Baseball Salary Activity

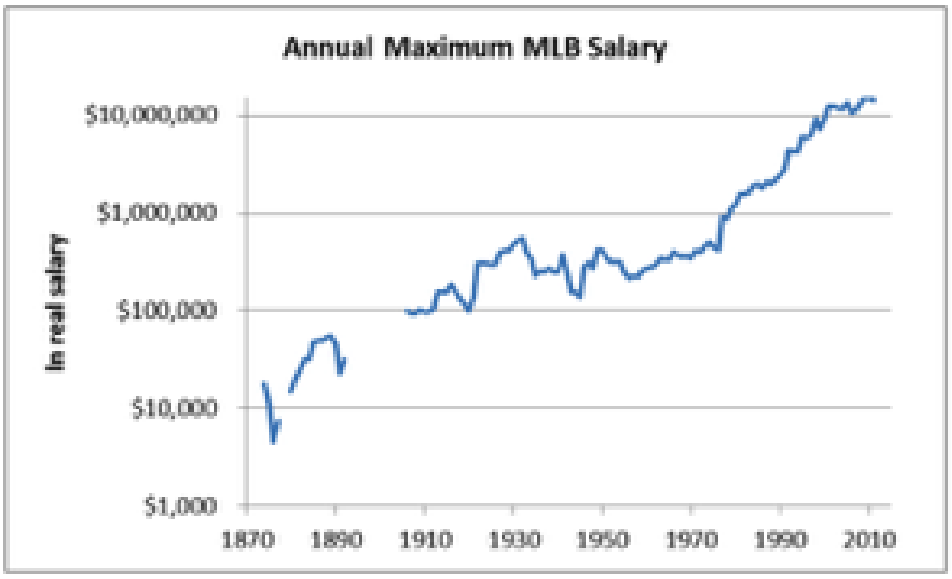
Jack Tubbs

November 2021

Contents

1 Introduction

The Sashelp.Baseball data set contains salary and performance information for Major League Baseball players (excluding pitchers) who played at least one game in both the 1986 and 1987 seasons (Time Inc. 1987). The salaries are for the 1987 season, and the performance measures are from the 1986 season. The data set contains 322 observations.
In 1986 the minimum salary was $60.5k and the average salary was $412.5k. The maximum salary was $2,412.5k that was paid to Jim Rice of the Boston Red Sox in the American league. In 1976, Hank Aaron was highest paid player at $240k.
The following graph reflects the significant increase in salaries beginning in the 1980's.



2 Plan for the Analysis

The baseball data will be used to illustrate issues related to linear regression models. The analysis that I will use is highly restrictive in terms of position players and their tenure in the major league. The independent variables are career baseball related statistics, as opposed to the same statistics in the previous year (1986). The dependent variable of interest is salary in 1987. Salary are not normally distributed (the density is shaped like an exponential curve). Log salary was suggested as normally distributed replacement. I confirmed this choice with a boxcox transformation.

In this document I illustrate the SAS code and output for linear, quadratic and multiple regression models with the respective diagnostic information. The last model illustrates what is called "model selection procedures". Here, I use stepwise selection with two different SAS procedures before considering a newer method call LASSO selection.

The analysis is just for illustrative purposes only. The use for career independent variables is problematic since these increase for each year that you play. Hence, years in the majors is a nuisance variable for both the independent and dependent variable, log Salary.

3 SAS

3.1 Code

```
options center nodate pagesize=80 ls=70;
title "1986 Baseball Data";
data baseball; set sashelp.baseball;
run;
```

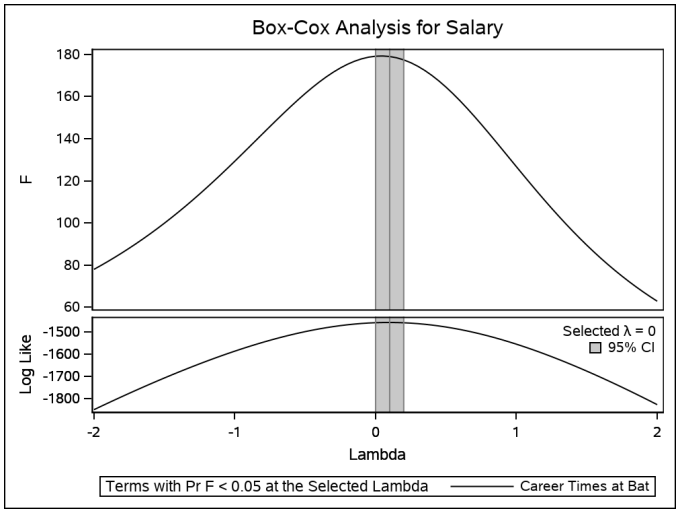
```
data baseball; set baseball;
in_fielder = (position in ('1B' '2B' 'SS' '3B'));
out_fielder = (position in ('CF' 'RF' 'LF' 'OF'));
catcher = (position = 'C');
CrHits2 = CrHits*CrHits;
run;

/*
proc contents data=baseball short;
run;

proc freq data=baseball;
table (in_fielder out_fielder catcher)*YrMajor;
run;
*/

proc transreg data=baseball
              plots=(transformation(dependent) obp);
   model BoxCox(Salary / convenient lambda=-2 to 2 by 0.05) =
         identity(CrAtBat);
run;

proc sgplot data=baseball;
histogram logSalary;
density logSalary;
density logSalary/ type= kernel;
run;

proc sgplot data=baseball;
vbox logSalary/group=out_fielder;
run;

proc sgscatter data=baseball; where out_fielder = 1;
matrix logSalary CrAtBat CrBB CrHits CrHome CrRbi CrRuns;
run;

title2 'Linear Regression';
proc reg data=baseball; where out_fielder = 1 and 3 < YrMajor < 10;
model logSalary = CrHits;
run;

title2 'Quadratic Regression';
proc reg data=baseball; where out_fielder = 1 and 3 < YrMajor < 10;
model logSalary = CrHits CrHits2;
run;

title2 'Multiple Regression';
title3 'Position -- Catcher';
proc reg data=baseball plots=none; where Catcher = 1 and 3 < YrMajor < 10;
```

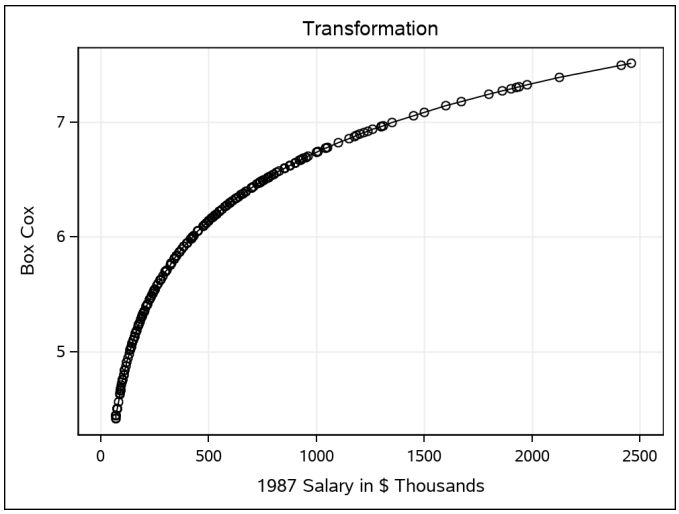
```
model logSalary = CrAtBat CrBB CrHits CrHome CrRbi CrRuns/ss2 ;
run;

title2 'Multiple Regression with Check for Collinearity';
proc reg data=baseball;* plots=none; where Catcher = 1 and 3 < YrMajor < 10;
model logSalary =   CrHits CrHome CrRuns/ss2 VIF collinoint;
run;

title2 'Model Selection Regression';
proc reg data=baseball plots=none; where Catcher = 1 and 3 < YrMajor < 10;
model logSalary = CrAtBat CrBB CrHits CrHome CrRbi CrRuns
               /ss2 best=5 selection=stepwise aic bic details=summary;
run;

proc glmselect data=baseball plot=CriterionPanel;
                          where Catcher = 1 and 3 < YrMajor < 10;
   model logSalary =
                 yrMajor crAtBat crHits crHome crRuns crRbi
                 crBB
               / selection=stepwise(select=SL) stats=all;
run;

proc glmselect data=baseball plot=CriterionPanel;
                          where Catcher = 1 and 3 < YrMajor < 10;
   model logSalary =
                 yrMajor crAtBat crHits crHome crRuns crRbi
                 crBB
/ selection=LASSO(choose=CP steps=4);
run;
 quit;
ods latex close;
```

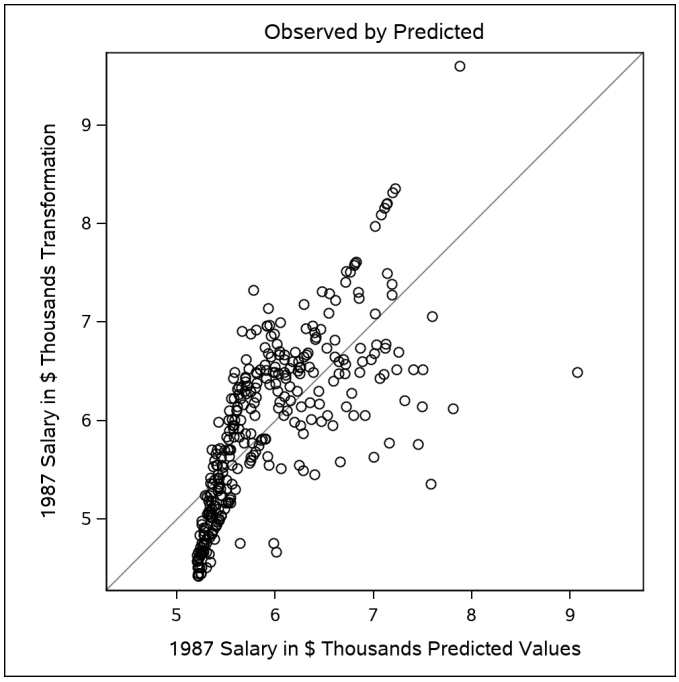
3.2 The TRANSREG Procedure

Code

```
proc transreg data=baseball
             plots=(transformation(dependent) obp);
   model BoxCox(Salary / convenient lambda=-2 to 2 by 0.05) =
         identity(CrAtBat);
run;
```


Box-Cox Analysis for Salary
F
180
160
140
120
100
80
60
Log Like
-1500
-1600
-1700
-1800
Selected λ = 0
95% CI
-2
-1
0
1
2
Lambda
Terms with Pr F < 0.05 at the Selected Lambda
Career Times at Bat

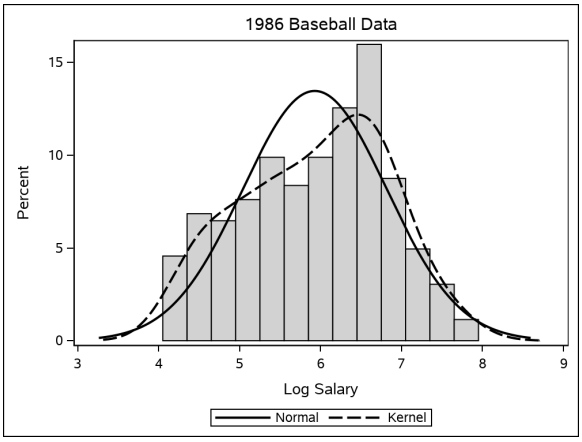

Transformation
Box Cox
7
6
5
0
500
1000
1500
2000
2500
1987 Salary in $ Thousands

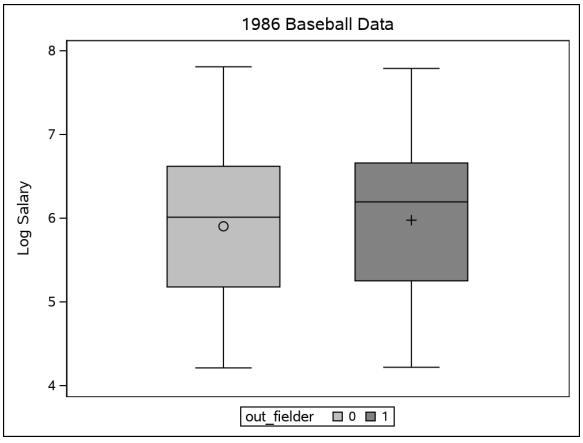


3.3 SGplots

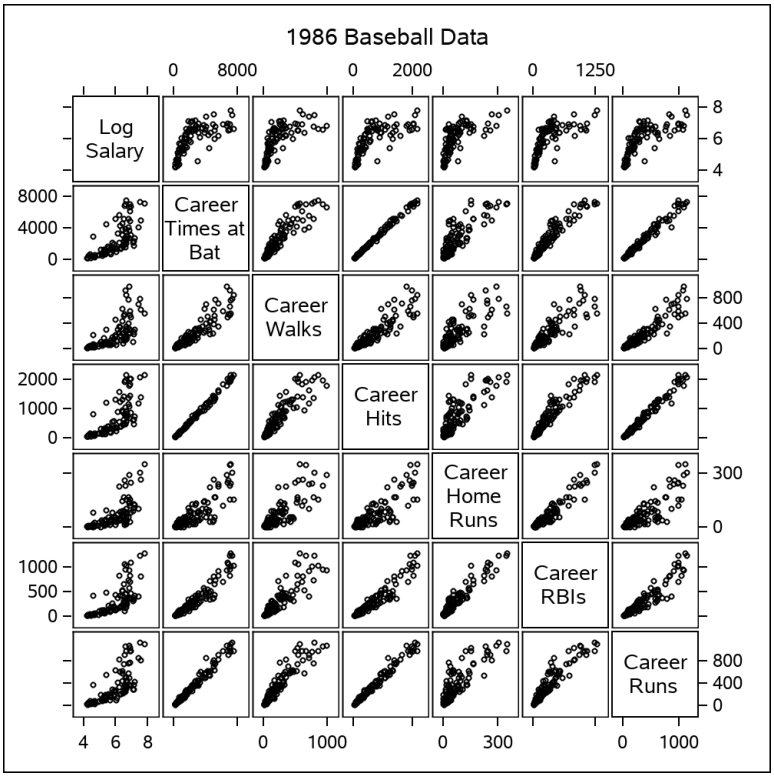
Code

```
proc sgplot data=baseball;
histogram logSalary;
density logSalary;
density logSalary/ type= kernel;
run;

proc sgplot data=baseball;
vbox logSalary/group=out_fielder;
run;

proc sgscatter data=baseball; where out_fielder = 1;
matrix logSalary CrAtBat CrBB CrHits CrHome CrRbi CrRuns;
run;
```


1986 Baseball Data
Percent
0
5
10
15
3
4
5
6
7
8
9
Log Salary
Normal
Kernel


1986 Baseball Data
Log Salary
4
5
6
7
8
out_fielder
0
1

Scatter Plots



3.4 Linear Regression

Code

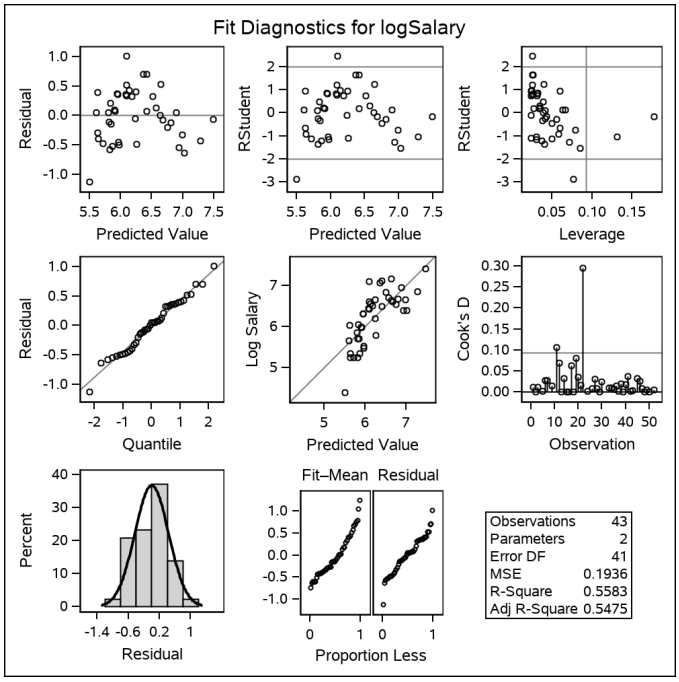
```
title2 'Linear Regression';
proc reg data=baseball; where out_fielder = 1 and 3 < YrMajor < 10;
model logSalary = CrHits;
run;
```

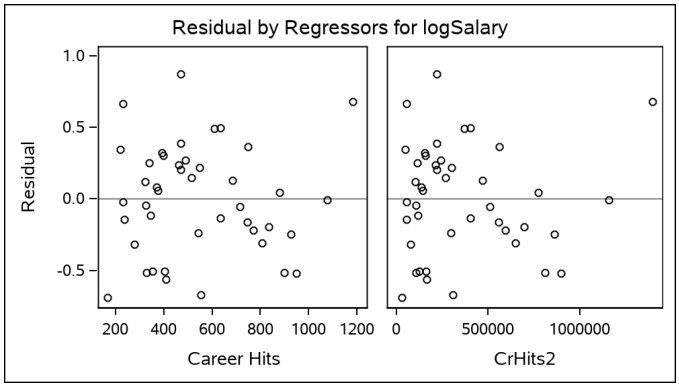
Number of Observations Read	52
Number of Observations Used	43
Number of Observations with Missing Values	9

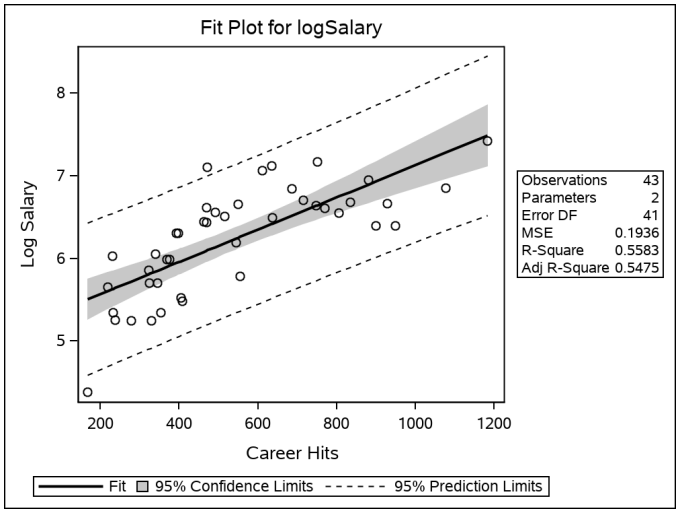
Analysis of Variance					
Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	1	10.03510	10.03510	51.82	<.0001
Error	41	7.93901	0.19363		
Corrected Total	42	17.97411			

Root MSE	0.44004	*R-Square*	0.5583
Dependent Mean	6.24016	*Adj R-Sq*	0.5475
Coeff Var	7.05173		

Parameter Estimates						
Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	Intercept	1	5.18043	0.16178	32.02	<.0001
CrHits	Career Hits	1	0.00195	0.00027102	7.20	<.0001


Fit Diagnostics for logSalary
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Log Salary
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 43
Parameters 2
Error DF 41
MSE 0.1936
R-Square 0.5583
Adj R-Square 0.5475


Residual by Regressors for logSalary
Residual
1.0
0.5
0.0
-0.5
200
400
600
800
1000
1200
Career Hits
0
500000
1000000
CrHits2


Fit Plot for logSalary
Log Salary
8
7
6
5
200
400
600
800
1000
1200
Career Hits
Observations 43
Parameters 2
Error DF 41
MSE 0.1936
R-Square 0.5583
Adj R-Square 0.5475
Fit
95% Confidence Limits
95% Prediction Limits

3.5 Quadratic Regression

Code

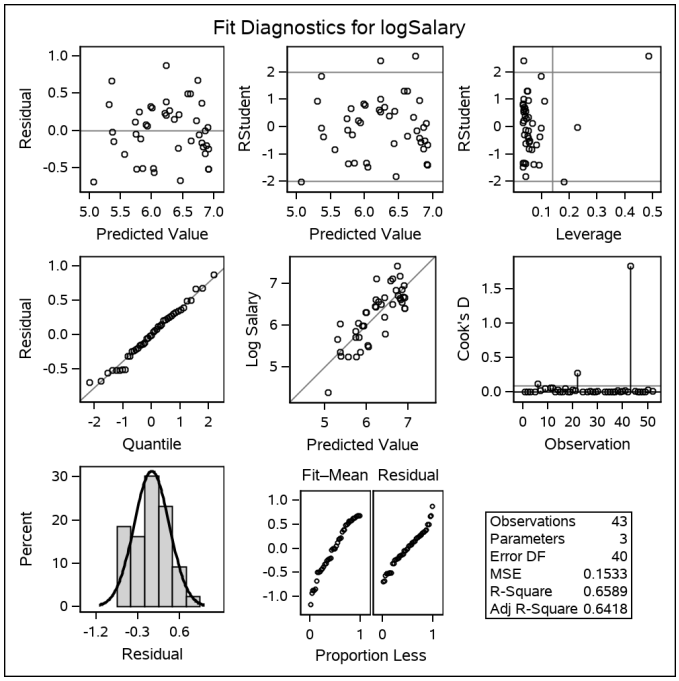
```
title2 'Quadratic Regression';
proc reg data=baseball; where out_fielder = 1 and 3 < YrMajor < 10;
model logSalary = CrHits CrHits2;
run;
```

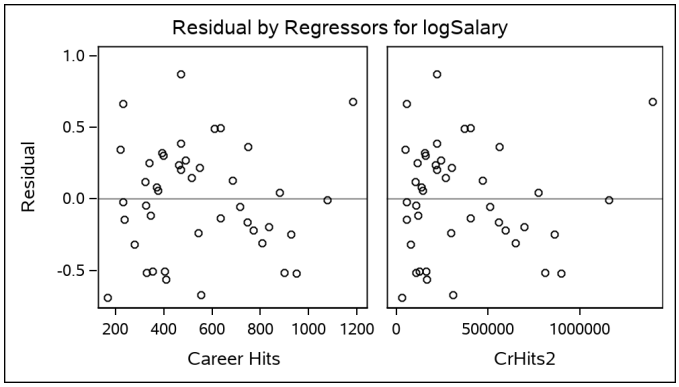
Number of Observations Read	52
Number of Observations Used	43
Number of Observations with Missing Values	9

Analysis of Variance					
Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	11.84233	5.92116	38.63	<.0001
Error	40	6.13178	0.15329		
Corrected Total	42	17.97411			

Root MSE	0.39153	*R-Square*	0.6589
Dependent Mean	6.24016	*Adj R-Sq*	0.6418
Coeff Var	6.27434		

Parameter Estimates						
Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	Intercept	1	4.19287	0.32163	13.04	<.0001
CrHits	Career Hits	1	0.00578	0.00114	5.07	<.0001
CrHits2		1	−0.00000307	8.930039E−7	−3.43	0.0014


Fit Diagnostics for logSalary
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Log Salary
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 43
Parameters 3
Error DF 40
MSE 0.1533
R-Square 0.6589
Adj R-Square 0.6418


Residual by Regressors for logSalary
Residual
Career Hits
CrHits2

3.6 Multiple Regression

Code

```
title2 'Multiple Regression';
title3 'Position -- Catcher';
proc reg data=baseball plots=none; where Catcher = 1 and 3 < YrMajor < 10;
model logSalary = CrAtBat CrBB CrHits CrHome CrRbi CrRuns/ss2 ;
run;
```

Number of Observations Read	18
Number of Observations Used	15
Number of Observations with Missing Values	3

Analysis of Variance					
Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	6	6.35116	1.05853	11.84	0.0013
Error	8	0.71531	0.08941		
Corrected Total	14	7.06647			

Root MSE	0.29902	*R-Square*	0.8988
Dependent Mean	6.14352	*Adj R-Sq*	0.8229
Coeff Var	4.86726		

Parameter Estimates							
Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Type II SS*
Intercept	Intercept	1	4.43875	0.25164	17.64	<.0001	27.81956
CrAtBat	Career Times at Bat	1	0.00120	0.00113	1.07	0.3172	0.10174
CrBB	Career Walks	1	0.00340	0.00148	2.31	0.0500	0.47526
CrHits	Career Hits	1	0.00242	0.00297	0.81	0.4388	0.05936
CrHome	Career Home Runs	1	0.01952	0.00914	2.14	0.0651	0.40824
CrRbi	Career RBIs	1	−0.01128	0.00478	−2.36	0.0461	0.49701
CrRuns	Career Runs	1	−0.00269	0.00600	−0.45	0.6652	0.01804

3.7 Multiple Regression with Check for Collinearity

Code

```
title2 'Multiple Regression with Check for Collinearity';
proc reg data=baseball;* plots=none; where Catcher = 1 and 3 < YrMajor < 10;
model logSalary =   CrHits CrHome CrRuns/ss2 VIF collinoint;
run;
```

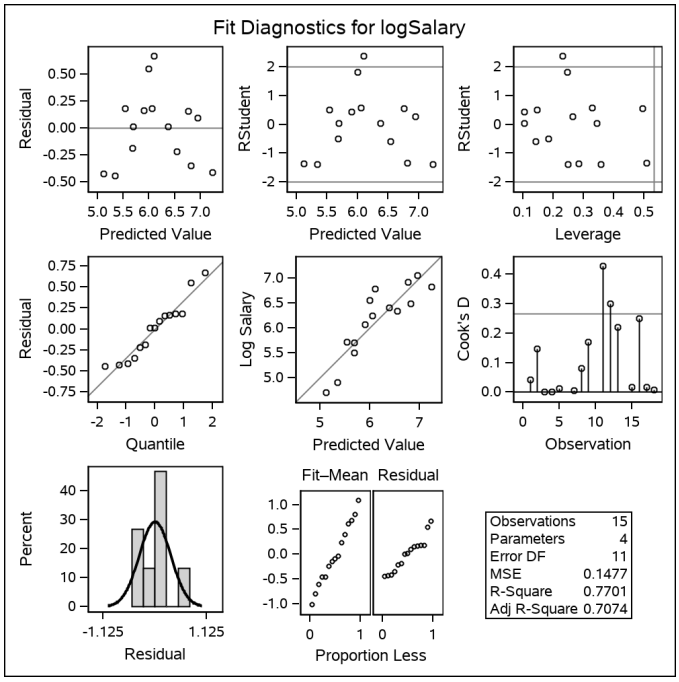
Number of Observations Read	18
Number of Observations Used	15
Number of Observations with Missing Values	3

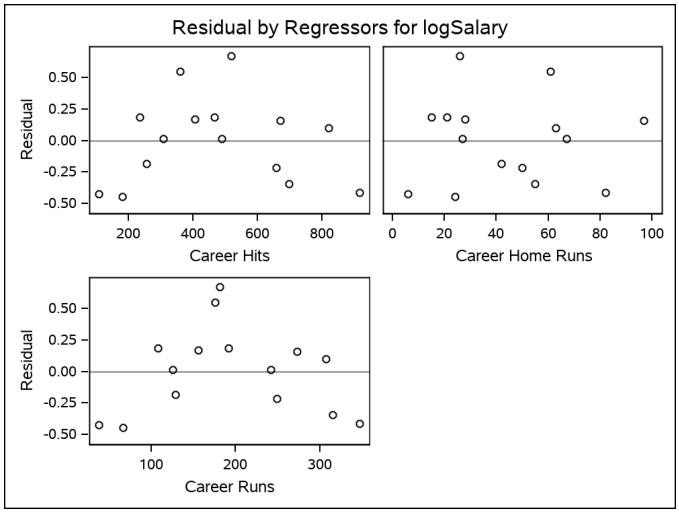
Analysis of Variance					
Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	3	5.44188	1.81396	12.28	0.0008
Error	11	1.62458	0.14769		
Corrected Total	14	7.06647			

Root MSE	0.38430	*R-Square*	0.7701
Dependent Mean	6.14352	*Adj R-Sq*	0.7074
Coeff Var	6.25544		

Parameter Estimates								
Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Type II SS*	*Variance Inflation*
Intercept	Intercept	1	4.86568	0.23758	20.48	<.0001	61.94533	0
CrHits	Career Hits	1	0.00109	0.00194	0.56	0.5858	0.04654	20.64082
CrHome	Career Home Runs	1	0.00256	0.00678	0.38	0.7127	0.02110	3.01114
CrRuns	Career Runs	1	0.00335	0.00565	0.59	0.5648	0.05203	26.17974

Collinearity Diagnostics (intercept adjusted)					
Number	*Eigenvalue*	*Condition Index*	*Proportion of Variation*		
			CrHits	*CrHome*	*CrRuns*
1	2.67140	1.00000	0.00625	0.03665	0.00515
2	0.30696	2.95005	0.03587	0.74775	0.01015
3	0.02165	11.10934	0.95787	0.21560	0.98470


Fit Diagnostics for logSalary
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
Log Salary
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 15
Parameters 4
Error DF 11
MSE 0.1477
R-Square 0.7701
Adj R-Square 0.7074


Residual by Regressors for logSalary
Residual
Career Hits
Career Home Runs
Residual
Career Runs

3.8 Model Selection Regression - PROC REG

Code

```
title2 'Model Selection Regression';
proc reg data=baseball plots=none; where Catcher = 1 and 3 < YrMajor < 10;
model logSalary = CrAtBat CrBB CrHits CrHome CrRbi CrRuns
                /ss2 best=5 selection=stepwise aic bic details=summary;
run;
```

Number of Observations Read	18
Number of Observations Used	15
Number of Observations with Missing Values	3

Summary									
Step	*Entered*	*Removed*	*Label*	*# In*	*Partial R^2*	*Model R^2*	*C(p)*	*F Value*	*Pr > F*
1	CrAtBat		Career Times at Bat	1	0.7665	0.7665	7.4571	42.66	<.0001
2	CrBB		Career Walks	2	0.0586	0.8250	4.8293	4.02	0.0682

Number of Observations Read	18
Number of Observations Used	15
Number of Observations with Missing Values	3

Analysis of Variance					
Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	2	5.82994	2.91497	28.29	<.0001
Error	12	1.23653	0.10304		
Corrected Total	14	7.06647			

Root MSE	0.32100	*R-Square*	0.8250
Dependent Mean	6.14352	*Adj R-Sq*	0.7959
Coeff Var	5.22510		

Parameter Estimates							
Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Type II SS*
Intercept	Intercept	1	4.72604	0.20599	22.94	<.0001	54.24256
CrAtBat	Career Times at Bat	1	0.00049980	0.00015227	3.28	0.0066	1.11010
CrBB	Career Walks	1	0.00314	0.00157	2.00	0.0682	0.41378

3.9 Model Selection Regression - PROC GLMSELECT

3.9.1 Stepwise

Code

```
proc glmselect data=baseball plot=CriterionPanel;
                          where Catcher = 1 and 3 < YrMajor < 10;
   model logSalary =
                  yrMajor crAtBat crHits crHome crRuns crRbi
                  crBB
                / selection=stepwise(select=SL) stats=all;
run;
```

1986 Baseball Data

Model Selection Regression

The GLMSELECT Procedure

Data Set	WORK.BASEBALL
Dependent Variable	logSalary
Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.15
Stay Significance Level (SLS)	0.15
Effect Hierarchy Enforced	None

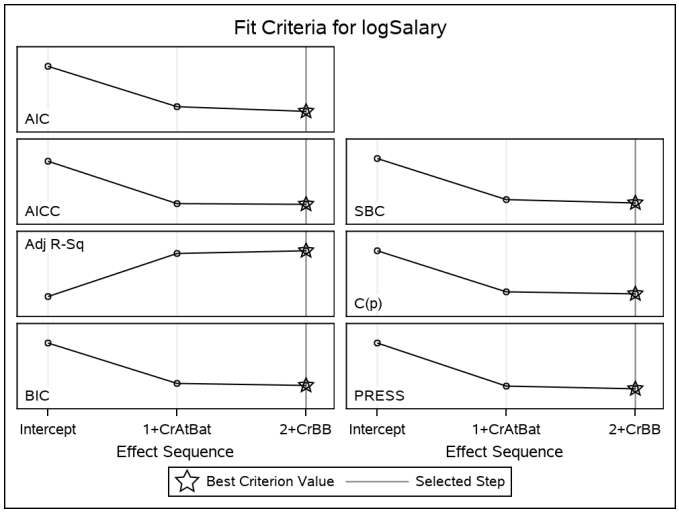
Number of Observations Read	18
Number of Observations Used	15

Dimensions	
Number of Effects	8
Number of Parameters	8

Step	*Enter*	*# In*	*Model* R^2	*Adj* R^2	*AIC*	*AICC*	*BIC*	*CP*	*SBC*	*PRESS*	*ASE*
0	β_0	1	0.0000	0.0000	7.7097	8.7097	−10.1654	61.6465	−8.5823	8.1120	0.4711
1	*CrAtBat*	2	0.7665	0.7485	−12.1063	−9.9245	−27.7035	6.4330	−27.6902	2.3475	0.1100
2	*CrBB*	3	0.8250	0.7959*	−14.4361*	−10.4361*	−28.5900*	4.0620*	−29.3120*	2.0192*	0.0824

Selection stopped because the candidate for entry has SLE > 0.15 and the candidate for removal has SLS < 0.15.

Stop Details					
Candidate For	*Effect*	*Candidate Significance*		*Compare Significance*	
Entry	YrMajor	0.1858	>	0.1500	(SLE)
Removal	CrBB	0.0682	<	0.1500	(SLS)



Selected Model

Note	The selected model is the model at the last step (Step 2).

Effects:	Intercept CrAtBat CrBB

Analysis of Variance				
Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*
Model	2	5.82994	2.91497	28.29
Error	12	1.23653	0.10304	
Corrected Total	14	7.06647		

Root MSE	0.32100
Dependent Mean	6.14352
R-Square	0.8250
Adj R-Sq	0.7959
AIC	−14.43614
AICC	−10.43614
BIC	−28.58997
C(p)	4.06204
PRESS	2.01923
SBC	−29.31199
ASE	0.08244

Parameter Estimates				
Parameter	*DF*	*Estimate*	*Standard Error*	*t Value*
Intercept	1	4.726037	0.205986	22.94
CrAtBat	1	0.000500	0.000152	3.28
CrBB	1	0.003140	0.001567	2.00

3.9.2 LASSO

Code

```
proc glmselect data=baseball plot=CriterionPanel;
                          where Catcher = 1 and 3 < YrMajor < 10;
   model logSalary =
                  yrMajor crAtBat crHits crHome crRuns crRbi
                  crBB
/ selection=LASSO(choose=CP steps=4);
run;
```

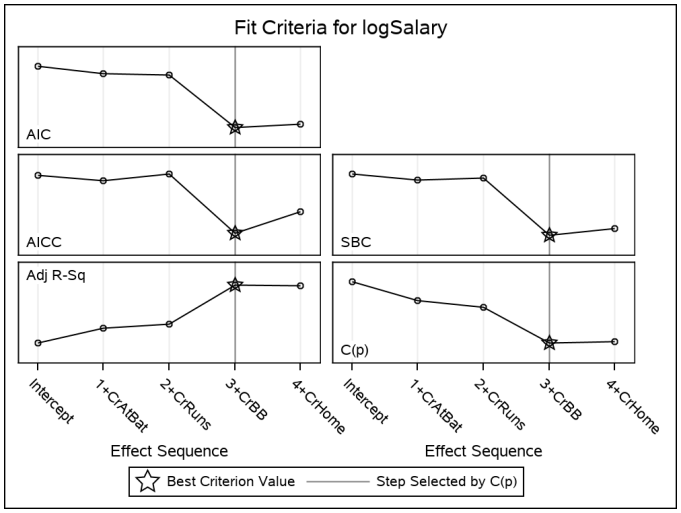
Data Set	WORK.BASEBALL
Dependent Variable	logSalary
Selection Method	LASSO
Stop at Specified Number of Steps	4
Choose Criterion	C(p)
Effect Hierarchy Enforced	None

Number of Observations Read	18
Number of Observations Used	15

Dimensions	
Number of Effects	8
Number of Parameters	8

LASSO Selection Summary				
Step	*Effect Entered*	*Effect Removed*	*Number Effects In*	*CP*
0	*Intercept*		1	61.6465
1	*CrAtBat*		2	44.9536
2	*CrRuns*		3	38.9718
3	*CrBB*		4	7.1642*
4	*CrHome*		5	8.3955
** Optimal Value of Criterion*				

Selection stopped at the specified number of steps (4).



Selected Model

Note	The selected model, based on C(p), is the model at Step 3.

Effects:	Intercept CrAtBat CrRuns CrBB

Analysis of Variance				
Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*
Model	3	5.72560	1.90853	15.66
Error	11	1.34086	0.12190	
Corrected Total	14	7.06647		

Root MSE	0.34914
Dependent Mean	6.14352
R-Square	0.8102
Adj R-Sq	0.7585
AIC	−11.22104
AICC	−4.55438
BIC	−25.75593
C(p)	7.16418
SBC	−25.38884

Parameter Estimates		
Parameter	*DF*	*Estimate*
Intercept	1	4.928902
CrAtBat	1	0.000278
CrRuns	1	0.001564
CrBB	1	0.002516