Model Selection

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

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

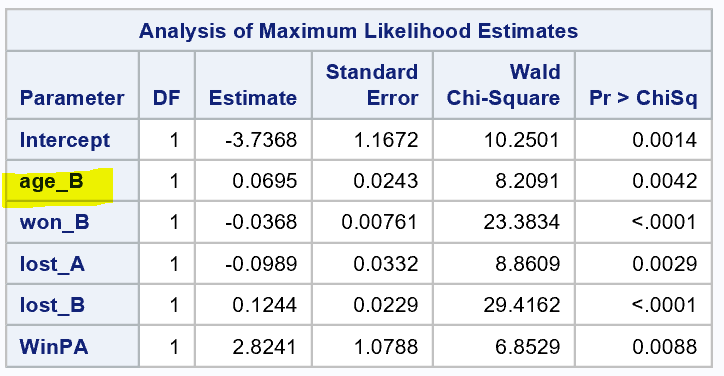
class Stance Over35AgeA Over35AgeB Over15lbA Over15lbB;

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;

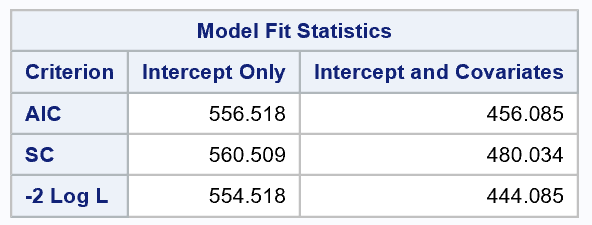
output out=boxinglogregout predprobs=I p=probpreb;

**run**;



Goodness of Fit

Including Age\_B

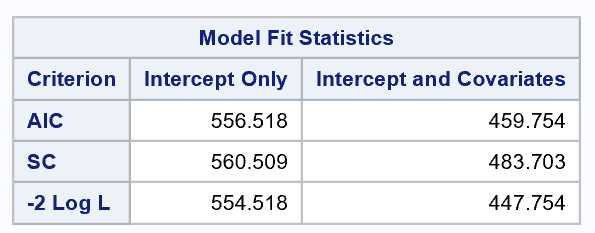


Though the model chosen with lowest AIC and BIC fits, it doesn’t quite make sense that the stepwise selection includes age\_B but excludes age\_A. These variables have equal meaning and weight. If we simply swap a and b from left to right the result would be different for no logical reason. This could be indicating that there is indeed a difference in the meaning of fighter\_A and fighter\_B (i.e chcallenger versus incumbent) however we have no context to the data that confirms this. For now we will assume there is no meaning.

Trial and error showed that adding Age\_A or removeing Age\_B does not help.

Creating a new variable “AdvAgeA” as Age\_A – Age\_B accounts for both of them and seems to have less a negative effect on the fit of the model than simply including Age\_A and Age\_B separately. The fit statistics aren’t quite as good but the model makes more sense.

Replace Age\_B with AdvAge\_A



Also noted that stepwise selection included Won\_B but not Won\_A which would raise similar concern. More evidence that there may be meaning to the slots A or B.

However WinPA (percentage of prior wins for fighter A) is included

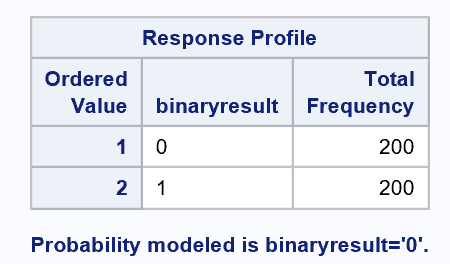
|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Actual | |
|  |  | Fighter A Wins (0) | Fighter B Wins (1) |
| Predicted | Fighter A Wins (0) | 144 | 56 |
| Fighter B Wins (1) | 63 | 137 |

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

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

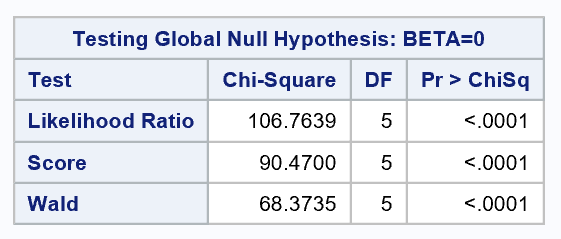
/\*output out=boxinglogregout predprobs=I p=probpreb;\*/

**run**;



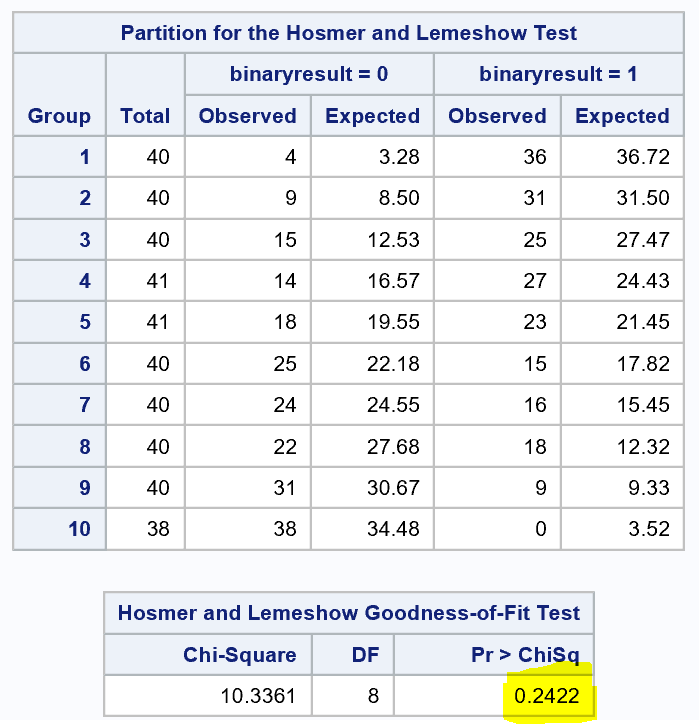
0 meaning that fighter A wins

We can reject the null hypotheses that BETA=0. Our variable are statistically significant in prediction 0,1 FughterAWin FighterBwin



|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0.500** | 144 | 137 | 63 | 56 | 70.3 | 72.0 | 68.5 | 30.4 | 29.0 |

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 that the model fits the data p-value .2422. The model is a good fit.