A Look at Information Criteria for Regression Models using SAS®

Jonas V. Bilenas

March 18, 2020

The standard for regression modeling selection has drifted from R-Square and Adjusted R-Square to Information Criteria metrics; AIC, AICC, and SBC which will be reviewed in the quick blog post. We will look at a simulated data and build a regression model using PROC GENMOD, PROC REG, and PROC GLMSELECT. Note that there are other regressions procedures in SAS which will also generate the desired metrics. I have also notice that metrics may change depending on the procedure being used which we will identify in this document.

Metric Calculations:

- **AIC**, Akaike Information Criteria, was developed by Akaike(1973) is a function of the number of observations n, number of regression terms (k) , and model Log Likelihood (LL). For normal distributions you can also use also use Sum of Squares Error (SSE). The lower the AIC the better the model fit.

$$\boldsymbol{AIC = -2 * LL + 2 * k}$$

For normal distribution, this can be used:

$$\boldsymbol{AIC = n * LN\left(\frac{SSE}{n}\right) + 2 * k}$$

- **AICC** is used (Hurvich & Tsai, 1989) for smaller number of n:

$$\boldsymbol{AICC = -2 * LL + 2 * k * \frac{n}{n - k - 1}}$$

- There is also a Bayesian Information Criteria, **BIC**, developed by Sawa (1978). But a larger penalization is using the Schwarz Bayesian Criteria Schwarz (1978) is further penalization based on n and k. **SBC:**

$$\boldsymbol{SBC = -2 * LL + k * LN(n)}$$

For normal distribution, this can be used:

$$\boldsymbol{SBC = n * LN\left(\frac{SSE}{n}\right) + k * LN(n)}$$

The next section we will look at predicting weight based on sex and height to a modified data set from SASHELP.CLASS dataset where n=190.
Data generation:

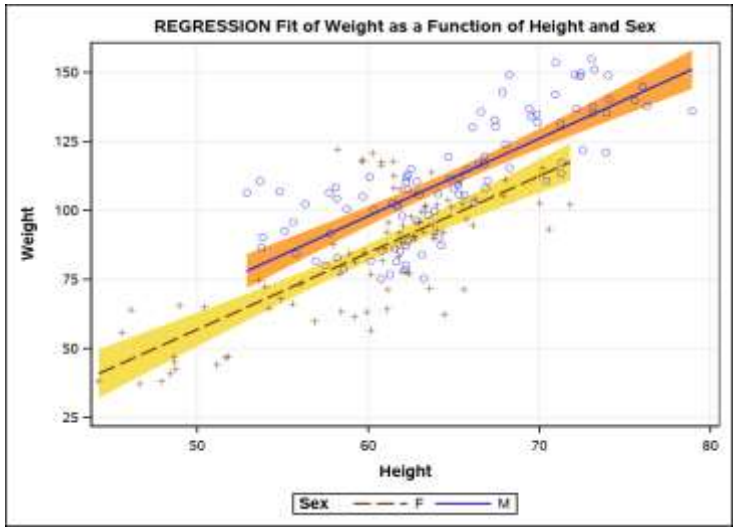
```
data class;
  set sashelp.class;
  call streaminit(20171105);
  do enhance=1 to 10;
     height = height+rand("normal")*2;
     weight = weight+rand("Normal")*4;
     output;
  end;
run;
```

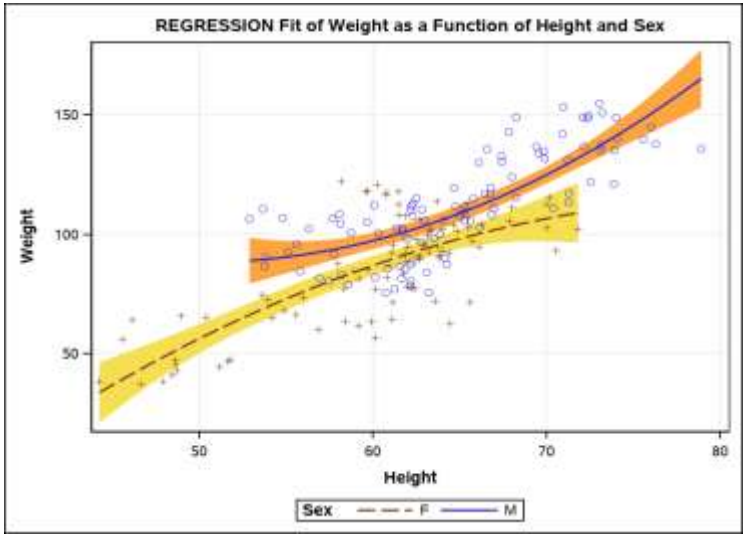
Let’s run some exploratory plots:

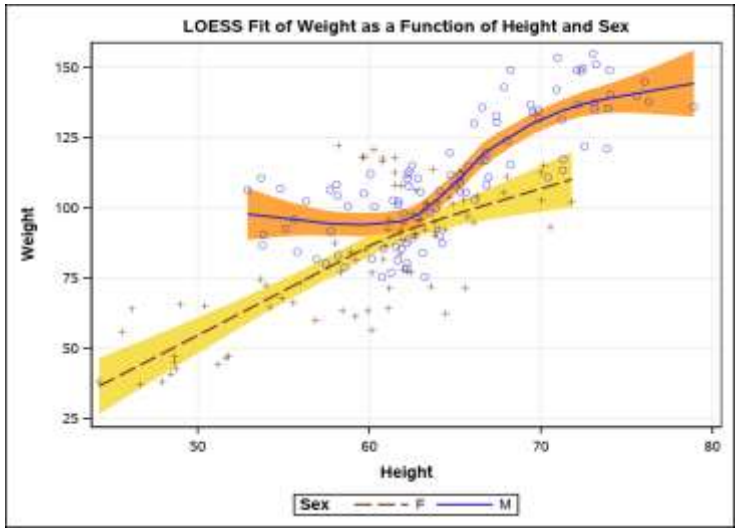
```
proc sgplot data=class;
  REG x=height y=weight
      /group=sex
       CLM alpha=0.05;
  xaxis grid;
  yaxis grid;
  title REGRESSION Fit of Weight as a Function of Height and Sex;
run;

proc sgplot data=class;
  LOESS x=height y=weight
      /group=sex
       CLM alpha=0.05;
  xaxis grid;
  yaxis grid;
  title LOESS Fit of Weight as a Function of Height and Sex;
run;

proc sgplot data=class;
  REG x=height y=weight
      /group=sex degree=2
       CLM alpha=0.05;
  xaxis grid;
  yaxis grid;
  title REGRESSION Fit of Weight as a Function of Height and Sex;
run;
```


REGRESSION Fit of Weight as a Function of Height and Sex
Weight
150
125
100
75
50
25
50
60
70
80
Height
Sex F M


REGRESSION Fit of Weight as a Function of Height and Sex
Weight
150
100
50
50
60
70
80
Height
Sex F M


LOESS Fit of Weight as a Function of Height and Sex
Weight
150
125
100
75
50
25
50
60
70
80
Height
Sex F M

Let's try using a PROC GENMOD to build a regression model using SEX, HEIGHT, and HEIGHT*HEIGHT with up to 2-way interactions and look at the output. Note GENMOD does not provide R-Square but that can be easy to get via ODS OUTPUT and PROC CORR to capture the correlation of Actual Weight and Predicted Weight and taking the square of that correlation. This was not done in the code. I like GENMOD since one can include CLASS variables in the procedure and use the largest level frequency as the reference for the class variables.

```
proc genmod data=class namelen=38;
  class sex/order=freq
          param=ref ref=first missing;
  model weight=height | sex
               height*height
               height*height*sex
     /dist=NOR wald;
  title GENMOD Fit of Weight as a Function of Height and
        Sex;
run; quit;
```

The results are shown on the next page. If we use the Log Likelihood calculation it looks like GENMOD includes the regression parameters including the intercept term and the error term (k=7 as opposed to k=6). If we use the SSE formula and k=6 the results match up to PROC REG and PROC GLMSELECT. Not sure if GENMOD is including the SCALE term which is not part of the final model. Using the Log Likelihood calculation, it looks like BIC is SBC in GENMOD. I generally only use one procedure when model building so I sort of do a manual backward elimination. GENMOD does not offer a selection rule and does not do stepwise which is ok since there are issues with stepwise regressions. See Flom and Cassell (2007).

GENMOD Fit of Weight as a Function of Height and Sex

The GENMOD Procedure

Model Information	
Data Set	WORK.CLASS
Distribution	Normal
Link Function	Identity
Dependent Variable	Weight
Number of Observations Read	190
Number of Observations Used	190

Class Level Information		
Class	Value	Design Variables
Sex	M	0
	F	1

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	184	36690.2858	199.4037
Scaled Deviance	184	190.0000	1.0326
Pearson Chi-Square	184	36690.2858	199.4037
Scaled Pearson X2	184	190.0000	1.0326
Log Likelihood		-769.6064	
Full Log Likelihood		-769.6064	
AIC (smaller is better)		1553.2129	
AICC (smaller is better)		1553.8282	
BIC (smaller is better)		1575.9420	

Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	322.8391	146.1902	36.3115	609.3667	4.88	0.0272
Height		1	-9.3572	4.5039	-18.1846	-0.5298	4.32	0.0377
Sex	F	1	-584.252	180.2199	-937.477	-231.028	10.51	0.0012
Height*Sex	F	1	18.4377	5.8132	7.0440	29.8313	10.06	0.0015
Height*Height		1	0.0933	0.0345	0.0257	0.1609	7.31	0.0068
Height*Height*Sex	F	1	-0.1479	0.0469	-0.2399	-0.0560	9.94	0.0016
Scale		1	13.8963	0.7129	12.5670	15.3661		

Code for PROC REG is shown below:

```
data for_reg;
  set class;
  sexclass    = (sex='F');
  height2     = height*height;
  height2_sex = sexclass*height2;
  height_sex  = height*sexclass;
run;

proc reg data=for_reg PLOTS=NONE;
  model weight = sexclass height height2 height_sex height2_sex
                 / Aic bic sbc;
  title REG Fit of Weight as a Function of Height and Sex;
run; quit;

proc reg data=for_reg PLOTS=NONE;
  model weight =  height sexclass height_sex height2  height2_sex
                 / selection =adjrsq aic bic sbc;
  title REG SELECTION=aic bic sbc;
run; quit;
```

Output:

REG Fit of Weight as a Function of Height and Sex

The REG Procedure
Model: MODEL1

Dependent Variable: Weight

Number of Observations Read	190
Number of Observations Used	190

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	89423	17885	89.69	<.0001
Error	184	36690	199.40373		
Corrected Total	189	126113			

Root MSE	14.12104	R-Square	0.7091
Dependent Mean	99.00719	Adj R-Sq	0.7012
Coeff Var	14.26264		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	322.83911	148.55463	2.17	0.0310
sexclass	1	-584.25236	183.13473	-3.19	0.0017
Height	1	-9.35722	4.57671	-2.04	0.0423
height2	1	0.09328	0.03505	2.66	0.0085
height_sex	1	18.43766	5.90723	3.12	0.0021
height2_sex	1	-0.14791	0.04767	-3.10	0.0022

REG SELECTION=aic bic sbc

The REG Procedure
Model: MODEL1

Dependent Variable: Weight

Adjusted R-Square Selection Method

Number of Observations Read	190
Number of Observations Used	190

Note:

Number in Model	Adjusted R-Square	R-Square	AIC	BIC	SBC	Variables in Mo
5	0.7012	0.7091	1012.0162	1014.4054	1031.49836	Height sexclass height_sex height2 height2_
4	0.6960	0.7025	1014.2843	1016.3808	1030.51944	sexclass height_sex height2 height2_sex
4	0.6913	0.6979	1017.1911	1019.1345	1033.42622	Height sexclass height_sex height2_sex
2	0.6901	0.6933	1016.0180	1017.8951	1025.75909	Height sexclass
2	0.6896	0.6929	1016.3080	1018.1760	1026.04908	Height height_sex
2	0.6890	0.6923	1016.6907	1018.5467	1026.43178	sexclass height2
2	0.6888	0.6921	1016.7829	1018.6360	1026.52397	Height height2_sex
3	0.6887	0.6937	1017.8208	1019.6655	1030.80893	sexclass height_sex height2
3	0.6887	0.6937	1017.8270	1019.6714	1030.81508	Height sexclass height2
3	0.6885	0.6934	1017.9665	1019.8051	1030.95458	Height sexclass height2_sex
3	0.6884	0.6934	1018.0124	1019.8491	1031.00049	Height sexclass height_sex
3	0.6882	0.6931	1018.1608	1019.9913	1031.14889	sexclass height2 height2_sex
3	0.6880	0.6929	1018.2670	1020.0931	1031.25509	Height height_sex height2_sex
3	0.6879	0.6929	1018.3080	1020.1323	1031.29607	Height height_sex height2
3	0.6873	0.6923	1018.6865	1020.4951	1031.67464	Height height2 height2_sex
4	0.6872	0.6938	1019.7046	1021.5166	1035.93969	Height sexclass height_sex height2
4	0.6870	0.6937	1019.8185	1021.6247	1036.05366	Height sexclass height2 height2_sex
3	0.6867	0.6916	1019.0729	1020.8654	1032.06098	height_sex height2 height2_sex
4	0.6863	0.6930	1020.2457	1022.0296	1036.48077	Height height_sex height2 height2_sex
2	0.6863	0.6896	1018.3000	1020.1057	1028.04104	height_sex height2
2	0.6840	0.6873	1019.7226	1021.4839	1029.46363	height2 height2_sex
1	0.6335	0.6354	1046.8794	1048.1137	1053.37340	Height
1	0.6334	0.6353	1046.9452	1048.1782	1053.43927	height2
2	0.6325	0.6364	1048.4035	1049.2998	1058.14458	Height height2
3	0.4694	0.4779	1119.1400	1117.3822	1132.12810	sexclass height_sex height2_sex
2	0.4677	0.4733	1118.7776	1117.8310	1128.51865	sexclass height_sex
2	0.4628	0.4685	1120.5109	1119.5243	1130.25198	sexclass height2_sex
2	0.4529	0.4587	1123.9953	1122.9289	1133.73632	height_sex height2_sex
1	0.2640	0.2679	1179.3636	1178.3095	1185.85768	sexclass
1	0.1980	0.2022	1195.6784	1194.4073	1202.17241	height_sex
1	0.1434	0.1480	1208.1768	1206.7489	1214.67082	height2_sex

GLMSELECT code is shown here with results following:

```
proc glmselect data=class namelen=38;
  class sex/order=freq
          param=ref ref=first missing;
  model weight=height | sex
              height*height
              height*height*sex
              / SELECTION=none;
  title GLMSELECT SELECTION=NONE;
run;quit;

proc glmselect data=class namelen=38;
  class sex/order=freq
          param=ref ref=first missing;
  model weight=height | sex
              height*height
              height*height*sex
              / SELECTION=backward choose=sbc;
  title GLMSELECT SELECTION=sbc;
run;quit;
```

GLMSELECT SELECTION=NONE

The GLMSELECT Procedure

Least Squares Summary

Step	Effect Entered	Number Effects In	SBC
0	Intercept	1	1239.8504
1	Height	2	1053.3734
2	Sex	3	1025.7591*
3	Height*Sex	4	1031.0005
4	Height*Height	5	1035.9397
5	Height*Height*Sex	6	1031.4984

* Optimal Value of Criterion

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	89423	17885	89.69	<.0001
Error	184	36690	199.40373		
Corrected Total	189	126113			

Root MSE	14.12104
Dependent Mean	99.00719
R-Square	0.7091
Adj R-Sq	0.7012
AIC	1204.01621
AICC	1204.6316
SBC	1031.49836

Parameter Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	322.839105	148.554632	2.17	0.0310
Height		1	-9.357215	4.576712	-2.04	0.0423
Sex	F	1	-584.252358	183.134733	-3.19	0.0017
Height*Sex	F	1	18.437660	5.907226	3.12	0.0021
Height*Height		1	0.093279	0.035054	2.66	0.0085
Height*Height*Sex	F	1	-0.147905	0.047672	-3.10	0.0022

GLMSELECT SELECTION=sbc

The GLMSELECT Procedure

Backward Selection Summary

Step	Effect Removed	Number Effects In	SBC
0		6	1031.4984
1	Height	5	1030.5194*

*** Optimal Value of Criterion**

Selection stopped at a local minimum of the SBC criterion.

Stop Details

Candidate For Removal	Effect	Candidate SBC		Compare SBC
	Height*Height*Sex	1030.8089	>	1030.5194

Note: The selected model, based on SBC, is the model at Step 1.

Effects:	Intercept Sex Height*Sex Height*Height Height*Height*Sex

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	4	88590	22147	109.19
Error	185	37524	202.83142	
Corrected Total	189	126113		

Root MSE	14.24189
Dependent Mean	99.00719
R-Square	0.7025
Adj R-Sq	0.696
AIC	1206.28432
AICC	1206.74333
SBC	1030.51944

Parameter Estimates					
Parameter		DF	Estimate	Standard Error	t Value
Intercept		1	19.540111	7.924019	2.47
Sex	F	1	-280.953364	108.304224	-2.59
Height*Sex	F	1	9.080444	3.766802	2.41
Height*Height		1	0.021708	0.001837	11.82
Height*Height*Sex	F	1	-0.076334	0.032638	-2.34

Note that the GLMSELECT model dropped the HEIGHT term. Also note no p-values but an ABS(t)>=2 is a p of 0.05 or lower. I generally don't like to remove a main effect if the interaction term is significant and would test with a joint test of the main effect and interaction.

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki (Eds.), Second international symposium on information theory, 267-281. Budapest: Academiai Kiado.
- Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716-723.
- Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. Biometrika, 297-307.
- Peter L. Flom and David L. Cassell (2007). Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use. NESUG 2007, https://www.lexjansen.com/pnwsug/2008/DavidCassell-StoppingStepwise.pdf
- Sawa, T. (1978). Information criteria for discriminating among alternative regression models. Econometrica, 46, 1273-1282. Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

Disclaimers: