Cross Validation

Friday, June 15, 2018

8:49 PM

We ran 3 different Cross Validation models.

All 3 methods produced the same model selection.

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*Cross Validation

Cross-validation: method to try to determine if a model will have

good predictive ability, even without a second set of data

Leave K out of validation (use for smaller data sets)

Options:

1. Test set is preferred when a large number of observations,

so that test set represents population well.

2. Leave one out is most computationally intensive.

3. Split and block requires that test sets are representative

of population (no intrinsic order in data).

4. Random may lead to two different researchers producing two

different models on the same data.

\*/

/\*Import dataset with formatted columns\*/

data auto;

infile '/home/carollr0/DataSets/Automobile\_data.csv' dlm=',' firstobs=2;

input symboling make $ fueltype $ aspiration $ numofdoors $

bodystyle $ drivewheels $ enginelocation $ wheelbase length width

height curbweight enginetype $ /\*numofcylinders $ \*/ enginesize fuelsystem $ bore

stroke compressionratio horsepower peakrpm citympg highwaympg price;

run;

/\*Remove Normalized Losses feature - too many missing data points, not reliable\*/

data auto2;

set auto;

keep symboling make fueltype aspiration numofdoors

bodystyle drivewheels enginelocation wheelbase length width

height curbweight enginetype numofcylinders enginesize fuelsystem bore

stroke compressionratio horsepower peakrpm citympg highwaympg price;

run;

/\*Delete all rows with missing data - 199 records left\*/

data auto\_clean;

set auto2;

if nmiss(of \_numeric\_, 1) + cmiss(of \_character\_, '?') then delete;

run;

/\*Observations 47 is shown to be leverage and outliers. This observations will be removed.\*/

data auto\_remobs;

set auto\_clean;

if \_N\_=47 then delete;

run;

quit;

/\*Leave one out CV

Effects: Intercept width enginesize stroke compressionratio horsepower peakrpm

Root MSE 3026.09372

Dependent Mean 13131

R-Square 0.8581

Adj R-Sq 0.8536

AIC 3312.70140

AICC 3313.47978

PRESS 1873371303

SBC 3139.57641

\*/

proc glmselect data=auto\_remobs;

model price = symboling wheelbase length width height curbweight enginesize bore

stroke compressionratio horsepower peakrpm citympg highwaympg /

selection=stepwise(stop=press);

run;

/\*larger press--->increase bias, decease in variablility-fairly stable model

/\*10 fold Cross Validation

Effects: Intercept width enginesize stroke compressionratio horsepower peakrpm

Root MSE 3026.09372

Dependent Mean 13131

R-Square 0.8581

Adj R-Sq 0.8536

AIC 3312.70140

AICC 3313.47978

SBC 3139.57641

CV PRESS 2002069260

\*/

proc glmselect data=auto\_remobs;

model price = symboling wheelbase length width height curbweight enginesize bore

stroke compressionratio horsepower peakrpm citympg highwaympg /

selection=stepwise(choose=cv) ;

run;

/\*5 fold Cross Validation 1/5 of obs will be taken out at a time to test

Effects: Intercept width enginesize stroke compressionratio horsepower peakrpm

Root MSE 3026.09372

Dependent Mean 13131

R-Square 0.8581

Adj R-Sq 0.8536

AIC 3312.70140

AICC 3313.47978

SBC 3139.57641

CV PRESS 1871729861

\*/

proc glmselect data=auto\_remobs;

model price = symboling wheelbase length width height curbweight enginesize bore

stroke compressionratio horsepower peakrpm citympg highwaympg /

selection=forward(choose=cv) CVMETHOD=random(5);

run;

/\*Using a stopping criterion, by default SAS does a 5fold cross validation (1/5 of the data at random=CVMETHOD=random(10)), effect that gives the best value of the criterion

Effects: Intercept width enginesize stroke compressionratio horsepower peakrpm

Root MSE 3026.09372

Dependent Mean 13131

R-Square 0.8581

Adj R-Sq 0.8536

AIC 3312.70140

AICC 3313.47978

SBC 3139.57641

CV PRESS 1938077509

\*/

proc glmselect data=auto\_remobs;

model price = symboling wheelbase length width height curbweight enginesize bore

stroke compressionratio horsepower peakrpm citympg highwaympg /

selection=stepwise(stop=cv);

run;

/\*all 3 cross validation methods end up with the same model each time\*/

/\*Check\*/

proc glm data=work.auto\_remobs plots=all;

Model price= width enginesize stroke compressionratio horsepower peakrpm/ cli solution;

output out = resultsC p=Predicted;

run;

/\*Leave one out CV

Effects: Intercept width enginesize stroke compressionratio horsepower peakrpm

Root MSE 3026.09372

Dependent Mean 13131

R-Square 0.8581

Adj R-Sq 0.8536

AIC 3312.70140

AICC 3313.47978

PRESS 1873371303

SBC 3139.57641

\*/

proc glmselect data=auto\_remobs;

model price = symboling wheelbase length width height curbweight enginesize bore

stroke compressionratio horsepower peakrpm citympg highwaympg /

selection=stepwise(stop=press);

run;

Machine generated alternative text:
Effæty 
GLMSELECT 
is at lut step (Step 
"Sdth 
188±3 
187 
Total 
1712404478 
12071143070 
g 157243 
o.u.81 
De n 
Adj R-sq 
Al cc 
PRESS 
Inte ræpt 
ratio 
331270140 
387 
10.38 
273 
2.2t743t 
Sta rd 
13323721 
7473726" 
13.42030 
oz02731 

/\*larger press--->increase bias, decease in variablility-fairly stable model

/\*10 fold Cross Validation

Effects: Intercept width enginesize stroke compressionratio horsepower peakrpm

Root MSE 3026.09372

Dependent Mean 13131

R-Square 0.8581

Adj R-Sq 0.8536

AIC 3312.70140

AICC 3313.47978

SBC 3139.57641

CV PRESS 2002069260

\*/

proc glmselect data=auto\_remobs;

model price = symboling wheelbase length width height curbweight enginesize bore

stroke compressionratio horsepower peakrpm citympg highwaympg /

selection=stepwise(choose=cv) ;

run;

GLMSELECT 
Model 
mdel. Cross is mdel at Step 6_ 
Effæty "Sdth r.tjc *3krpm 
187 
Total 
of 
S qua ru 
1712404478 
12071143070 
Square 
g 157243 
085-81 
De n 
Adj R-sq 
Al cc 
CV PRESS 
p rabo 
331270140 
188±3 
387 
10.38 
273 
22t743t 
Sta rd 
13323721 
7473726" 
13.42030 
oz02731 

/\*5 fold Cross Validation 1/5 of obs will be taken out at a time to test

Effects: Intercept width enginesize stroke compressionratio horsepower peakrpm

Root MSE 3026.09372

Dependent Mean 13131

R-Square 0.8581

Adj R-Sq 0.8536

AIC 3312.70140

AICC 3313.47978

SBC 3139.57641

CV PRESS 1871729861

\*/

proc glmselect data=auto\_remobs;

model price = symboling wheelbase length width height curbweight enginesize bore

stroke compressionratio horsepower peakrpm citympg highwaympg /

selection=forward(choose=cv) CVMETHOD=random(5);

run;

E ュ 鬻 〔 ゴ 「 富 響 ・ 黔 h 当 g 、 な 璽 日 贏 ョ を 2 を ョ 国 & Ⅲ 0W2 「 讐 「 9 ョ 
ⅲ d ョ ロ del. b 国 52d00 C 「 055 、 da ロ 0. 五 呼 2 ョ ロ de 一 耳 Step 円 
c ロ ヨ を 2 ロ っ 0 1 2 の 3 朝 4332D 讐 3753D5 
C ロ - 「 2 d To 一 1 の 3 12D711 3070 
CV PRESS 
ACC 
早 R Sq 
R ・ Sq に 国 「 2 
p 当 d 当 t 、 2 当 
GLMSELECT P - ロ ædu 
187 1712404478 
103 「 8738522 172 4 432 18 円 53 
P 国 「 ョ 県 E ⅲ ョ 国 5 
コ 竺 五 Of 国 ュ 国 コ 
1 ・ 371 22D455 747 朝 72555 ・ 4 7 
S. - d ′ - ロ d 竺 
141 い 22785 1 響 525721 
3 朝 43215 1 響 42D353 273 
Squ 国 「 25 
2257435 05D2731 
of 
1 の 12431821 
313 叫 57 41 
331 47 の 78 
331 卩 701 0 
CIE 「 81 
13131 
の 157243 
S コ d 国 「 d 
Squa 「 2 F 竺 
E 「 「 0 「 - 《 国 ど e 
1038 

/\*Using a stopping criterion, by default SAS does a 5fold cross validation (1/5 of the data at random=CVMETHOD=random(10)), effect that gives the best value of the criterion

Effects: Intercept width enginesize stroke compressionratio horsepower peakrpm

Root MSE 3026.09372

Dependent Mean 13131

R-Square 0.8581

Adj R-Sq 0.8536

AIC 3312.70140

AICC 3313.47978

SBC 3139.57641

CV PRESS 1938077509

\*/

proc glmselect data=auto\_remobs;

model price = symboling wheelbase length width height curbweight enginesize bore

stroke compressionratio horsepower peakrpm citympg highwaympg /

selection=stepwise(stop=cv);

run;

ョ 0 「 ロ ミ 旱 0 一 2 お 332D ま い 753 
CO - 「 ま d Total 1 お 12D711 010 
Re 望 、 崧 
7 訂 GLMSELECT p 言 u ー 
い 一 0 鬻 8 に 鬻 02 二 4 嚮 い 8 に 
、 ~ 一 五 象 Va ュ 当 
3 04 〕 715 
E ョ 動 
S を 当 第 u 響 ・ 
い 313 4 
い 31 い 70 一 呂 

/\*all 3 cross validation methods end up with the same model each time

This is good because it gives us the best model using different criterion but with the Stepwise selection model\*/

/\*Check\*/

proc glm data=work.auto\_remobs plots=all;

Model price= width enginesize stroke compressionratio horsepower peakrpm/ cli solution;

output out = resultsC p=Predicted;

run;

The GLMProcedure 
NumberofObservations Read 1 

山 0 元 》 2 製 0 - 2 ! d np 、 u - 言 」 一 u 」 2 い u 山 240 ロ い * の 6 - e コ 型 5 山 工 
望 ← 心 28 ← 
望 ← 心 28 ← 
93 」 E に 」 80 ← 
E に に に 円 
一 O 」 国 一 噐 富 
! n' 働 00 」 d 襾 - 24 ト 
富 国 国 02 
O 」 朝 O 」 朝 52 ← 00000 00 2 ← 
円 」 」 」 33 」 朝 円 2 ← 000001 」 朝D「 
国 」 3 新 円 5 「 3E3D9 イ 「 8 」 「 00000 02 」 0 円 
3 」 O O 鬲 」 新 2 国 国 心 円 円 00000 0882 円 
朝 響 00000•z 」 新明 
2 」 新 339 菖 鬲 8 円 」 」 
3 イ 」 い 339 O 鬲 8 円 」 
38BO 朝 2 「 3 朝 円 00000 朝 」 2 
3 冒 ← 尊 3 冒 鬲 」 ← い 00000 新 82 
O 5 ← に 「 00000 朝 5 円 
新 円 「 」 00000 朝 「 瀛 
080C 朝 8 OEO い 」 朝 00000 新 58 」 
080; 朝 響 080 い 」 朝 00000 新 5D5 
080; 朝 響 080 い 」 朝 00000 新 88 
2 製 0 - 2 ! d 
」 」 に 朝 円 円 ← 00000 00 2 ← 
u04 市 と 

Sum of 
Sum of 
Sum of - Effor SS 
PRESS statistic 
1712404478 
1873371303 
Durbin-Watnn O 
Fit Diagnostics for price 
1 ocno 
scoo 
-3 
20900 40000 
Predicted Value 
Quantile 
Residual 
Predicted Value 
20000 
predicted Value 
an Residual 
Proportion Less 
y, 
0:10 
000 oos 010 01S 
everage 
SO ISO 
Observation 
194 
Parameters 
E„or 
9.166 
R. square 0.8581 
Adj *Square 08536 
020 
200 

Residual Plots for price 
600 
62 s 
650 
67. s 
Width 
stroke 
70.0 
72S 
so 
200 
engine size 
compressionratio 
250 

Residual Plots for price 
200 
horseponer 
4000 
4500 
SSOO 
peakrpm 
6000 
6sco 

The model

Price = -51,910.74 + 620.61(width) + 141.40(engine size) -3716.22(stroke) + 293.54(compression ratio) + 36.64 (horsepower) + 2.27(peak rpm)

Comparison to 1st model selection

The sum of squared residuals is smaller for the 2nd model selection which means the residuals are closer to 0.

The Press Statistic is lower. Not sure if we learned what this means so I looked up the definition

PRESS statistic

From Wikipedia, the free encyclopedia

[Jump to navigation](https://en.wikipedia.org/wiki/PRESS_statistic#mw-head)[Jump to search](https://en.wikipedia.org/wiki/PRESS_statistic#p-search)

In [statistics](https://en.wikipedia.org/wiki/Statistics), the **predicted residual error sum of squares (PRESS) statistic** is a form of [cross-validation](https://en.wikipedia.org/wiki/Cross-validation_(statistics)) used in [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) to provide a summary measure of the fit of a model to a sample of observations that were not themselves used to estimate the model. It is calculated as the sums of squares of the prediction residuals for those observations.[[1]](https://en.wikipedia.org/wiki/PRESS_statistic#cite_note-1)[[2]](https://en.wikipedia.org/wiki/PRESS_statistic#cite_note-2)[[3]](https://en.wikipedia.org/wiki/PRESS_statistic#cite_note-3)

A *fitted model* having been produced, each observation in turn is removed and the model is refitted using the remaining observations. The out-of-sample predicted value is calculated for the omitted observation in each case, and the PRESS statistic is calculated as the sum of the squares of all the resulting prediction errors:[[4]](https://en.wikipedia.org/wiki/PRESS_statistic#cite_note-4)

Given this procedure, the PRESS statistic can be calculated for a number of candidate model structures for the same dataset, with the lowest values of PRESS indicating the best structures. Models that are over-parameterised ([over-fitted](https://en.wikipedia.org/wiki/Overfitting)) would tend to give small residuals for observations included in the model-fitting but large residuals for observations that are excluded.

From <[*https://en.wikipedia.org/wiki/PRESS\_statistic*](https://en.wikipedia.org/wiki/PRESS_statistic)>

Adj R Square is a few points higher from 82.229 to 85.36