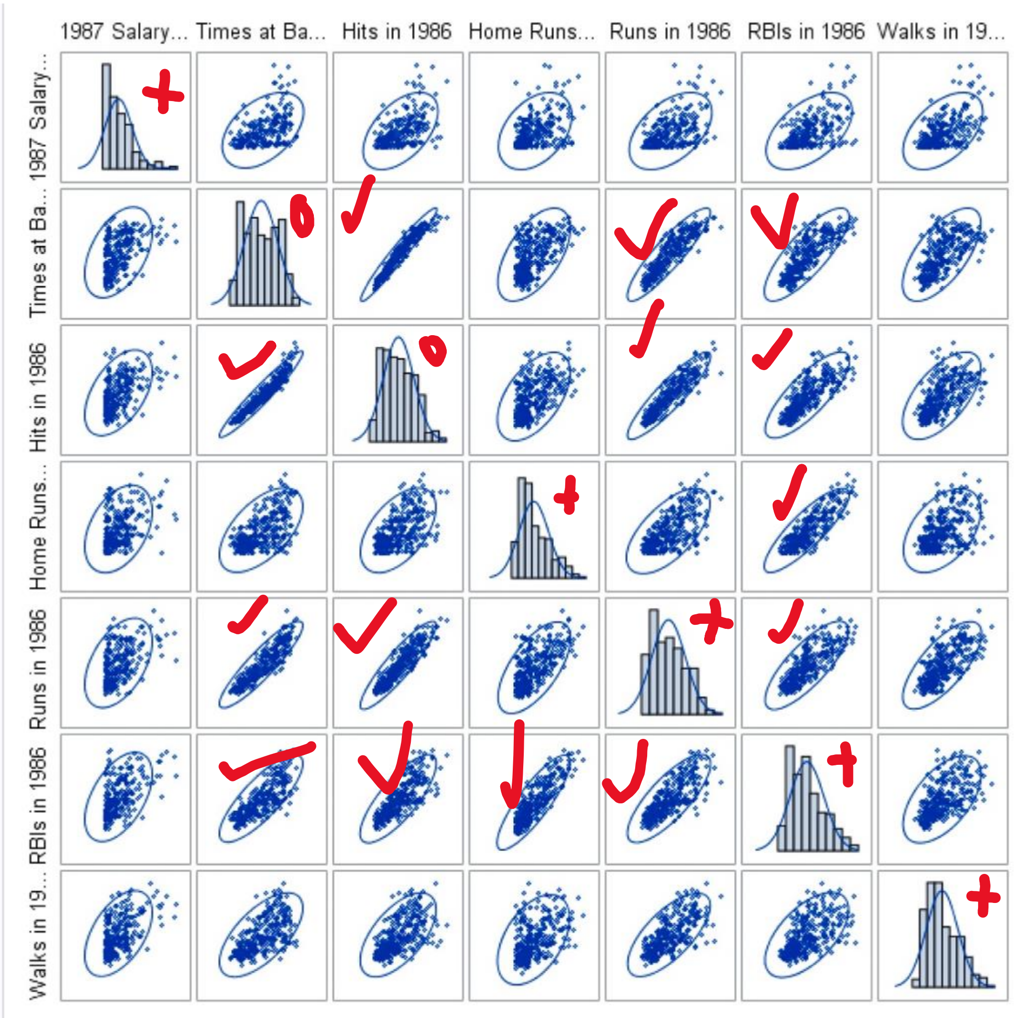
MATH 312

FINAL PROJECT

Hayden Trautmann

Senior at Lehigh University

1)



Looking at the scatter matrices above, we can see that the plots which appear most linear with minimal outliers have the strongest correlation. We designated the strongly correlated scatter matrices by adding a check mark in the box. We can observe the strongest correlation is between X1 and X2, Times at Bat and Hits and 1986. We will also be able to observe this strong correlation between these two variables later when we generate the Pearson Correlation Coefficients.

We also designated the relatively symmetric distributions in the diagonals above with a blue circle, and the non-symmetric distributions with a + for right-skewed and a dash for left-skewed. This same labeling process is used in the box plots below.

Just as with the scatter matrices, we see a many of the numerical variables with skew to the right, and only two numerical variables which appear to have a symmetric distribution. These observations are consistent with our observations from the scatter matrices.

Frequency Tables for Visuals

Computing Frequencies and Percentages Using PROC FREQ

Team at the End of 1986

Team	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Atlanta	11	3.42	11	3.42
Baltimore	15	4.66	26	8.07
Boston	10	3.11	36	11.18
California	13	4.04	49	15.22
Chicago	24	7.45	73	22.67
Cincinnati	12	3.73	85	26.40
Cleveland	12	3.73	97	30.12
Detroit	12	3.73	109	33.85
Houston	11	3.42	120	37.27
Kansas City	14	4.35	134	41.61
Los Angeles	14	4.35	148	45.96
Milwaukee	14	4.35	162	50.31
Minneapolis	13	4.04	175	54.35
Montreal	14	4.35	189	58.70
New York	24	7.45	213	66.15
Oakland	12	3.73	225	69.88
Philadelphia	12	3.73	237	73.60
Pittsburgh	11	3.42	248	77.02
San Diego	13	4.04	261	81.06
San Francisco	14	4.35	275	85.40
Seattle	12	3.73	287	89.13
St Louis	11	3.42	298	92.55
Texas	13	4.04	311	96.58
Toronto	11	3.42	322	100.00

Position(s) in 1986

Position	Frequency	Percent	Cumulative Frequency	Cumulative Percent
13	1	0.31	1	0.31
1B	31	9.63	32	9.94
1O	1	0.31	33	10.25
23	1	0.31	34	10.56
2B	31	9.63	65	20.19
2S	1	0.31	66	20.50
32	1	0.31	67	20.81
3B	32	9.94	99	30.75
3O	1	0.31	100	31.06
3S	3	0.93	103	31.99
C	40	12.42	143	44.41
CD	1	0.31	144	44.72
CF	26	8.07	170	52.80
CS	1	0.31	171	53.11
DH	16	4.97	187	58.07
DO	2	0.62	189	58.70
LF	25	7.76	214	66.46
O1	4	1.24	218	67.70
OD	1	0.31	219	68.01
OF	30	9.32	249	77.33
OS	2	0.62	251	77.95
RF	26	8.07	277	86.02
S3	1	0.31	278	86.34
SS	30	9.32	308	95.65
UT	14	4.35	322	100.00

League at the End of 1986

League	Frequency	Percent	Cumulative Frequency	Cumulative Percent
American	175	54.35	175	54.35
National	147	45.65	322	100.00

Division at the End of 1986

Division	Frequency	Percent	Cumulative Frequency	Cumulative Percent
East	157	48.76	157	48.76
West	165	51.24	322	100.00

We can see that the most frequent teams are New York and Chicago, both containing 24 observations. Also, C is the most frequent position, with 40 observations accounting for 12.42% of the data.

2) Correlation Analysis

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	CrRbi	CrBB	nOuts	nAssts	nError	Salary
CrRbi Career RBIs	1.00000 322	0.88500 <.0001 322	0.10088 0.0706 322	-0.09126 0.1021 322	-0.12324 0.0270 322	0.61871 <.0001 263
CrBB Career Walks	0.88500 <.0001 322	1.00000 322	0.04573 0.4134 322	-0.04550 0.4158 322	-0.14027 0.0117 322	0.54574 <.0001 263
nOuts Put Outs in 1986	0.10088 0.0706 322	0.04573 0.4134 322	1.00000 322	-0.02520 0.6523 322	0.10974 0.0491 322	0.30048 <.0001 263
nAssts Assists in 1986	-0.09126 0.1021 322	-0.04550 0.4158 322	-0.02520 0.6523 322	1.00000 322	0.70635 <.0001 322	0.02544 0.6814 263
nError Errors in 1986	-0.12324 0.0270 322	-0.14027 0.0117 322	0.10974 0.0491 322	0.70635 <.0001 322	1.00000 322	-0.00540 0.9305 263

Salary 1987 Salary in $ Thousands	0.61871 <.0001 263	0.54574 <.0001 263	0.30048 <.0001 263	0.02544 0.6814 263	-0.00540 0.9305 263	1.00000 263

The highlighted p – values were below .05 and show strong correlation between the attributes.

Spearman Correlation Statistics (Fisher's z Transformation)							
Variable	With Variable	N	Sample Correlation	Fisher's z	95% Confidence Limits		p Value for H0:Rho=0
CrRbi	CrBB	322	0.93483	1.69536	0.919475	0.947331	<.0001
CrRbi	nOuts	322	0.07157	0.07170	-0.038023	0.179468	0.2004
CrRbi	nAssts	322	-0.04514	-0.04517	-0.153678	0.064479	0.4198
CrRbi	nError	322	-0.06607	-0.06616	-0.174107	0.043547	0.2373
CrRbi	Salary	263	0.79807	1.09327	0.749457	0.838119	<.0001
CrBB	nOuts	322	0.04345	0.04348	-0.066164	0.152026	0.4374
CrBB	nAssts	322	-0.03522	-0.03523	-0.143963	0.074366	0.5292
CrBB	nError	322	-0.09379	-0.09406	-0.201026	0.015671	0.0929
CrBB	Salary	263	0.77176	1.02466	0.717808	0.816495	<.0001
nOuts	nAssts	322	0.15424	0.15548	0.045716	0.259173	0.0055
nOuts	nError	322	0.16830	0.16991	0.060103	0.272581	0.0024
nOuts	Salary	263	0.21112	0.21434	0.092527	0.323809	0.0005
nAssts	nError	322	0.74823	0.96892	0.695835	0.792700	<.0001
nAssts	Salary	263	0.05730	0.05736	-0.064103	0.177027	0.3550
nError	Salary	263	0.01899	0.01899	-0.102205	0.139623	0.7595

The highlighted p – values were below .05 and show strong correlation between the attributes.

Projecting the result of the regression model if it were to be implemented, I would expect it to be inaccurate using all the data together, but if we created a train set and isolated the strongly correlated attributes highlighted above I would expect the regression model to be more accurate.

3) Regression Analysis

Durbin-Watson D	2.064
Pr < DW	0.6914
Pr > DW	0.3086

Number of Observations	263
1st Order Autocorrelation	-0.033

Note: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for testing negative autocorrelation.

No autocorrelation, since looking at the Durbin-Watson D, both probabilities Pr < DW and Pr > DW are not significant. So we can assume they are independent.

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.959539	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.085866	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	0.519185	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	2.924255	**Pr > A-Sq**	<0.0050

D'AGOSTINO TEST OF NORMALITY FOR VARIABLE D, N=263
G1=0.29108 SQRTB1=0.28942 Z= 1.93134 P=0.0534
G2=2.84364 B2=5.76713 Z= 4.67922 P=0.0000
K**2=CHISQ(2 DF)=25.62514 P=0.0000

All the hypothesis tests for normality were below 0.05, so there is a serious violation against normality.

4) Regression Analysis - Full Model

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	16	32389239	2024327	23.79	<.0001
Error	246	20929874	85081		
Corrected Total	262	53319113			

The F – Test was below 0.05, so this shows us it could worthwhile to proceed because there is a significant regression effect, however we already identified there is a serious violation against normality so we will not proceed with the LSE model.

Root MSE	291.68611	**R-Square**	0.6075
Dependent Mean	1.27952E-13	**Adj R-Sq**	0.5819
Coeff Var	2.279658E17		

Adj R-Squared is 0.5819 which indicates that just under 42% of the variability in the data cannot be accounted for in the model. There was a serious violation against normality so I will not report the LSE.

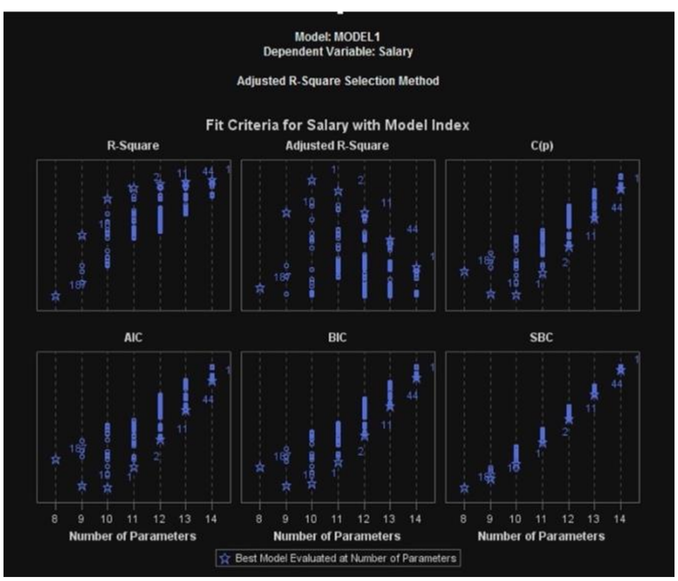
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-15.69078	18.26105	-0.86	0.3910	0
nAtBat	Times at Bat in 1986	1	-1.71718	0.58550	-2.93	0.0037	21.47655
nHits	Hits in 1986	1	7.87708	2.18472	3.61	0.0004	28.44674
nHome	Home Runs in 1986	1	0.78940	5.67692	0.14	0.8895	7.73102
nRuns	Runs in 1986	1	-2.99654	2.74069	-1.09	0.2753	14.54214
nRBI	RBIs in 1986	1	0.20938	2.37304	0.09	0.9298	11.46548
nBB	Walks in 1986	1	6.12442	1.66925	3.67	0.0003	3.96894
YrMajor	Years in the Major Leagues	1	5.04190	11.43257	0.44	0.6596	9.23684
CrAtBat	Career Times at Bat	1	-0.16163	0.12437	-1.30	0.1950	249.85140
CrHits	Career Hits	1	-0.02058	0.61915	-0.03	0.9735	497.07282
CrHome	Career Home Runs	1	-0.07315	1.48200	-0.05	0.9607	50.06939
CrRuns	Career Runs	1	1.55888	0.68290	2.28	0.0233	161.01942
CrRbi	Career RBIs	1	0.71968	0.63692	1.13	0.2596	134.74454
CrBB	Career Walks	1	-0.66416	0.30050	-2.21	0.0280	20.47714
nOuts	Put Outs in 1986	1	0.25382	0.07216	3.52	0.0005	1.25638
nAssts	Assists in 1986	1	0.18212	0.20472	0.89	0.3745	2.71651
nError	Errors in 1986	1	-1.42081	4.04170	-0.35	0.7255	2.19559

Not all the Variance inflations are greater than 10, so all the attributes plotted together do not show serious multicollinearity.

Collinearity Diagnostics									
			Proportion of Variation						
Number	Eigenvalue	Condition Index	Intercept	nAtBat	nHits	nHome	nRuns	nRBI	nBB
1	7.19981	1.00000	0.00001542	0.00022740	0.00016867	0.00074988	0.00034185	0.00064749	0.00149
2	4.18796	1.31117	0.00203	0.00169	0.00122	0.00164	0.00234	0.00203	0.00328
3	1.71776	2.04729	0.00162	0.00011851	0.00004212	0.00865	0.00019793	0.00151	0.00077083
4	0.98589	2.70238	0.84908	0.00006500	0.00002143	0.00055146	0.00001055	0.00005375	0.00012030
5	0.85768	2.89732	0.14037	0.00012450	0.00013637	0.00539	0.00114	0.00164	0.00101
6	0.68778	3.23546	0.00063711	0.00067675	0.00084344	0.04326	0.00484	0.01163	0.12487
7	0.53112	3.68183	0.00021398	0.00666	0.01018	0.01305	0.00223	0.00001396	0.16986
8	0.25387	5.32539	0.00015767	8.559946E-8	0.00022726	0.00467	0.00719	0.00422	0.00305
9	0.18269	6.27768	0.00192	0.00528	0.00351	0.08470	0.00046736	0.01607	0.03178
10	0.13105	7.41201	0.00030540	0.00789	0.00327	0.13586	0.12629	0.13840	0.13617
11	0.09631	8.64609	0.00251	0.00070928	0.00011771	0.00109	0.08093	0.11697	0.00148
12	0.06246	10.73601	3.486237E-7	0.00437	0.02896	0.04430	0.01488	0.04996	0.27955
13	0.05609	11.32930	0.00035202	0.30304	0.00106	0.28212	0.30598	0.32323	0.02755
14	0.02889	15.78737	0.00027127	0.41953	0.61678	0.19451	0.10794	0.17624	0.09339
15	0.01462	22.19211	0.00001517	0.04227	0.05531	0.01199	0.11406	0.04626	0.05316
16	0.00480	38.70918	0.00002878	0.10855	0.01497	0.09955	0.03109	0.08278	0.00117
17	0.00119	77.78890	0.00047828	0.09880	0.26317	0.06791	0.20009	0.02833	0.07129

The cutoff for the condition index is about 30, so for that number if the condition number exceeded 30 then we know there is serious multicollinearity that needs to be dealt with.



The best model evaluate appears to be SBC since it is the most linear with tightly fit data points.

Model Index	Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	BIC	SBC	Variables in Model
1	**9**	0.5917	0.6057	4.0919	2988.9973	2992.2766	3024.71880	nAtBat nHits nRuns nBB CrAtBat C CrRbi CrBB nOuts
2	**10**	0.5911	0.6067	5.4704	2990.3349	2993.8013	3029.62861	nAtBat nHits nRuns nBB CrAtBat C CrRbi CrBB nOuts nAssts
3	**10**	0.5904	0.6060	5.9173	2990.8113	2994.2358	3030.10502	nAtBat nHits nRuns nBB YrMajor C CrRuns CrRbi CrBB nOuts

This table shows the top three performing models from Adjusted R2 , AIC, BIC, SBC, and Cp.

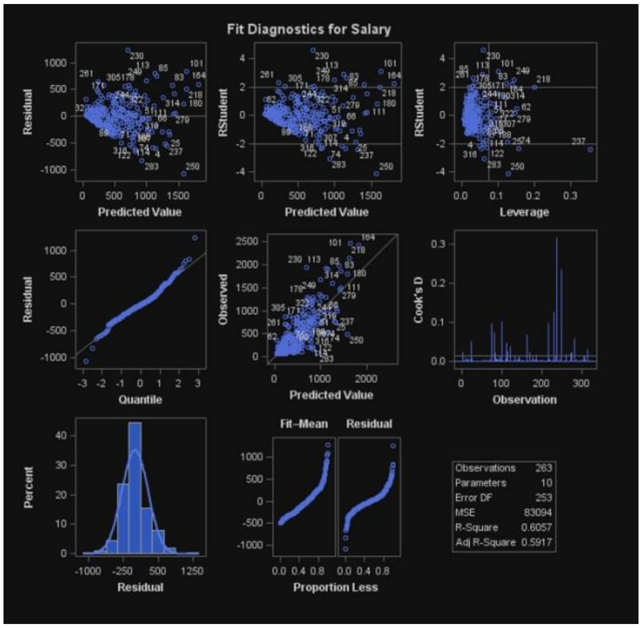
LASSO Selection Summary				
Step	**Effect Entered**	**Effect Removed**	**Number Effects In**	**CV PRESS**
0	**Intercept**		1	53918978.6
1	**CrRbi**		2	33793555.5

This table shows the top performing model from Group Lasso.

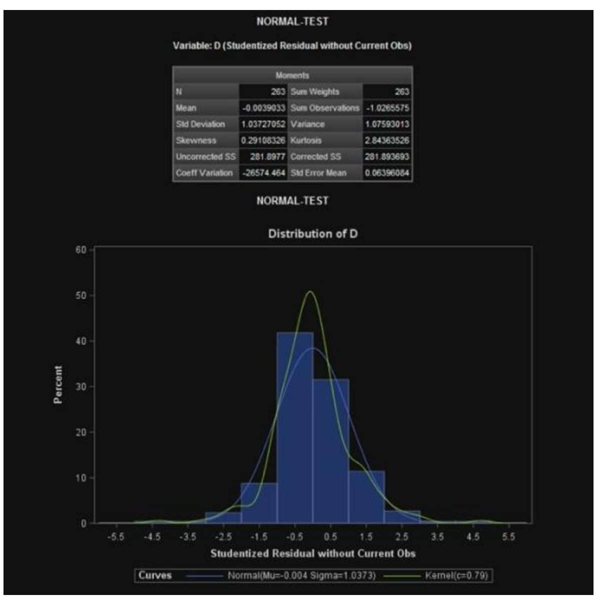
Elastic Net Selection Summary				
Step	**Effect Entered**	**Effect Removed**	**Number Effects In**	**CV PRESS**
0	**Intercept**		1	53784858.1
1	**CrRbi**		2	33559881.4

This table shows the top performing model from Elastic Net

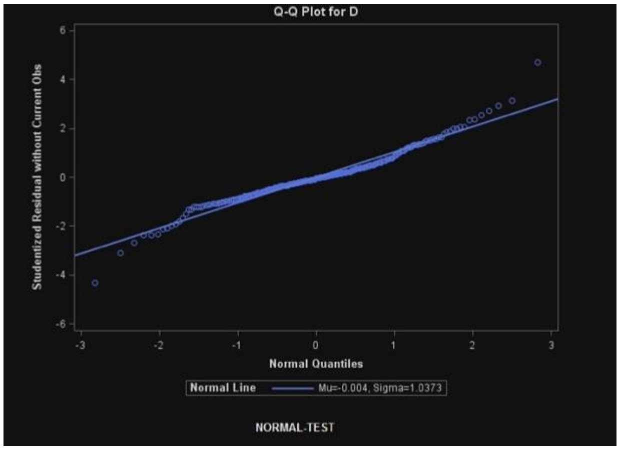
Fit these top models to test set – After hours of trial and error I could not figure out how to do this in SAS



There does not appear to be constant variance after looking at the Homoscedastic Plot. Many observations fall outside the designated cutoffs at 2 and -2. Looking at the QQ-plot there appears the scatters falling along a relatively straight line, so there is no serious violation against normality.

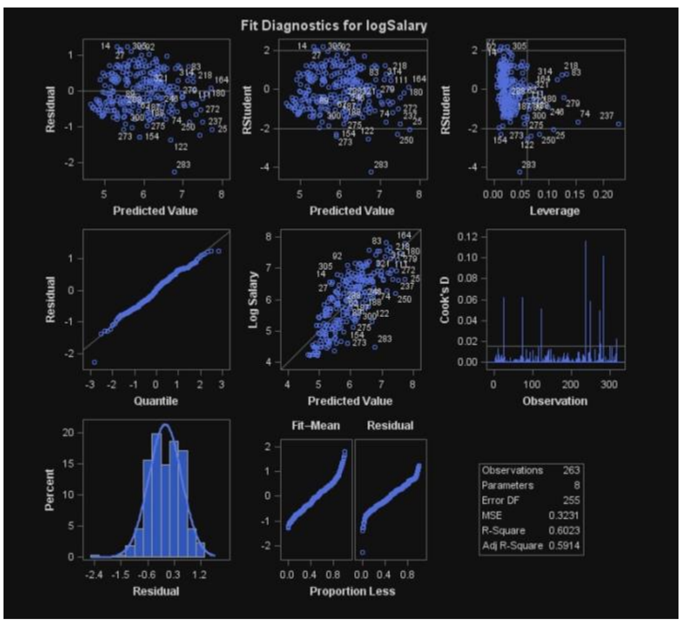


The distribution appears to be symmetric.



The QQ-Plot shows the scatters falling along a relatively straight line, so there is no serious violation against normality.

Regressing logSalary



Unlike the Fit Diagnostics for Salary, the diagnostics for logSalary appear to be constant variance after looking at the Homoscedastic Plot. Not many observations fall outside the designated cutoffs at 2 and -2. Looking at the QQ-plot there appears the scatters falling along a relatively straight line, so there is no serious violation against normality.

Therefore, logSalary would be a better option for regression analysis than salary

Extra Credit

5) Center Regressors

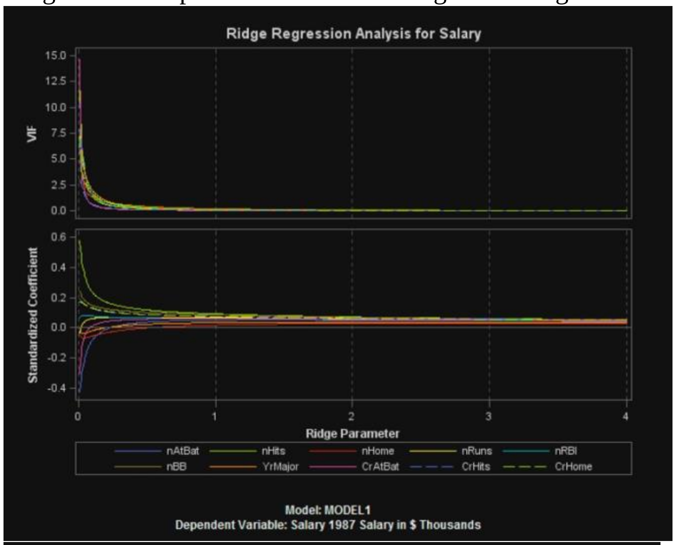
```
/*CENTERING*/
PROC STDIZE DATA=Num OUT=Num02 METHOD=MEAN;
  VAR Salary nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat
RUN;

PROC REG DATA=Num02 PLOTS=NONE;
  MODEL Salary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtB
RUN;
QUIT;
```

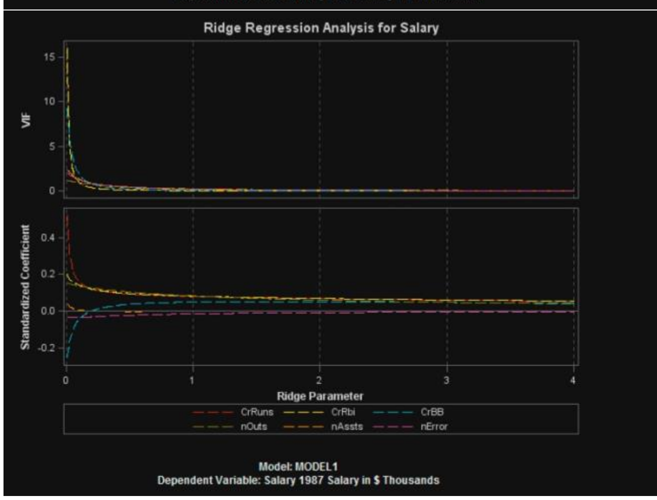
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	10	124.25217	12.42522	37.77
Error	252	82.90156	0.32897	
Corrected Total	262	207.15373		

Root MSE	0.57356
Dependent Mean	5.92722
R-Square	0.5998
Adj R-Sq	0.5839
AIC	-16.63353
AICC	-15.38553
SBC	-242.33984
CV PRESS	90.27767

Centering the regressors slightly improved the Adj R-Squared value from 0.5819 to 0.5998, but this still indicates almost 40% of the variability in the data cannot be accounted for in the model

6) Ridge Trace for possible deletion of insignificant regressors





The best ridge parameter appears to be nError since its line is closest to 0.

Appendix

```
/*Normality check*/
%MACRO NORMTEST(VAR,DATA);
/*********************************************************************************/
/* Macro NORMTEST is revised from the code in D'Agostino's paper.              */
/* "A Suggestion for Using Powerful and Informative Tests of Normality"         */
/* Author(s): Ralph B. D'Agostino, Albert Belanger, and Ralph B. D'Agostino Jr.  */
/* Source: The American Statistician, Vol. 44, No. 4 (Nov., 1990), pp. 316-321   */

/* It provides five hypothesis tests                                */
/* (1) Shapiro-Wilk test                                         */
/* (2) Kolmogorov-Smirnov test                                      */
/* (3) Cramer-von Mises test                                      */
/* (4) Anderson-Darling                                         */
/* (5,6,7) D'Agostino's K^2                                          */
/* For details about the first four tests, users are referred to SAS online doc  */
/* under UNIVARIATE procedure. As for D'Agostino's test, please refer to the art.*/
/* mentioned above.                                              */
/* Revised by Ping-Shi Wu Dec. 2015 @ Lehigh University                  */
/*********************************************************************************/
 ODS NOPROCTITLE;
 ODS GRAPHICS /BORDER=OFF;
 ODS SELECT Moments Histogram QQPlot CDFPlot;
 TITLE "NORMAL-TEST";
 PROC UNIVARIATE DATA=&DATA NORMAL;
  VAR &VAR;
       HISTOGRAM &VAR/NORMAL(MU=EST SIGMA=EST) KERNEL;
  QQPLOT &VAR/NORMAL(MU=EST SIGMA=EST);
  CDFPLOT &VAR/NORMAL(MU=EST SIGMA=EST);
  OUTPUT OUT=XXSTAT N=N MEAN=XBAR STD=S SKEWNESS=G1 KURTOSIS=G2;
 RUN;
 ODS SELECT TestsForNormality;
 PROC UNIVARIATE DATA=&DATA NORMAL;
  VAR &VAR;
 RUN;
 TITLE;
 OPTIONS LS=80;
 DATA _NULL_;
  SET XXSTAT;
  SQRTB1=(N-2)/SQRT(N*(N-1))*G1;
  Y=SQRTB1*SQRT((N+1)*(N+3)/(6*(N-2)));
  BETA2=3*(N*N+27*N-70)*(N+1)*(N+3)/((N-2)*(N+5)*(N+7)*(N+9));
  W=SQRT(-1+SQRT(2*(BETA2-1)));
  DELTA=1/SQRT(LOG(W));
  ALPHA=SQRT(2/(W*W-1));
       Z_B1=DELTA*LOG(Y/ALPHA+SQRT((Y/ALPHA)**2+1));
  B2=3*(N-1)/(N+1)+(N-2)*(N-3)/((N+1)*(N-1))*G2;
  MEANB2=3*(N-1)/(N+1);
  VARB2= 24*N*(N-2)*(N-3)/((N+1)*(N+1)*(N+3)*(N+5));
  X=(B2-MEANB2)/SQRT(VARB2);
```

```
   MOMENT=6*(N*N-5*N+2)/((N+7)*(N+9))*SQRT(6*(N+3)*(N+5)/(N*(N-2)*(N-3)));
   A=6+8/MOMENT*(2/MOMENT+SQRT(1+4/(MOMENT**2)));
   Z_B2=(1-2/(9*A)-((1-2/A)/(1+ X*SQRT(2/(A-4))))**(1/3))/SQRT(2/(9*A));
   PRZB1=2*(1-PROBNORM(ABS(Z_B1)));
   PRZB2=2*(1-PROBNORM(ABS(Z_B2)));
   CHITEST=Z_B1*Z_B1 + Z_B2*Z_B2;
   PRCHI=1-PROBCHI(CHITEST,2);
   FILE PRINT;
   PUT @22 "D'AGOSTINO TEST OF NORMALITY FOR VARIABLE &VAR, "
   N = /@20 G1=8.5 @33 SQRTB1 =8.5 @50 "Z=" Z_B1 8.5 @65 "P=" PRZB1 6.4
     /@20 G2=8.5 @33 B2=8.5 @50 "Z=" Z_B2 8.5 @65 "P=" PRZB2 6.4
     /@20 "K**2=CHISQ(2 DF)=" CHITEST 8.5 @65 "P=" PRCHI 6.4;
  RUN;
  TITLE;
%MEND NORMTEST;

ODS RTF FILE='PLAY.RTF';
DATA MLB86;
  SET SASHELP.BASEBALL;
RUN;

PROC CONTENTS DATA=MLB86 VARNUM;
RUN;

PROC SURVEYSELECT DATA=MLB86
         SAMPRATE=0.80
         SEED=818559125
         OUT=SAMPLE OUTALL
      METHOD=SRS NOPRINT;
RUN;


/*Split variables into three smaller groups of (6,5,5)*/
DATA S1(KEEP= Salary nAtBat nHits nHome nRuns nRBI nBB);
  SET MLB86;
RUN;
DATA S2(KEEP= Salary YrMajor CrAtBat CrHits CrHome CrRuns);
  SET MLB86;
RUN;
DATA S3(KEEP= Salary CrRbi CrBB nOuts nAssts nError);
  SET MLB86;
RUN;

DATA TEST(DROP=SELECTED);
  SET SAMPLE;
  WHERE SELECTED^=1;
RUN;

/*DATA TRAIN_1;*/
/* INPUT y x1 x2 x3 x4 x5;*/
/*  LABEL y = 'Y- var'*/
/*      x1= 'Name'*/
/*      x2= 'Team'*/
```

```
/*      x3= 'natbat'*/
/*                  x4= 'nHits';*/
/*                  x5= 'nHome';*/



/*proc sgscatter data=S1;*/
/*  compare y=(Salary)*/
/*        x=(nAtBat nHits nHome nRuns nRBI nBB)*/
/*        / reg ellipse=(type=mean) spacing=4;*/
/*run;*/

PROC SGSCATTER DATA = S1;
  MATRIX Salary nAtBat nHits nHome nRuns nRBI nBB / ellipse diagonal = (histogram normal);
RUN;
PROC SGSCATTER DATA = S2;
  MATRIX Salary YrMajor CrAtBat CrHits CrHome CrRuns / ellipse diagonal = (histogram normal);
RUN;
PROC SGSCATTER DATA = S3;
  MATRIX Salary CrRbi CrBB nOuts nAssts nError / ellipse diagonal = (histogram normal);
RUN;

/*Scatter matrix plus Correlation analysis*/
PROC CORR DATA= S1 SPEARMAN FISHER(BIASADJ=NO) PLOTS=MATRIX(HISTOGRAM
NVAR=6);
RUN;

PROC CORR DATA= S2 SPEARMAN FISHER(BIASADJ=NO) PLOTS=MATRIX(HISTOGRAM
NVAR=5);
RUN;
PROC CORR DATA= S3 SPEARMAN FISHER(BIASADJ=NO) PLOTS=MATRIX(HISTOGRAM
NVAR=5);
RUN;

/*1.b Generating Boxplots*/
/*Boxplots*/
PROC SGPLOT DATA=S1;
  VBOX Salary;
RUN;
PROC SGPLOT DATA=S1;
  VBOX nAtBat;
RUN;
PROC SGPLOT DATA=S1;
  VBOX nHits;
RUN;
PROC SGPLOT DATA=S1;
  VBOX nHome;
RUN;
PROC SGPLOT DATA=S1;
  VBOX nRuns;
RUN;
PROC SGPLOT DATA=S1;
```

```
  VBOX nRBI;
RUN;
PROC SGPLOT DATA=S2;
  VBOX YrMajor;
RUN;
PROC SGPLOT DATA=S2;
  VBOX CrAtBat;
RUN;
PROC SGPLOT DATA=S2;
  VBOX CrHits;
RUN;
PROC SGPLOT DATA=S2;
  VBOX CrHome;
RUN;
PROC SGPLOT DATA=S2;
  VBOX CrRuns;
RUN;
/*TODO- still need s3 data, display as 3 collumns*/

/*getting categorical data from MLB86*/
/*1.c Frequency Table for Categorical Data*/
title "Computing Frequencies and Percentages Using PROC FREQ";
proc freq data=MLB86;
tables Team Position League Division;
run;



/*2*/
/*Correlation Analysis on numerical variables*/
DATA Num(DROP= Name Team League Division Position logSalary);
  SET MLB86;
RUN;

DATA Num_logSal(DROP= Salary Name Team League Division Position);
  SET MLB86;
RUN;

PROC CORR DATA=S1;
RUN;
QUIT;
PROC CORR DATA=S2;
RUN;
QUIT;
PROC CORR DATA=S3;
RUN;
QUIT;


/*3.a and 3.c*/
/*Full Model of Salary on all numerical except logSalary*/
/*Show evidence of multicollinearity among regressors*/
PROC REG DATA=Num;
```

```
  MODEL Salary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/DWPROB COLLIN VIF;
  OUTPUT OUT=CFM_FIT RSTUDENT=D;
RUN;
QUIT;

%NORMTEST(D,CFM_FIT)

/*CENTERING*/
PROC STDIZE DATA=Num OUT=Num02 METHOD=MEAN;
  VAR Salary nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns CrRbi
CrBB nOuts nAssts nError;
RUN;

PROC REG DATA=Num02 PLOTS=NONE;
  MODEL Salary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/COLLIN VIF;
RUN;
QUIT;

/*TRY RIDGE TRACE NEXT TO SEE IF DELETION OF INSIGNIFICANT VARIABLE CAN HELP */
PROC REG DATA=Num OUTEST=EST_RIDGE RIDGE=0.01 TO 4 BY 0.005 OUTVIF;
  MODEL Salary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError;
RUN;
QUIT;

/*/*TRY DELETING X3*/*/
/*PROC REG DATA=CEMENT PLOTS=NONE;*/
/*  MODEL Y= X1 X2 X4/VIF COLLIN;*/
/*RUN;*/
/*QUIT;*/

/*DELETION X3 ALLEVIATES THE COLLINEARITY A LOT, BUT NOT COMPLETELY REMOVE
*/
/*(1) PASS SINCE IT HELPS TO ALLEVIATE THE COLLINEARITY*/
/*(2) NO PASS;*/ /* PROCEED WITH */
/*(2-A1) RIDGE REGRESSION IF NO SELECTION OF FEATURES IS INTENDED*/
/*(2-A1) PC REGRESSION IF NO SELECTION OF FEATURES IS INTENDED*/
/*(2-B) (GROUP) LASSO IF SELECTION OF FEATURES IS INTENDED*/

/* 1. ALL REGRESSION MODELS */
PROC REG DATA=Num PLOTS(ONLY)=CRITERIONPANEL(UNPACK LABELVARS);
PROC REG DATA=Num PLOTS(ONLY)=CRITERIONPANEL;
  MODEL Salary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=ADJRSQ AIC BIC SBC CP;
  PLOT CP.*NP./VAXIS=0 TO 250 BY 50 HAXIS=0 TO 4 BY 1 CHOCKING=RED NOMODEL
NOSTAT;
RUN;
QUIT;

PROC REG DATA=Num PLOTS(ONLY)=CRITERIONPANEL;
```

```
  MODEL Salary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=F DETAILS;
RUN;
QUIT;

PROC REG DATA=Num PLOTS(ONLY)=CRITERIONPANEL;
  MODEL Salary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=B DETAILS;
RUN;
QUIT;

PROC REG DATA=Num PLOTS(ONLY)=CRITERIONPANEL;
  MODEL Y=X1 X2 X4/SELECTION=STEPWISE DETAILS;
RUN;
QUIT;



/*3.b and 3.c*/
/*Full Model of logSalary on all numerical except salary*/
/*Show evidence of multicollinearity among regressors*/

PROC REG DATA=Num_logSal;
  MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/DWPROB COLLIN VIF;
  OUTPUT OUT=FC_FIT RSTUDENT=D;
RUN;
QUIT;

%NORMTEST(D,FC_FIT)


PROC REG DATA=Num_logSal;
  MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/DWPROB COLLIN VIF;
  OUTPUT OUT=CFM_FIT RSTUDENT=D;
RUN;
QUIT;

%NORMTEST(D,CFM_FIT)

/*CENTERING*/
PROC STDIZE DATA=Num_logSal OUT=Num03 METHOD=MEAN;
  VAR logSalary nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns CrRbi
CrBB nOuts nAssts nError;
RUN;

PROC REG DATA=Num03 PLOTS=NONE;
  MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/COLLIN VIF;
RUN;
QUIT;
```

```
/*TRY RIDGE TRACE NEXT TO SEE IF DELETION OF INSIGNIFICANT VARIABLE CAN HELP */
PROC REG DATA=Num_logSal OUTEST=EST_RIDGE RIDGE=0.01 TO 4 BY 0.005 OUTVIF;
  MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError;
RUN;
QUIT;

/*/*TRY DELETING X3*/*/
/*PROC REG DATA=CEMENT PLOTS=NONE;*/
/*  MODEL Y= X1 X2 X4/VIF COLLIN;*/
/*RUN;*/
/*QUIT;*/

/*DELETION X3 ALLEVIATES THE COLLINEARITY A LOT, BUT NOT COMPLETELY REMOVE
*/
/*(1) PASS SINCE IT HELPS TO ALLEVIATE THE COLLINEARITY*/
/*(2) NO PASS;*/ /* PROCEED WITH */
/*(2-A1) RIDGE REGRESSION IF NO SELECTION OF FEATURES IS INTENDED*/
/*(2-A1) PC REGRESSION IF NO SELECTION OF FEATURES IS INTENDED*/
/*(2-B) (GROUP) LASSO IF SELECTION OF FEATURES IS INTENDED*/

/* 1. ALL REGRESSION MODELS */
PROC REG DATA=Num PLOTS(ONLY)=CRITERIONPANEL(UNPACK LABELVARS);
PROC REG DATA=Num_logSal PLOTS(ONLY)=CRITERIONPANEL;
  MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=ADJRSQ AIC BIC SBC CP;
  PLOT CP.*NP./VAXIS=0 TO 250 BY 50 HAXIS=0 TO 4 BY 1 CHOCKING=RED NOMODEL
NOSTAT;
RUN;
QUIT;

PROC REG DATA=Num PLOTS(ONLY)=CRITERIONPANEL;
  MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=F DETAILS;
RUN;
QUIT;

PROC REG DATA=Num_logSal PLOTS(ONLY)=CRITERIONPANEL;
  MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=B DETAILS;
RUN;
QUIT;

PROC REG DATA=Num_logSal PLOTS(ONLY)=CRITERIONPANEL;
  MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=STEPWISE DETAILS;
RUN;
QUIT;

/*X1 X2 IS THE MAJOR WINNER*/
PROC REG DATA=Num_logSal;
  MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/DWPROB COLLIN VIF;
```

```
  OUTPUT OUT=FC_FIT RSTUDENT=D;
RUN;
QUIT;

%NORMTEST(D,FC_FIT)

/*3.c and show evidence of multicollinearity among regressors*/
/*TODO*/

/*3.d model selection*/
/*Model Selection for Salary on numerical data*/
PROC REG DATA=Num PLOTS(LABEL)=CRITERIA;
  MODEL Salary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=ADJRSQ CP AIC BIC SBC;
RUN;
QUIT;

PROC REG DATA=Num PLOTS(LABEL)=CRITERIA;
  MODEL Salary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=FORWARD;
RUN;
QUIT;

PROC REG DATA=Num PLOTS(LABEL)=CRITERIA;
  MODEL Salary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=BACKWARD;
RUN;
QUIT;

PROC REG DATA=Num PLOTS(LABEL)=CRITERIA;
  MODEL Salary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=STEPWISE;
RUN;
QUIT;

PROC GLMSELECT DATA=Num PLOTS=ALL;
   MODEL Salary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=LASSO(CHOOSE=CV STOP=NONE)
CVMETHOD=RANDOM(10);
RUN;

PROC GLMSELECT DATA=Num PLOTS=ALL;
   MODEL Salary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=ELASTICNET(CHOOSE=CV STOP=NONE)
CVMETHOD=RANDOM(10);
RUN;


/*Model Selection for logSalary on numerical data*/
/*3.d model selection*/
/*Model Selection for Salary on numerical data*/
PROC REG DATA=Num_logSal PLOTS(LABEL)=CRITERIA;
```

```
  MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=ADJRSQ CP AIC BIC SBC;
RUN;
QUIT;

PROC REG DATA=Num_logSal PLOTS(LABEL)=CRITERIA;
  MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=FORWARD;
RUN;
QUIT;

PROC REG DATA=Num_logSal PLOTS(LABEL)=CRITERIA;
  MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=BACKWARD;
RUN;
QUIT;

PROC REG DATA=Num_logSal PLOTS(LABEL)=CRITERIA;
  MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=STEPWISE;
RUN;
QUIT;

PROC GLMSELECT DATA=Num_logSal PLOTS=ALL;
   MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=LASSO(CHOOSE=CV STOP=NONE)
CVMETHOD=RANDOM(10);
RUN;

PROC GLMSELECT DATA=Num_logSal PLOTS=ALL;
   MODEL logSalary=nAtBat nHits nHome nRuns nRBI nBB YrMajor CrAtBat CrHits CrHome CrRuns
CrRbi CrBB nOuts nAssts nError/SELECTION=ELASTICNET(CHOOSE=CV STOP=NONE)
CVMETHOD=RANDOM(10);
RUN;
```