**4.4.1- SAS: Nonparametric Regression Methods**

**(LOESS, Regression Trees, and Random Forests)**

Example: (Baseball, same as Handout 4.1.1)

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

**AmerLg = (League="American");**

**EastDv = (Division="East");**

**run;**

**/\* loess \*/**

**proc loess data=baseball plots=(fitpanel fitplot contourfitpanel contourfit);**

**model logSalary = crAtBat nBB**

**/ degree=2 select=AICC scale=sd;**

**output out=out1 p=predloess;**

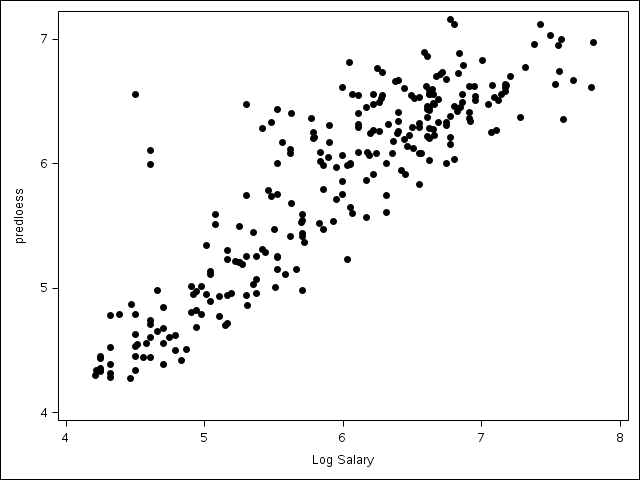
**run;**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***The LOESS Procedure***     | *Optimal Smoothing Criterion* | | | --- | --- | | *AICC* | *Smoothing Parameter* | | -0.57876 | 0.60646 | |

**proc sgplot data=out1;**

**scatter x=logSalary y=predloess / markerattrs=(symbol=circlefilled size=6pt);**

**run;**



**/\* regression tree \*/**

**proc hpsplit data=baseball seed=123 maxdepth=15 maxbranch=2;**

**class league division;**

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

**yrMajor crAtBat crHits crHome crRuns crRbi**

**crBB league division nOuts nAssts nError;**

**output out=out2;**

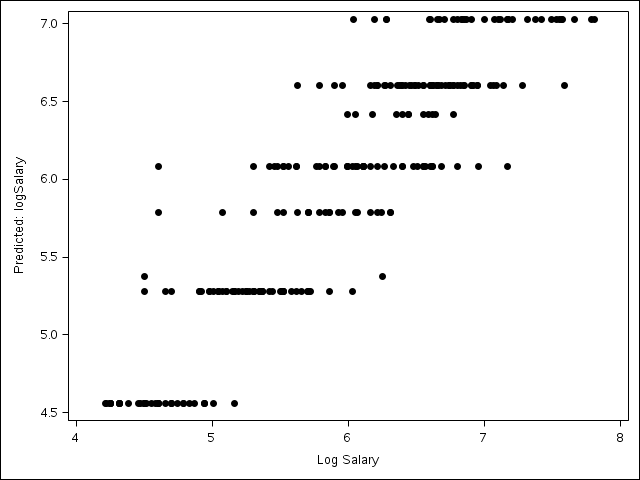
**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***The HPSPLIT Procedure***         | *Model-Based Fit Statistics for Selected Tree* | | | | --- | --- | --- | | *N Leaves* | *ASE* | *RSS* | | 8 | 0.1443 | 37.9587 |  | *Variable Importance* | | | | | | --- | --- | --- | --- | --- | | *Variable* | *Variable Label* | *Training* | | *Count* | | *Relative* | *Importance* | | *CrAtBat* | Career Times at Bat | 1.0000 | 11.2539 | 1 | | *nBB* | Walks in 1986 | 0.3546 | 3.9905 | 2 | | *CrRbi* | Career RBIs | 0.3414 | 3.8415 | 2 | | *nAtBat* | Times at Bat in 1986 | 0.2168 | 2.4397 | 1 | | *CrRuns* | Career Runs | 0.2161 | 2.4316 | 1 | |

**proc sgplot data=out2;**

**scatter x=logSalary y=p\_logSalary / markerattrs=(symbol=circlefilled size=6pt);**

**run;**



**Question: What is going on in this plot? Do these patterns in the prediction make sense? If yes, why do they make sense?**

**Question: Recalling Output in Handout #23, what do the “important” variables have in common?**

**/\* random forest \*/**

**proc hpforest data=baseball seed=134 scoreprole=oob;**

**input nAtBat nHits nHome nRuns nRBI nBB**

**yrMajor crAtBat crHits crHome crRuns crRbi**

**crBB league division nOuts nAssts nError;**

**target Salary;**

**ods output FitStatistics=fitstats**

**VariableImportance=varimp;**

**run;**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The HPFOREST Procedure   |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | | **Model Information** | | | | --- | --- | --- | | **Parameter** | **Value** |  | | **Variables to Try** | 4 | (Default) | | **Maximum Trees** | 100 | (Default) | | **Missing Value Handling** | . | Valid value | | | **Number of Observations** | | | --- | --- | | **Type** | **N** | | **Number of Observations Read** | 322 | | **Number of Observations Used** | 263 | |  | **Loss Reduction Variable Importance** | | | | | | | --- | --- | --- | --- | --- | --- | | **Variable** | **Number of Rules** | **MSE** | **OOB MSE** | **Absolute Error** | **OOB Absolute Error** | | **CrHits** | 907 | 27941.87 | 20687.57 | 48.803825 | 34.608172 | | **CrRbi** | 1160 | 22995.54 | 12521.15 | 35.533126 | 19.290786 | | **CrRuns** | 1072 | 23108.48 | 10892.41 | 39.211686 | 18.379497 | | **CrAtBat** | 751 | 18859.52 | 10140.97 | 32.764124 | 20.230476 | | **CrBB** | 1364 | 16893.90 | 4896.42 | 31.277359 | 11.410166 | | **nBB** | 606 | 12942.85 | 4625.19 | 14.772798 | 3.751437 | | **CrHome** | 804 | 13002.18 | 3062.38 | 18.501506 | 4.823677 | | **nHits** | 439 | 10636.46 | 2314.45 | 14.907649 | 3.961956 | | **YrMajor** | 455 | 5866.65 | 471.24 | 11.912504 | 2.927752 | | **nAtBat** | 414 | 10120.05 | 199.98 | 14.692048 | 0.552953 | | **Division** | 9 | 355.44 | -102.12 | 0.373370 | -0.103367 | | **League** | 15 | 117.50 | -174.16 | 0.244754 | -0.153395 | | **nRBI** | 572 | 11899.64 | -352.58 | 15.151606 | -0.354135 | | **nRuns** | 497 | 8491.47 | -1336.94 | 11.766502 | -0.471976 | | **nHome** | 423 | 5302.24 | -1882.58 | 8.979994 | -0.764283 | | **nError** | 1755 | 4534.88 | -3505.17 | 13.465747 | -3.311704 | | **nAssts** | 1582 | 3494.33 | -4257.11 | 12.493737 | -3.871985 | | **nOuts** | 1802 | 9530.72 | -4815.96 | 21.164897 | -4.546558 | |

**Question: What does it mean to have a negative out of bag mean square error? What does this provide evidence for?**

**data varimp; set varimp;**

**VarOrder=\_n\_;**

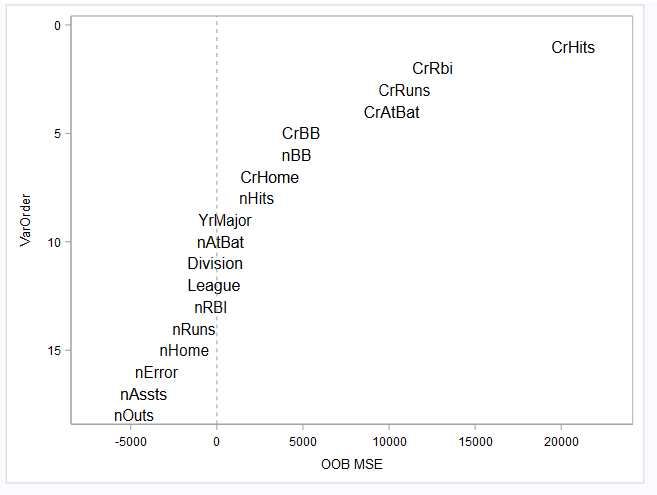
**proc sgplot data=varimp;**

**scatter x=MSEOOB y=VarOrder / markerchar=Variable markercharattrs=(size=12);**

**yaxis reverse;**

**refline 0 / axis = x LINEATTRS=(pattern=2);**

**run;**



**/\* Visualize effects of top predictors using a generalized additive model \*/**

**proc gampl data=baseball plots(unpack)=all;**

**model logSalary = s(crHits) s(CrRbi) s(CrRuns) s(CrAtBat)**

**/ dist=norm;**

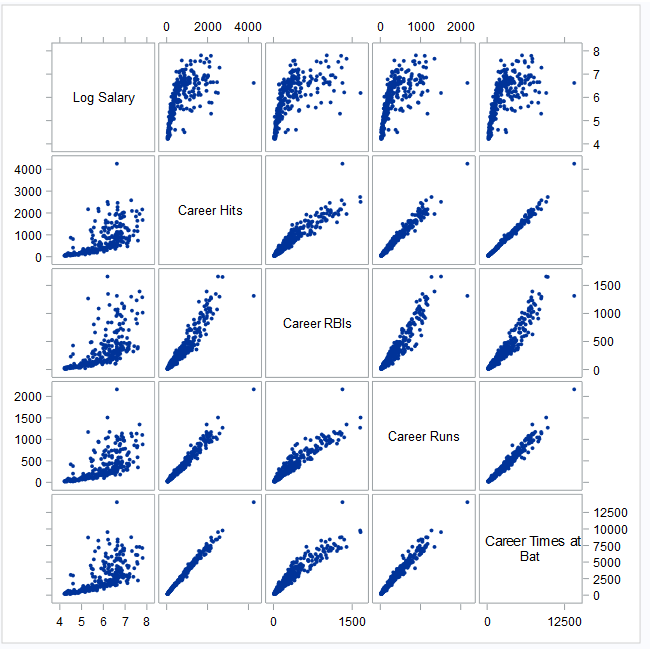
**run;**

|  |  |
| --- | --- |
| The GAMPL Procedure | |
|  |  |
|  |  |

**/\* Compare with simple scatter plot \*/**

**proc sgscatter data=baseball;**

**matrix logSalary crHits crRBI**

 **crRuns crAtBat /**

**markerattrs=(**

**symbol=CIRCLEFILLED**

**size=6pt);**

**run;**