Model Selection

Using trial and error (Forward, Backward, Stepwise, Variable selection the best selection criteria were identified using stepwise:

**PROC** **logistic** data= boxing;

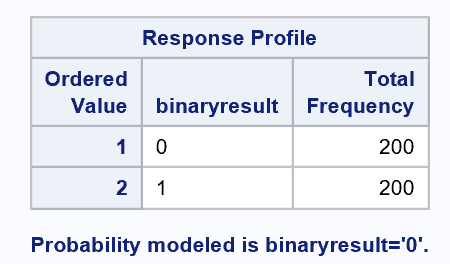
class Stance Over35AgeA Over35AgeB Over15lbA Over15lbB;

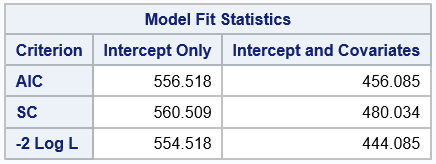
model binaryresult = age\_A age\_B height\_A height\_B reach\_A reach\_B weight\_A weight\_B won\_A won\_B lost\_A lost\_B kos\_A kos\_B AdvAgeA AdvHeightA AdvReachA AdvWgtA WinPA WinPB KoAPer KoBPer

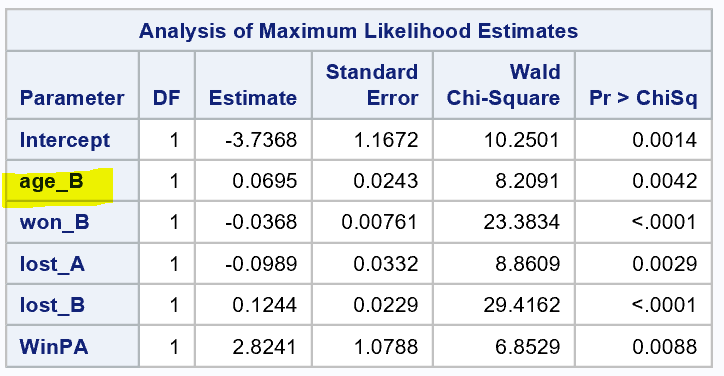
/ selection = stepwise;

output out=boxinglogregout predprobs=I p=probpreb;

**run**;







Goodness of Fit

Though the model that stepwise selection chose with lowest AIC and BIC fits, it doesn’t quite make sense that the stepwise selection includes age\_B but excludes age\_A. These variables have equal meaning and weight. If we simply swap a and b from left to right the result would be different for no logical reason.

This could be indicating that there is indeed a difference in the meaning of the fighter\_A and fighter\_B slots (i.e. challenger versus incumbent). However, we have no context or documentation to confirm this. It is more likely that there is interaction between the variables.

Trial and error shows that adding age\_A does not help. Removing age\_B gives us statistically significant coefficients but reduces fit.

Creating a new variable “AdvAgeA” as Age\_A – Age\_B accounts for both variables an allows us to model without an interaction term. Interestingly it actually has slightly better fit statistics than the interaction term. The fit statistics aren’t quite as good as what stepwise selection gave us but the model makes more sense.

|  |  |
| --- | --- |
| **Include age\_A and age\_B:** | **Exclude age\_A and age\_B:** |
|  |  |
|  |  |
| **Interaction age\_A\*age\_B** | **Cheat to represent interaction** AdvAgeA= Age\_A–Age\_B |
|  |  |
|  |  |

The Chosen Model

Common sense tells us that there is almost certainly interaction between age\_A and age\_B; however, **for the purpose of part 1 in this analysis we will not include an interaction term as instructed.**

It is also noted that stepwise selection included Won\_B but not Won\_A which would raise similar concerns as with age\_A and age\_B. This could be more evidence that there is interaction or there may be meaning to the slots A or B or. WinPA (percentage of prior wins for fighter A) was selected and probably accounts for this.

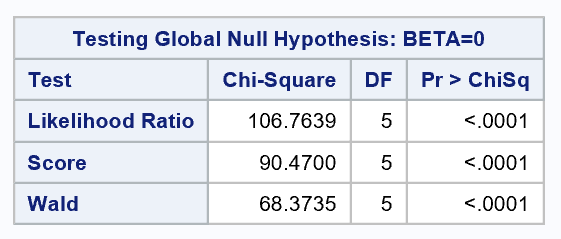
**PROC** **logistic** data= boxing;

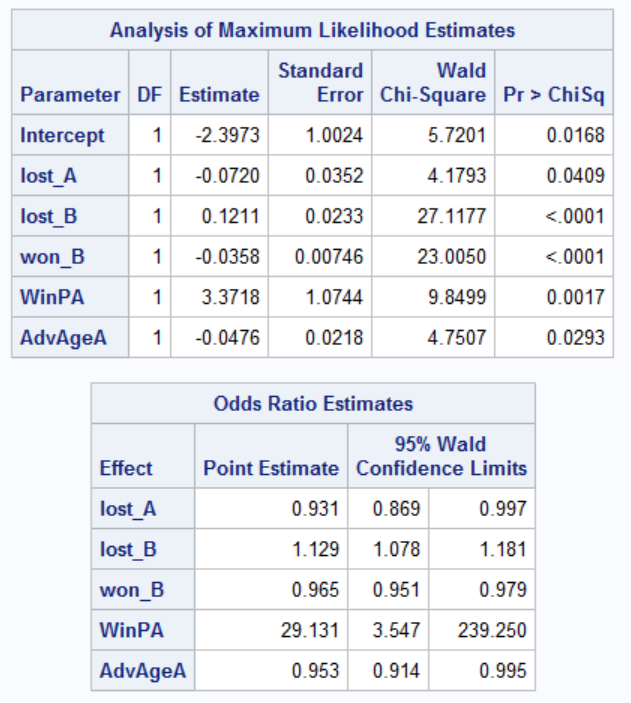
model binaryresult = lost\_A lost\_B won\_B WinPA AdvAgeA /LACKFIT CTABLE;

/\*output out=boxinglogregout predprobs=I p=probpreb;\*/

**run**;

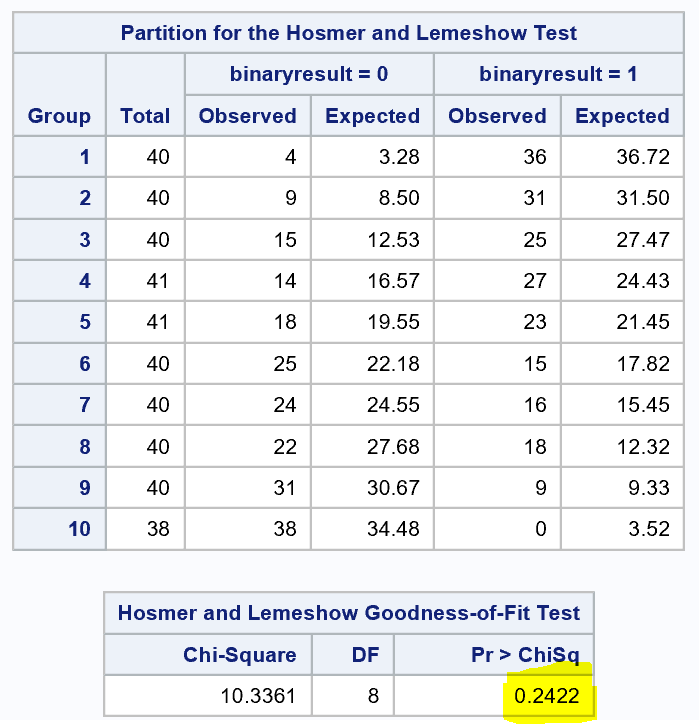
We can reject the null hypotheses that BETA=0. Our variables are statistically significant in predicting 0,1 Fighter A Wins versus Fighter B Wins





|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | Fighter A Wins (0) | Fighter B Wins (1) |
| Predicted | Fighter A Wins (0) | 144 | 56 |
| Fighter B Wins (1) | 63 | 137 |

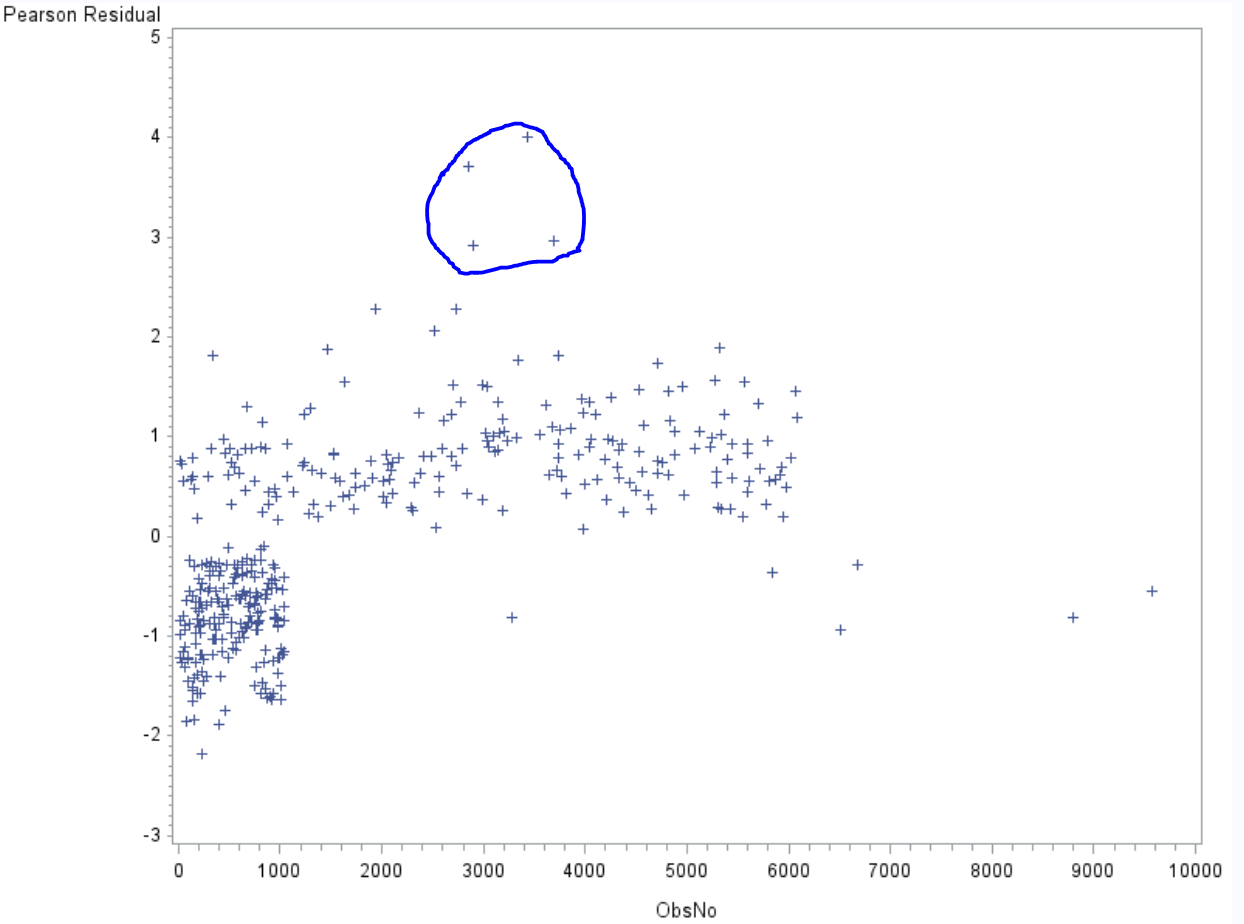
Using .5 accuracy because the negative impact of falsely predicting is equal. These are boxing matches, If this were cancer and not cancer and if treatment will have negative side effects then we would want to consider higher than .5

Use Hosmer Lemeshow because many continues variables.

Do not reject the null hypothesis p-value .2422; the model is a good fit.

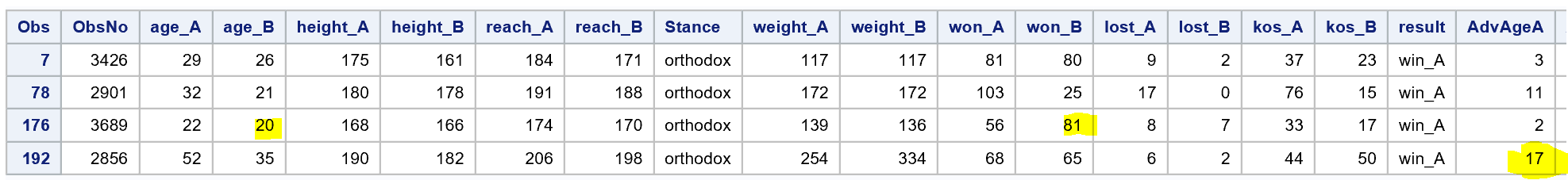
Residual Diagnostics

Something strange about these observations

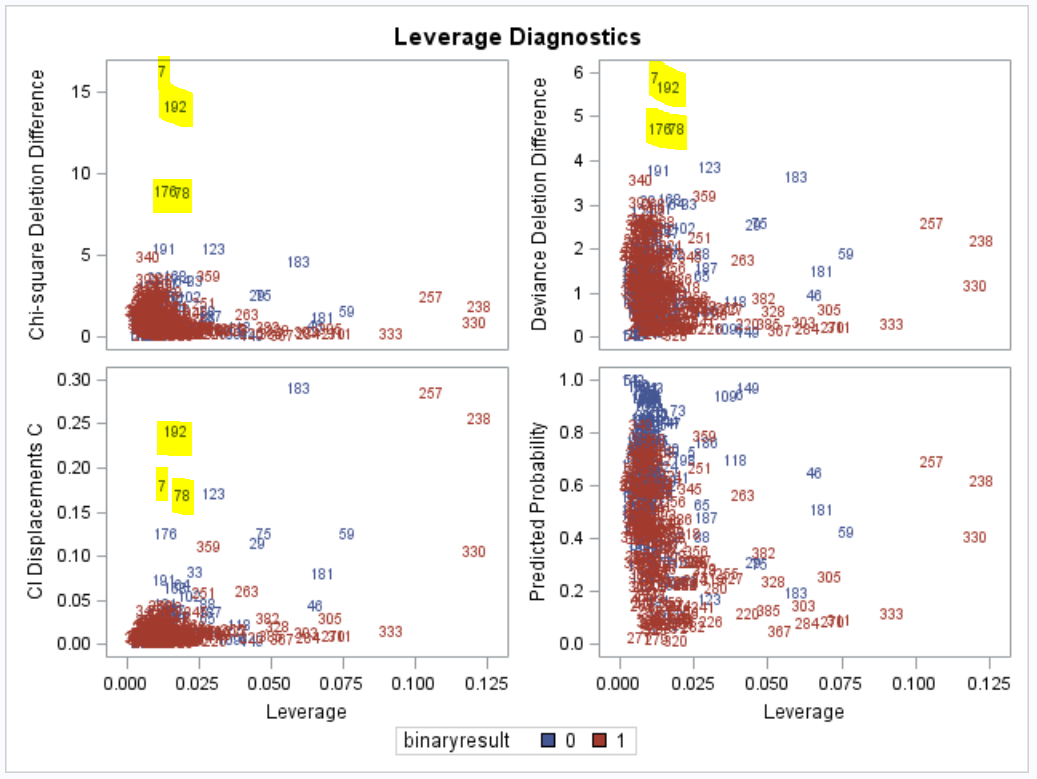


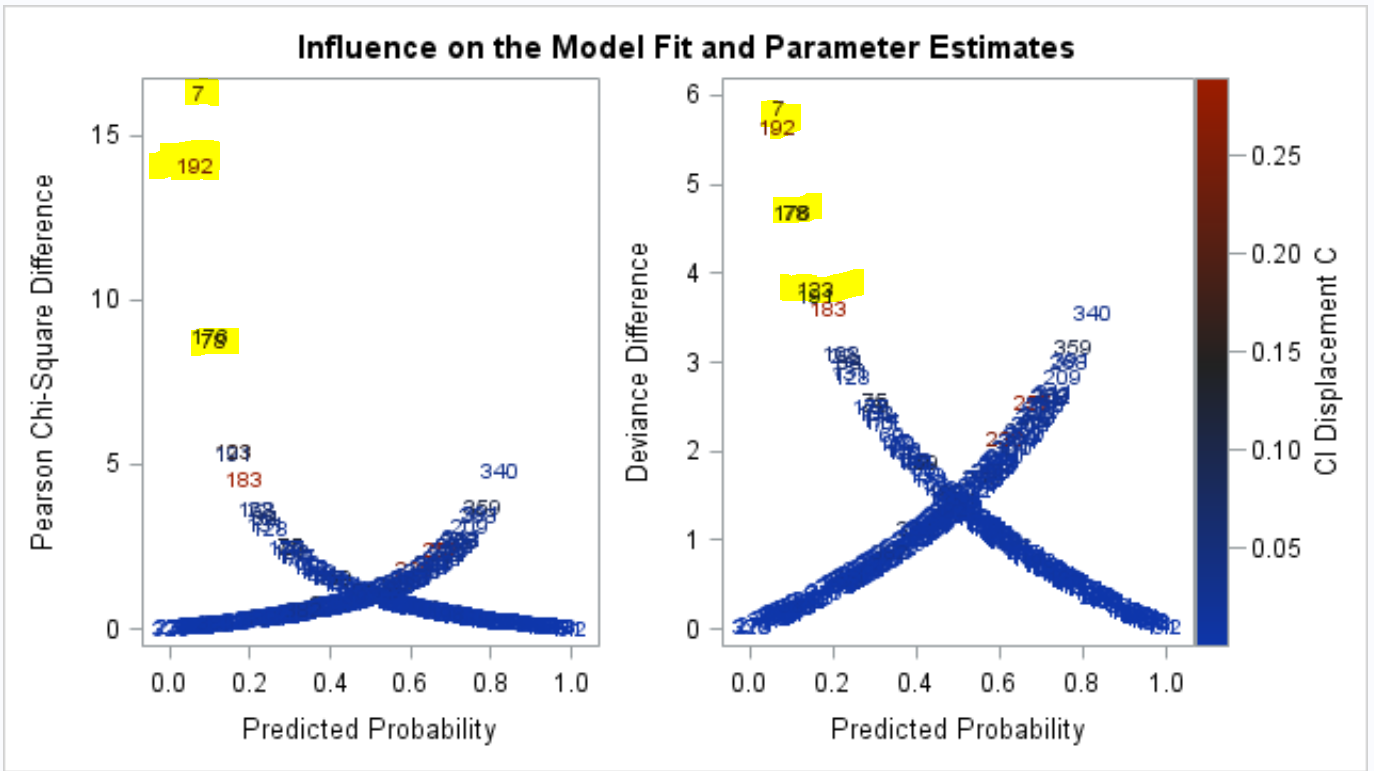
There are suspicious outliers;

* Observation No. 2856 age\_A=52 vs age\_B=35 seems extreme.
* Observation No. 3689 age\_B=20 and has had 88 fights total seems extreme given age.



Observation No. 7, 78, 176, 192 are standing out:

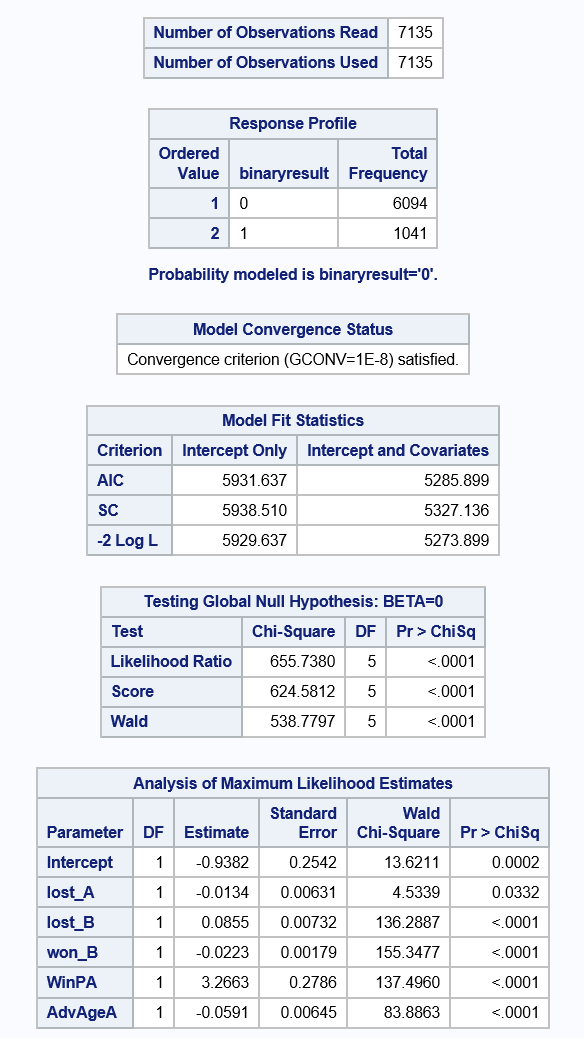


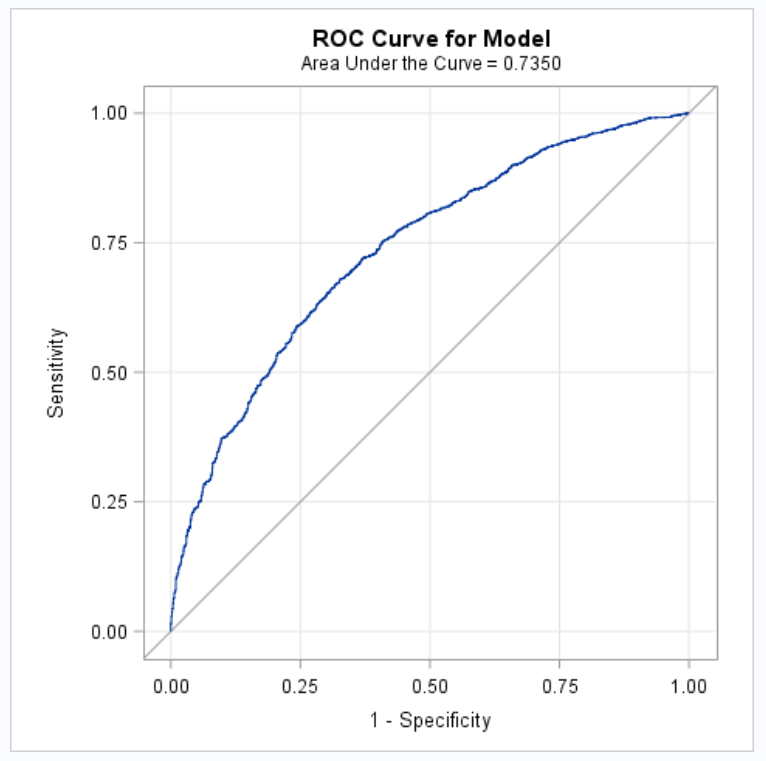


Try model without them and see if stats improve:

|  |  |
| --- | --- |
| With Outliers | Without Outliers |
|  |  |

Test Results





Final Model

The probability that Fighter A is the winner () is defined as:

SAS Code

/\* Assumptions \*/

/\* Multivariate normal distribution for entire set of variables \*/

/\* Univariate normal distribution on response \*/

/\* Linear relationships between scores on Y and scores on X for all variables \*/

/\* Uniform error variances for response (Y) across all values of X \*/

%***web\_drop\_table***(WORK.BOXING);

FILENAME REFFILE "C:/Users/danie/Documents/GitHub/6372BoxingProject/train.csv";

**PROC** **IMPORT** DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.BOXING;

GETNAMES=YES;

**RUN**;

**PROC** **PRINT** data=boxing;

/\* Model Selection \*/

**PROC** **logistic** data= boxing;

class Stance Over35AgeA Over35AgeB Over15lbA Over15lbB;

model binaryresult = age\_A age\_B height\_A height\_B reach\_A reach\_B weight\_A weight\_B won\_A won\_B lost\_A lost\_B kos\_A kos\_B AdvAgeA AdvHeightA AdvReachA AdvWgtA WinPA WinPB KoAPer KoBPer

/ selection = stepwise;

output out=boxinglogregout predprobs=I p=probpreb;

**run**;

/\* Chosen Model\*/

**PROC** **logistic** data= boxing plots(only label)=(leverage dpc);

model binaryresult = lost\_A lost\_B won\_B WinPA AdvAgeA /LACKFIT CTABLE;

output out=boxinglogregout predprobs=I p=probpreb resdev=resdev reschi=pearres;

**run**;

**proc** **print** data=boxing;

where obsno = **3426**;

**run**;

/\* Candidate 1 remove age\_B \*/

/\* PROC logistic data= boxing; \*/

/\* model binaryresult = lost\_A lost\_B won\_B WinPA /LACKFIT CTABLE; \*/

/\* output out=boxinglogregout predprobs=I p=probpreb resdev=resdev reschi=pearres; \*/

/\* run; \*/

/\* \*/

/\* Candidate 1.5 add age A \*/

/\* PROC logistic data= boxing; \*/

/\* model binaryresult = age\_A age\_B lost\_A lost\_B won\_B WinPA /LACKFIT CTABLE; \*/

/\* output out=boxinglogregout predprobs=I p=probpreb resdev=resdev reschi=pearres; \*/

/\* run; \*/

/\* \*/

/\* Candidate 2 age A and B interaction \*/

/\* PROC logistic data= boxing; \*/

/\* model binaryresult = lost\_A lost\_B won\_B WinPA age\_A\*age\_B /LACKFIT CTABLE; \*/

/\* output out=boxinglogregout predprobs=I p=probpreb resdev=resdev reschi=pearres; \*/

/\* run; \*/

**proc** **gplot** data=boxinglogregout;

plot resdev\*obsno;

plot pearres\*obsno;

**run**; **quit**;

**proc** **print** data=boxinglogregout;

where pearres > **2.5**;

**run**;

/\* Remove outliers and run again \*/

**DATA** boxingRemovedOutliers;

SET boxing;

IF obsno = **3426** THEN DELETE;

IF obsno = **2901** THEN DELETE;

IF obsno = **3689** THEN DELETE;

IF obsno = **2856** THEN DELETE;

**RUN**;

**PROC** **logistic** data= boxingRemovedOutliers plots(only label)=(leverage dpc);

model binaryresult = lost\_A lost\_B won\_B WinPA AdvAgeA /LACKFIT CTABLE;

output out=boxinglogregoutRemovedOutliers predprobs=I p=probpreb resdev=resdev reschi=pearres;

**run**;

/\* Test \*/

%***web\_drop\_table***(WORK.BOXINGTEST);

FILENAME REFFILE "C:/Users/danie/Documents/GitHub/6372BoxingProject/test.csv";

**PROC** **IMPORT** DATAFILE=REFFILE

DBMS=CSV

OUT=WORK.BOXINGTEST;

GETNAMES=YES;

**RUN**;

**proc** **logistic** data=BOXINGTEST rocoptions(crossvalidate) plots(only)=roc;

model binaryresult(event="0") = lost\_A lost\_B won\_B WinPA AdvAgeA;

**run**;