**4.1.1: SAS - Penalized Regression Methods**

**(Ridge Regression, LASSO, and Elastic Net)**

Example 1: (Ridge Regression; recall Handout 2.6.1 example) A study seeks to relate (in females) amount of body fat (Y) to triceps skinfold thickness (X1), thigh circumference (X2), and midarm circumference (X3). Amount of body fat is expensive to measure, requiring immersion of person in water. This expense motivates the desire for a predictive model based on these inexpensive predictors.

**/\* Input data \*/**

**data bodyfat;**

**input triceps thigh midarm body @@; cards;**

**19.5 43.1 29.1 11.9 24.7 49.8 28.2 22.8**

**30.7 51.9 37.0 18.7 29.8 54.3 31.1 20.1**

**19.1 42.2 30.9 12.9 25.6 53.9 23.7 21.7**

**31.4 58.5 27.6 27.1 27.9 52.1 30.6 25.4**

**22.1 49.9 23.2 21.3 25.5 53.5 24.8 19.3**

**31.1 56.6 30.0 25.4 30.4 56.7 28.3 27.2**

**18.7 46.5 23.0 11.7 19.7 44.2 28.6 17.8**

**14.6 42.7 21.3 12.8 29.5 54.4 30.1 23.9**

**27.7 55.3 25.7 22.6 30.2 58.6 24.6 25.4**

**22.7 48.2 27.1 14.8 25.2 51.0 27.5 21.1**

**;**

**run;**

**/\* Look at original fit \*/**

**proc reg data=bodyfat;**

**model body = triceps thigh midarm / vif;**

**title1 'Bodyfat Regression (original fit)';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | Bodyfat Regression (original fit) |      | **Parameter Estimates** | | | | | | | | --- | --- | --- | --- | --- | --- | --- | | **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** | | **Intercept** | **1** | 117.08469 | 99.78240 | 1.17 | 0.2578 | 0 | | **triceps** | **1** | 4.33409 | 3.01551 | 1.44 | 0.1699 | 708.84291 | | **thigh** | **1** | -2.85685 | 2.58202 | -1.11 | 0.2849 | 564.34339 | | **midarm** | **1** | -2.18606 | 1.59550 | -1.37 | 0.1896 | 104.60601 | |

**/\* Try ridge regression as a remedial measure \*/**

**proc reg data=bodyfat ridge=0 to .04 by .005**

**outvif outest=ridgests**

**plots(only)=ridge(VIFaxis=log);**

**model body = triceps thigh midarm / vif;**

**title1 'Bodyfat Ridge Regression';**

**title2 '(find adequate ridge parameter)';**

**run;**

**/\* What these options do:**

**ridge=0 to .04 by 0.005**

**run a regression with each of these ridge parameter**

**values**

**outvif outest=ridgests**

**ask for relevant output to be sent to a data set**

**called ridgests (will include VIF and standardized**

**coefficients for each ridge parameter)**

**plots(only)=ridge(VIFaxis=log);**

**make Ridge Trace and VIF plots only, with vertical axis**

**in VIF plot on log scale**

**\*/**

|  |  |  |
| --- | --- | --- |
| |  | | --- | | ***Bodyfat Ridge Regression*** | | ***(find adequate ridge parameter)*** | |

**/\* Now look at variable coeffs with ridge parameter 0.02 \*/**

**proc reg data=bodyfat outest=ridgenew outseb ridge=0.02**

**outvif noprint;**

**model body = triceps thigh midarm;**

**title1 'Bodyfat Ridge Regression (c=.02)';**

**run;**

**proc print data=ridgenew;**

**var \_type\_ \_rmse\_ triceps thigh midarm;**

**title1 'Ridge Estimates for Variable Coefficients,';**

**title2 'with ridge parameter c = 0.02';**

**run;**

**/\* PARMS and SEB give the result of the regular OLS regression.**

**RIDGE and RIDGESEB give the result of the ridge regression.**

**-- Note no intercept is given; need to use textbook**

**equation 7.46b to get intercept in ridge reg. (as below)**

**Note substantial drop in SE for estimates in ridge reg.**

**RIDGEVIF give the VIF after ridge regression.**

**\*/**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | R***idge Estimates for Variable Coefficients,*** | | ***with ridge parameter c = 0.02*** |  | **Obs** | **\_TYPE\_** | **\_RMSE\_** | **triceps** | **thigh** | **midarm** | | --- | --- | --- | --- | --- | --- | | **1** | PARMS | 2.47998 | 4.33409 | -2.85685 | -2.18606 | | **2** | SEB | 2.47998 | 3.01551 | 2.58202 | 1.59550 | | **3** | RIDGEVIF | . | 1.10255 | 1.08054 | 1.01051 | | **4** | RIDGE | 2.59924 | 0.55535 | 0.36814 | -0.19163 | | **5** | RIDGESEB | 2.59924 | 0.12465 | 0.11841 | 0.16436 | |

**/\* Get intercept term in ridge regression \*/**

**proc means data=bodyfat mean;**

**var body triceps thigh midarm;**

**title1 'Summary Statistics';**

**run;**

**data temp;**

**b0 = 20.195 - 0.55535\*25.305 - 0.36814\*51.17 + 0.19163\*27.62;**

**proc print data=temp;**

**var b0;**

**title1 'Ridge Regression Intercept';**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Summary Statistics*** |  | **Variable** | **Mean** | | --- | --- | | |  | | --- | | **body** | | **triceps** | | **thigh** | | **midarm** | | |  | | --- | | 20.1950000 | | 25.3050000 | | 51.1700000 | | 27.6200000 | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | ***Ridge Regression Intercept*** |  | **Obs** | **b0** | | --- | --- | | **1** | -7.40303 | |

**/\* Get predicted values in ridge regression \*/**

**data bodyfat; set bodyfat;**

**predbody = -7.40303 + 0.55535\*triceps**

**+ 0.36814\*thigh - 0.19163\*midarm;**

**proc sgplot data=bodyfat;**

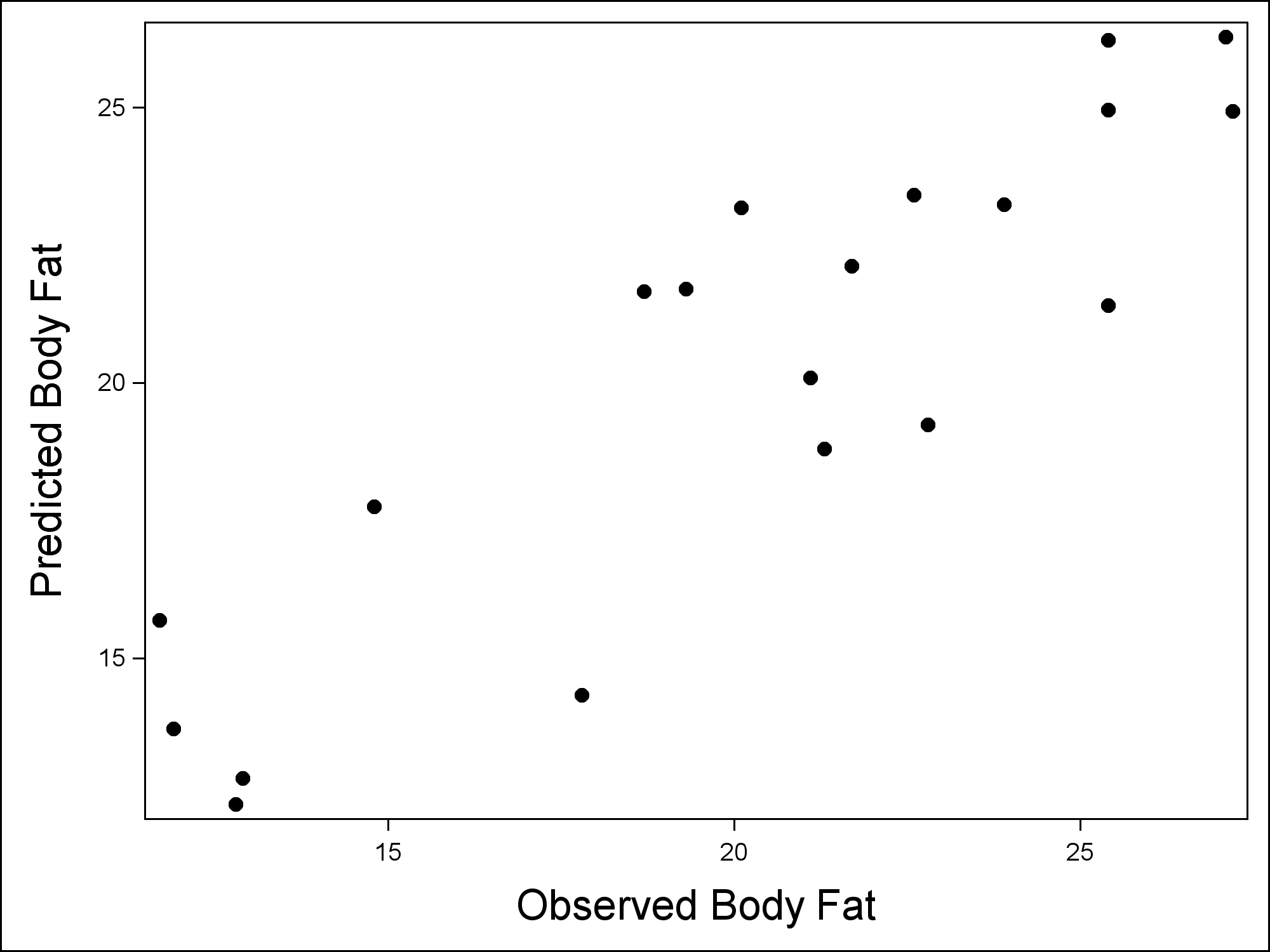
**scatter x=body y=predbody / markerattrs=(symbol=CIRCLEFILLED);**

**xaxis label='Observed Body Fat' labelattrs=(size=15pt);**

**yaxis label='Predicted Body Fat' labelattrs=(size=15pt);**

**title1;**

**run;**



Example 2: (Baseball) This data set (from the SAS Help) contains salary (for 1987) and performance (1986 and some career) data for 322 MLB players who played at least one game in both 1986 and 1987 seasons, excluding pitchers. How can salary be predicted from performance?

**data baseball; set sashelp.baseball;**

**proc contents varnum data=baseball;**

**ods select position;**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | **Variables in Creation Order** | | | | | | --- | --- | --- | --- | --- | | **#** | **Variable** | **Type** | **Len** | **Label** | | **1** | Name | Char | 18 | Player's Name | | **2** | Team | Char | 14 | Team at the End of 1986 | | **3** | nAtBat | Num | 8 | Times at Bat in 1986 | | **4** | nHits | Num | 8 | Hits in 1986 | | **5** | nHome | Num | 8 | Home Runs in 1986 | | **6** | nRuns | Num | 8 | Runs in 1986 | | **7** | nRBI | Num | 8 | RBIs in 1986 | | **8** | nBB | Num | 8 | Walks in 1986 | | **9** | YrMajor | Num | 8 | Years in the Major Leagues | | **10** | CrAtBat | Num | 8 | Career Times at Bat | | **11** | CrHits | Num | 8 | Career Hits | | **12** | CrHome | Num | 8 | Career Home Runs | | **13** | CrRuns | Num | 8 | Career Runs | | **14** | CrRbi | Num | 8 | Career RBIs | | **15** | CrBB | Num | 8 | Career Walks | | **16** | League | Char | 8 | League at the End of 1986 | | **17** | Division | Char | 8 | Division at the End of 1986 | | **18** | Position | Char | 8 | Position(s) in 1986 | | **19** | nOuts | Num | 8 | Put Outs in 1986 | | **20** | nAssts | Num | 8 | Assists in 1986 | | **21** | nError | Num | 8 | Errors in 1986 | | **22** | Salary | Num | 8 | 1987 Salary in $ Thousands | | **23** | Div | Char | 16 | League and Division | | **24** | logSalary | Num | 8 | Log Salary | |

**/\* lasso \*/**

**proc glmselect data=baseball plots=(criterion ase);**

**class league division;**

**model logSalary = nAtBat nHits nHome nRuns nRBI nBB**

**yrMajor crAtBat crHits crHome crRuns crRbi**

**crBB league division nOuts nAssts nError**

**/ selection=lasso(adaptive choose=sbc stop=none);**

**output out=out1 p=predlasso;**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | *Data Set* | WORK.BASEBALL | | *Dependent Variable* | logSalary | | *Selection Method* | Adaptive LASSO | | *Stop Criterion* | None | | *Choose Criterion* | SBC | | *Effect Hierarchy Enforced* | None |  |  |  | | --- | --- | | *Number of Observations Read* | 322 | | *Number of Observations Used* | 263 |  | *Class Level Information* | | | | --- | --- | --- | | *Class* | *Levels* | *Values* | | *League* | 2 | American National | | *Division* | 2 | East West | | | *LASSO Selection Summary* | | | | | | --- | --- | --- | --- | --- | | *Step* | *Effect Entered* | *Effect Removed* | *Number Effects In* | *SBC* | | *\* Optimal Value of Criterion* | | | | | | *0* | *Intercept* |  | 1 | -57.2041 | | *1* | *CrRuns* |  | 2 | -70.8348 | | *2* | *nHits* |  | 3 | -226.0696 | | *3* | *CrHits* |  | 4 | -238.6648 | | *4* | *YrMajor* |  | 5 | -248.4971 | | *5* | *nBB* |  | 6 | -260.5682\* | | *6* | *Division\_East* |  | 7 | -257.7020 | | *7* | *nOuts* |  | 8 | -254.3352 | | *8* |  | *CrRuns* | 7 | -260.1040 | | *9* | *nError* |  | 8 | -254.9990 | | *10* | *CrBB* |  | 9 | -249.9243 | | *11* | *CrRuns* |  | 10 | -245.7008 | | *12* | *nAtBat* |  | 11 | -241.6564 | | *13* | *nHome* |  | 12 | -236.3245 | | *14* | *League\_American* |  | 13 | -238.1068 | | *15* | *CrAtBat* |  | 14 | -234.0015 | | *16* |  | *CrHits* | 13 | -241.0870 | | *17* | *nAssts* |  | 14 | -235.9894 | | *18* | *CrHits* |  | 15 | -230.5456 | | *19* | *nRuns* |  | 16 | -225.5197 | | *20* | *nRBI* |  | 17 | -220.3634 | | *21* | *CrRbi* |  | 18 | -214.7952 | | *22* | *CrHome* |  | 19 | -209.2505 |   Selection stopped because all candidate effects for entry are linearly dependent on effects in the model. |

|  |  |
| --- | --- |
|  | |
| ***Selected Model***  ***The selected model, based on SBC, is the model at Step 5.***   |  |  | | --- | --- | | *Root MSE* | 0.57845 | | *Dependent Mean* | 5.92722 | | *R-Square* | 0.5849 | | *Adj R-Sq* | 0.5768 | | *AIC* | -17.00115 | | *AICC* | -16.56194 | | *SBC* | -260.56823 | | | *Parameter Estimates* | | | | --- | --- | --- | | *Parameter* | *DF* | *Estimate* | | *Intercept* | 1 | 4.229778 | | *nHits* | 1 | 0.007194 | | *nBB* | 1 | 0.005629 | | *YrMajor* | 1 | 0.062808 | | *CrHits* | 1 | 0.000222 | | *CrRuns* | 1 | 0.000136 | |

**/\* elastic net \*/**

**proc glmselect data=out1 plots=(criterion ase) seed=12;**

**class league division;**

**model logSalary = nAtBat nHits nHome nRuns nRBI nBB**

**yrMajor crAtBat crHits crHome crRuns crRbi**

**crBB league division nOuts nAssts nError**

**/ selection=elasticnet(stop=none choose=cv)**

**cvmethod=random(20);**

**output out=out2 p=predelasticnet;**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | *Data Set* | WORK.OUT1 | | *Dependent Variable* | logSalary | | *Selection Method* | ELASTICNET | | *Stop Criterion* | None | | *Choose Criterion* | Cross Validation | | *Cross Validation Method* | Random | | *Cross Validation Fold* | 20 | | *Effect Hierarchy Enforced* | None | | *Random Number Seed* | 12 |  |  |  | | --- | --- | | *Number of Observations Read* | 322 | | *Number of Observations Used* | 263 |  | *Class Level Information* | | | | --- | --- | --- | | *Class* | *Levels* | *Values* | | *League* | 2 | American National | | *Division* | 2 | East West | | | *Elastic Net Selection Summary* | | | | | | --- | --- | --- | --- | --- | | *Step* | *Effect Entered* | *Effect Removed* | *Number Effects In* | *CV PRESS* | | *\* Optimal Value of Criterion* | | | | | | *0* | *Intercept* |  | 1 | 209.2326 | | *1* | *CrRuns* |  | 2 | 123.1776 | | *2* | *CrHits* |  | 3 | 123.7433 | | *3* | *nHits* |  | 4 | 97.6956 | | *4* | *nBB* |  | 5 | 94.7216 | | *5* | *CrRbi* |  | 6 | 98.1015 | | *6* | *YrMajor* |  | 7 | 92.7082 | | *7* | *nRBI* |  | 8 | 94.5500 | | *8* | *Division\_East* |  | 9 | 93.3921 | | *9* | *nOuts* |  | 10 | 94.1530 | | *10* | *nError* |  | 11 | 93.8913 | | *11* | *nHome* |  | 12 | 94.2533 | | *12* | *League\_American* |  | 13 | 94.4968 | | *13* |  | *CrRbi* | 12 | 90.7314 | | *14* |  | *CrRuns* | 11 | 90.1957 | | *15* |  | *nRBI* | 10 | 89.6571 | | *16* | *CrBB* |  | 11 | 89.2733 | | *17* | *CrRuns* |  | 12 | 89.4515 | | *18* | *nAtBat* |  | 13 | 88.9017\* | | *19* | *CrAtBat* |  | 14 | 89.2818 | | *20* | *nAssts* |  | 15 | 89.7926 | | *21* | *CrHome* |  | 16 | 91.8598 | | *22* | *nRBI* |  | 17 | 92.6309 | | *23* | *nRuns* |  | 18 | 93.1973 | | *24* | *CrRbi* |  | 19 | 94.5881 |   *\* Optimal Value of Criterion*  Selection stopped because all candidate effects for entry are linearly dependent on effects in the model. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ***Selected Model***  ***The selected model, based on Cross Validation, is the model at Step 18.***   |  |  | | --- | --- | | *Root MSE* | 0.56923 | | *Dependent Mean* | 5.92722 | | *R-Square* | 0.6090 | | *Adj R-Sq* | 0.5902 | | *AIC* | -18.72037 | | *AICC* | -17.02682 | | *SBC* | -237.28237 | | *CV PRESS* | 88.90168 |  | *Parameter Estimates* | | | | --- | --- | --- | | *Parameter* | *DF* | *Estimate* | | *Intercept* | 1 | 4.195962 | | *nAtBat* | 1 | -0.000112 | | *nHits* | 1 | 0.006807 | | *nHome* | 1 | 0.003545 | | *nBB* | 1 | 0.007082 | | *YrMajor* | 1 | 0.070194 | | *CrHits* | 1 | 0.000247 | | *CrRuns* | 1 | 0.000212 | | *CrBB* | 1 | -0.000348 | | *League\_American* | 1 | -0.092575 | | *Division\_East* | 1 | 0.144062 | | *nOuts* | 1 | 0.000192 | | *nError* | 1 | -0.007767 | |

**proc sgscatter data=out2;**

**matrix logSalary predlasso predelasticnet /**

**markerattrs=(symbol=circlefilled size=6pt);**

**run;**

