6	9	10	131
Spline(CrRuns)	7.99999	1.9101	1.106E-6	9	10	225
Spline(CrRbi)	7.99998	2.2269	4.788E-6	9	10	224
Spline(CrBB)	7.99999	1.6910	5.865E-8	9	10	205
Spline(nOuts)	8.00000	0.5259	6.969E-7	9	10	199
Spline(nAssts)	7.99999	0.0395	2.478E-7	9	10	145
Spline(nError)	3.11471	409.7	0.1379	9	10	29

The GAMPL Procedure

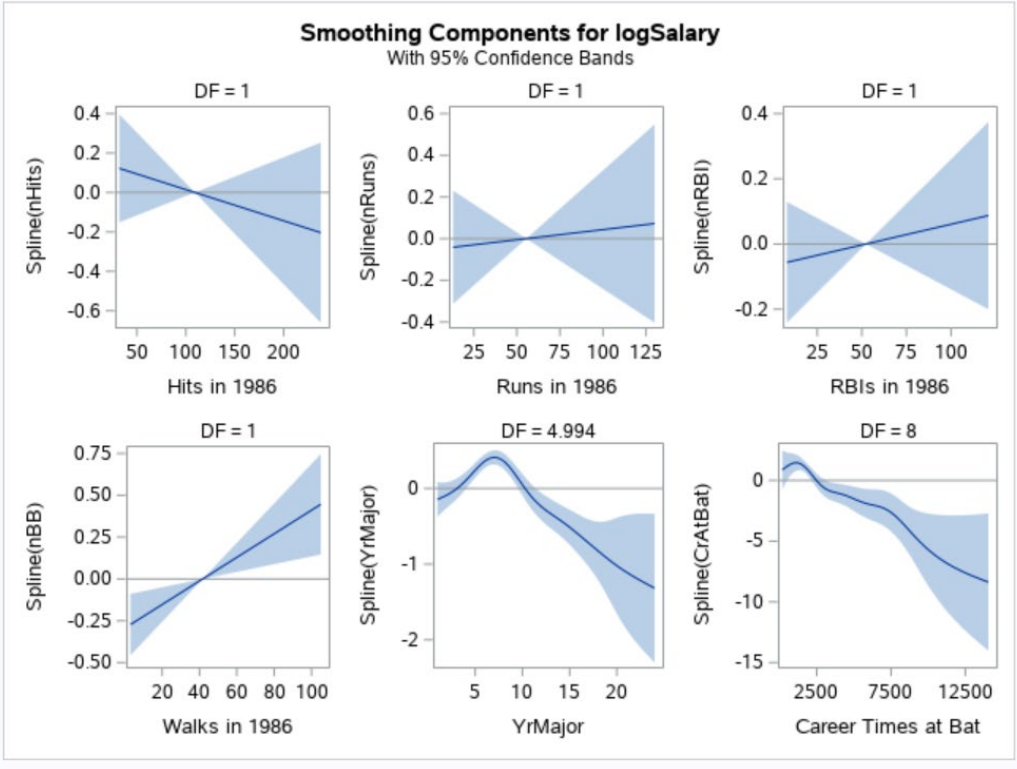
Tests for Smoothing Components				
Component	Effective DF	Effective DF for Test	F Value	Pr > F
Spline(nAtBat)	1.00000	1	1.12	0.2917
Spline(nHits)	1.88470	3	4.89	0.0028
Spline(nHome)	1.00009	1	0.02	0.8867
Spline(nRuns)	3.87135	5	7.94	<.0001
Spline(nRBI)	8	8	6.57	<.0001
Spline(nBB)	8	8	17.39	<.0001
Spline(YrMajor)	4.94963	6	85.31	<.0001
Spline(CrAtBat)	8.00000	8	32.08	<.0001
Spline(CrHits)	8.00000	8	19.47	<.0001
Spline(CrHome)	7.99998	8	32.35	<.0001
Spline(CrRuns)	7.99999	8	8.13	<.0001
Spline(CrRbi)	7.99998	8	37.08	<.0001
Spline(CrBB)	7.99999	8	2.40	0.0180
Spline(nOuts)	8.00000	8	25.13	<.0001
Spline(nAssts)	7.99999	8	5.57	<.0001
Spline(nError)	3.11471	4	4.70	0.0013

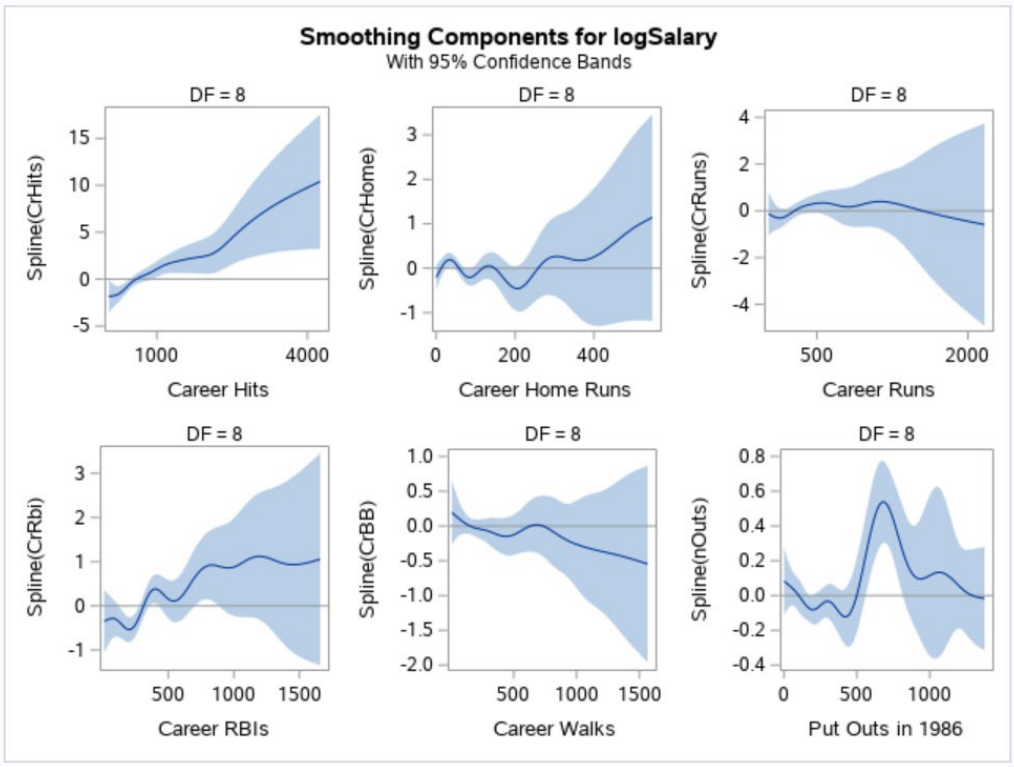
Refit the models using only the significant terms.

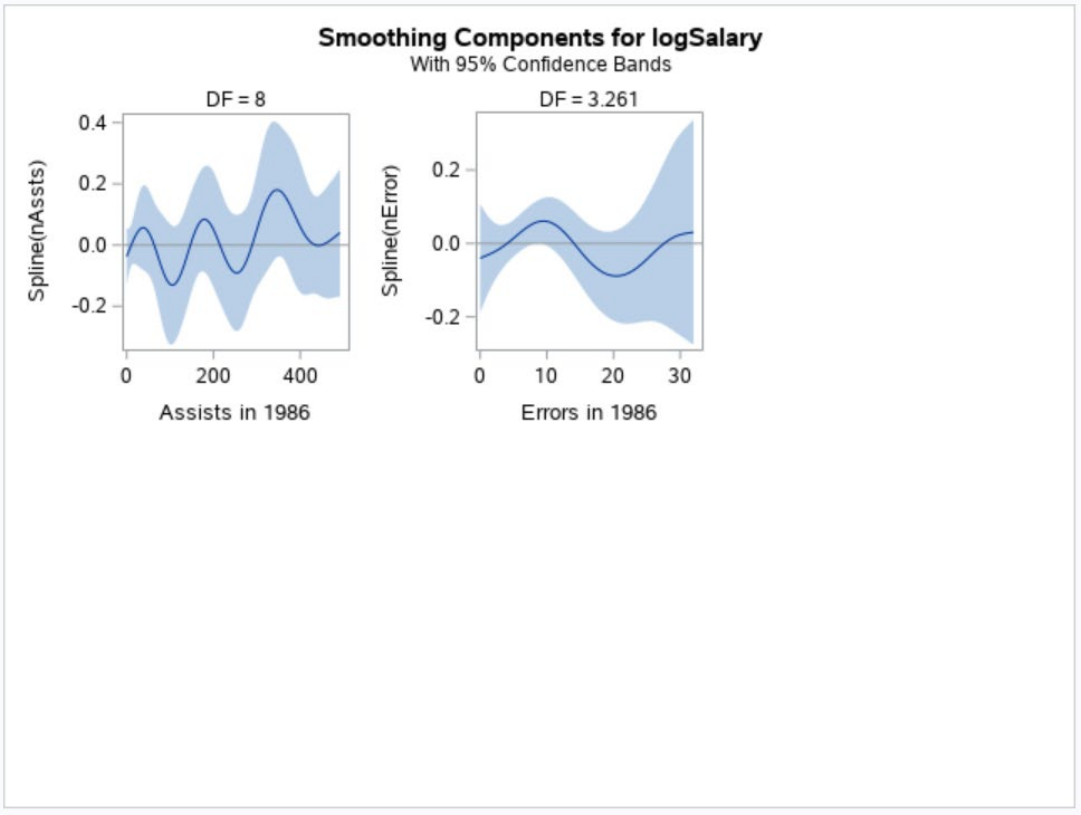
```
proc gampl data = baseball plots(unpack) = all seed=12345;
class league;
model logSalary = s(nHits) s(nRuns) s(nRBI) s(nBB)
 s(yrMajor)  s(crAtBat)  s(crHits)  s(crHome)  s(crRuns)  s(crRbi)
 s(crBB) s(nOuts) s(nAssts) s(nError) param(league);
run;
```

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	5.973712	0.027474	47275.6067	<.0001
League American	1	-0.087965	0.041139	4.5721	0.0325
League National	0	0	.	.	.
Dispersion	1	0.074193	0.104924		

Estimates for Smoothing Components						
Component	Effective DF	Smoothing Parameter	Roughness Penalty	Number of Parameters	Rank of Penalty Matrix	Number of Knots
Spline(nHits)	1.00000	9.305E32	1.27E-28	9	10	127
Spline(nRuns)	1.00001	3.936E10	2.161E-7	9	10	89
Spline(nRBI)	1.00000	1.195E15	4.67E-12	9	10	92
Spline(nBB)	1.00011	1.1857E9	5.789E-6	9	10	84
Spline(YrMajor)	4.99422	13.3813	0.4092	9	10	21
Spline(CrAtBat)	8.00000	0.9905	6.393E-9	9	10	257
Spline(CrHits)	8.00000	0.8120	1.562E-7	9	10	241
Spline(CrHome)	7.99998	0.1697	9.014E-6	9	10	131
Spline(CrRuns)	7.99999	0.9701	2.375E-7	9	10	225
Spline(CrRbi)	7.99999	0.8355	2.213E-6	9	10	224
Spline(CrBB)	8.00000	0.6534	8.725E-8	9	10	205
Spline(nOuts)	8.00000	0.6252	6.53E-7	9	10	199
Spline(nAssts)	7.99999	0.0380	2.248E-7	9	10	145
Spline(nError)	3.26130	385.2	0.1622	9	10	29




Smoothing Components for logSalary
With 95% Confidence Bands
DF = 8
Spline(CrHits)
Career Hits
DF = 8
Spline(CrHome)
Career Home Runs
DF = 8
Spline(CrRuns)
Career Runs
DF = 8
Spline(CrRbi)
Career RBIs
DF = 8
Spline(CrBB)
Career Walks
DF = 8
Spline(nOuts)
Put Outs in 1986


Smoothing Components for logSalary
With 95% Confidence Bands
DF = 8
Spline(nAssts)
Assists in 1986
DF = 3.261
Spline(nError)
Errors in 1986

```
/* 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;
```

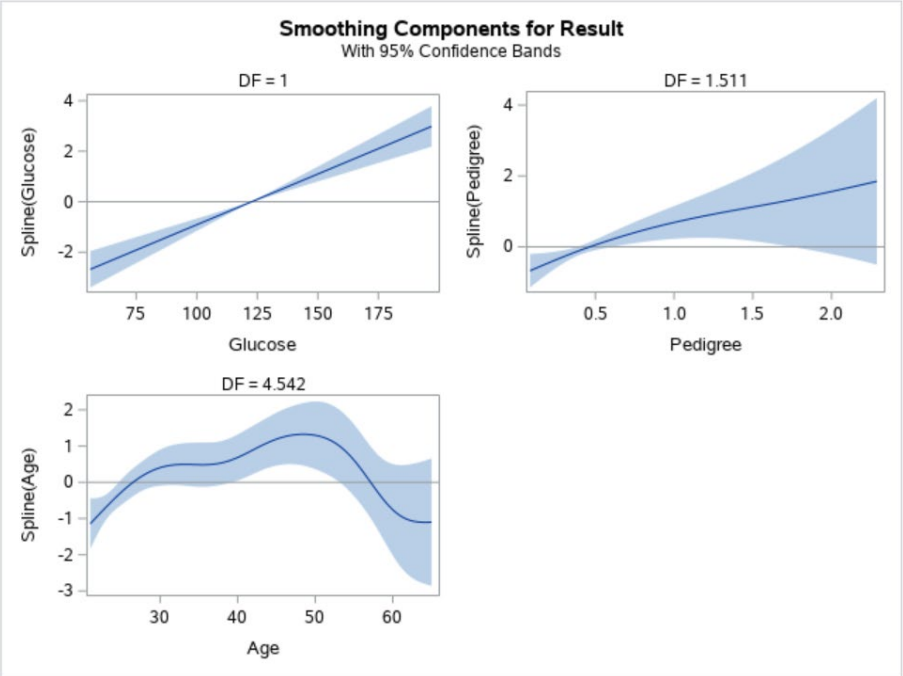
```
ods graphics on;
proc gampl data=DiabetesStudy plots seed=12345;
   model Result(event='1') = spline(Glucose)
                             spline(Pedigree) spline(Age) /
dist=binary;
   output out=DiabetesStudyOut;
   id Diabetes Test;
run;
```

Fit Statistics	
Penalized Log Likelihood	-149.85765
Roughness Penalty	2.85613
Effective Degrees of Freedom	8.05242
Effective Degrees of Freedom for Error	320.61181
AIC (smaller is better)	312.96402
AICC (smaller is better)	313.41826
BIC (smaller is better)	343.55593
UBRE (smaller is better)	-0.00230

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-0.750116	0.149923	25.0333	<.0001

Estimates for Smoothing Components						
Component	Effective DF	Smoothing Parameter	Roughness Penalty	Number of Parameters	Rank of Penalty Matrix	Number of Knots
Spline(Glucose)	1.00000	9.032E10	2.711E-7	9	10	110
Spline(Pedigree)	1.51071	0.4383	0.5086	9	10	283
Spline(Age)	4.54171	69.8810	2.3475	9	10	42

Tests for Smoothing Components				
Component	Effective DF	Effective DF for Test	Chi-Square	Pr > ChiSq
Spline(Glucose)	1.00000	1	53.0363	<.0001
Spline(Pedigree)	1.51071	2	9.9354	0.0070
Spline(Age)	4.54171	6	23.0661	0.0008



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;
```

The FREQ Procedure

Frequency
Percent

Table of Diabetes by Error

Diabetes	Error 0	Error 1	Total
0	130 64.36	17 8.42	147 72.77
1	35 17.33	20 9.90	55 27.23
Total	165 81.68	37 18.32	202 100.00