**Q&A for Project 2**

1. Is there an option in SAS to utilize the log / loss function for evaluation, or is that something we need to set up in Excel?

A1: You can do this in SAS or Excel.  To do this you need to use the output from the proc logistic an proc discrim analysis.  You are asked to provide a separate dataset with predictions, so it makes sense to do it Excel.  I would do it both in SAS and Excel as a check.

2. I took a log transform of the shot\_distance variable. Is this OK for evaluating odds ratios?

A2: Taking logs is fine when evaluating odds ratios.  Just be clear when you report them about whether there is a variable transformation.

3. When I run the stepwise model (same for forward and backward selection as well), the shot Zone Range variable always comes out with a positive odds ratio which doesn’t make a lot of sense to me.

A3: I suspect what you are observing with the stepwise proc logistic is that the shot range variable is correlated with some other variable.  That is why you are getting the strange results.  Try eliminating some variables correlated with shot range and see if you get more sensible odds ratios.

4. I think because of the large data set SAS does not show the output for proc logistic.

A4a: I think there are a few things going on with your coding and analysis. To get your code to start working in a reasonable amount of time, for analysis of coding problems, run it on a small subset of the data.  See my code below and the (obs=1000) option in the set statement.  Also watch your print statements.  Sometimes printing too much can slow your program execution.  You also can add an (obs=) to a proc print statement to print a subset of the data.

data full; set datIn(obs=1000); run;

proc logistic data = full descending;

class shot\_made\_flag/ param = ref;

model shot\_made\_flag = lat lon minutes\_remaining seconds\_remaining shot\_distance

https://ssl.gstatic.com/ui/v1/icons/mail/images/cleardot.gif attendance arena\_temp avgnoisedb/ctable pprob = .5;

output out = logisticOut predprobs=I p=predprob resdev=resdev reschi=pearres;

run;

A4b:  In your model you have just included all the variables.  All the class variables are bogging you down.  I eliminated the class variables and got the program to run.  Also, there appears to be a conflict with the loc\_x loc\_y and lat lon variables.  When I eliminated the loc\_x loc\_y variables the program gave sensible results.  These sets of variables are very similar so eliminating one set makes sense.

A4c:  If you want to include some of the class variables, I would first see if they can be re-coded into some kind of ordinal ranking variable that SAS would see as continuous.  If you have other class variables you think important in the modeling, enter them one at a time and see how the model responds.

A4d:  Think carefully about the variables and what makes sense to include.  Basketball is a game where the players are trying to put a ball in a hoop.  The closer they are to the basket, the easier it is to get the ball in the hoop.  The game lasts 48 minutes with 4, 12 minute quarters.  Weather conditions may affect performance.  Also, players’ performance may be better at different points in a game.  Playoff games are more important than regular season games because in playoff games the players are vying for a championship.

5. Since we are not clear about continuous variables (which to include in the model and which not to include) does it make sense to use different models with various techniques?

A5: Yes, it makes sense to evaluate different models. You may want to think about dividing the data into test and train data sets and trying different models. Use model fit statistics to compare different models.

6. In the project word file it says: "Evaluation: Compare each competing models with the AUC,..." Does this really mean AIC criterion? Is it a mistype?

A6: No. AUC is an acronym for Area Under the Curve. We will talk about this in our next session. This only pertains to models using logistic regression.

7. Where do priors come from in discriminate analysis?

A7: In the graduate school admissions data used in the class examples the priors come from the input data. See the file classExampesSession12.xlsx in the Session12 zip folder.

8. Do we have to answer each of these three propositions for each of the models.?

Part A. The odds of Kobe making a shot decrease with respect to the distance he is from the hoop.  If there is evidence of this, quantify this relationship.  (CIs, plots, etc.) –what other things other than CI, plots, coefficients and interpretations?

A. 8PartA: Based on our class discussion, shot\_distance should enter the logistic model. The coefficient for this variable in the model relates to the change in odds if the distance increases by one unit.

Part B. The probability of Kobe making a shot decreases linearly with respect to the distance he is from the hoop.    Is there is evidence of this relationship (CIs, plots, etc.)?

A. 8PartB: Looking at the coefficient in the logistic model gives an indication of whether the probability making a shot decreases linearly with respect to the distance he is from the hoop.

Part C.The relationship between the distance Kobe is from the basket and the odds of him making the shot is different if they are in the playoffs.  Quantify your findings with statistical evidence one way or the other. (Tests, CIs, plots, etc.)

A. 8PartC: You can calculate the odds Kobe makes a basket in the regular season and the playoffs and then form an odds ratio.  Then you can use a statistical test to test if that ratio is different from one (see session 10, slide 21).  You could also plot the probability of making a shot by distance for the regular season and the playoffs and then compare the slopes of the lines using the estimated slopes with regression and their standard errors. In this case, you will want to group the shots in some categories by distance. The test statistic would be a t test with the standard error given by a weighted average of the standard errors in the two regressions with the sample sizes as weights.

9. Page#3, What is “objective / loss function”?





