**Golf Model Selection Exercise 7.3**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Subset Size | Predictors | R2adj | AIC | AICC | BIC |
| 1 | GIR | 0.251 | -62.516 | -62.391 | -55.959 |
| 2 | GIR, PuttsPerRound | 0.486 | -135.220 | -135.010 | -125.385 |
| 3 | GIR,BirdieConversion,Scrambling | 0.538 | -155.310 | -154.994 | **-142.198** |
| 4 | GIR,BirdieConversion,SandSaves,Scrambling | 0.543 | -156.291 | -155.846 | -139.900 |
| 5 | GIR,BirdieConversion,SandSaves  ,Scrambling,PuttsPerRound | **0.546** | **-156.641** | **-156.045** | -136.972 |
| 6 | DrivingAccuracy,GIR,BirdieConversion,  SandSaves,Scrambling,PuttsPerRound | 0.544 | -154.730 | -153.960 | -131.783 |
| 7 | DrivingAccuracy,GIR,PuttingAverage,  BirdieConversion,SandSaves,Scrambling,  PuttsPerRound | 0.541 | -152.735 | -151.767 | -126.510 |

Table 1 Best values for R2adj, AIC, AICc and BIC on optimal subsets of different size.

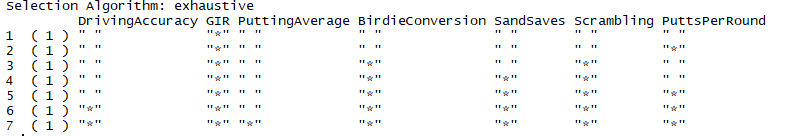


Figure 1 Predictors chosen for optimal model of each subset.

The optimal model using all possible subsets for R2adj, AIC, and AICC would be the subset size 5 one using the predictors GIR, BirdieConversion, SandSaves, Scrambling, and PuttsPerRound to predict the log of PrizeMoney.

The optimal model using BIC was the one with subset size 3. That used the predictors: GIR, BirdieConversion, Scrambling.

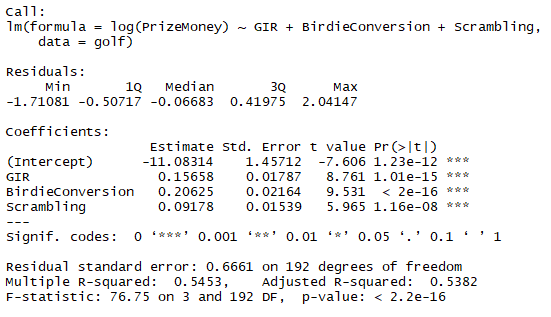


Figure 2 Summary of best BIC model (subset size 3).

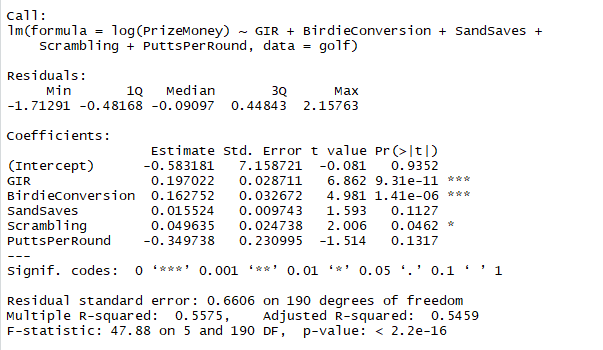


Figure 3 summary of best R2adj, AIC, AICc model (subset size 5).

**Backwards Subset Selection**

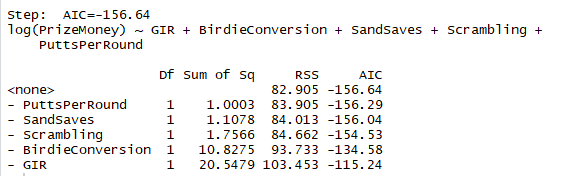
****

Figure 4 Backwards subset using AIC

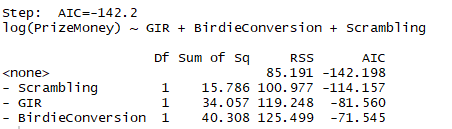


Figure 5 Backwards subset using BIC

**Forwards Subset Selection**

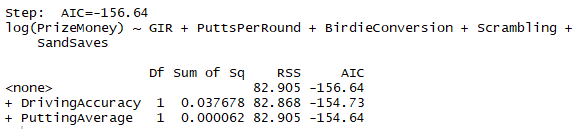
****

Figure 6 Forwards subset using AIC

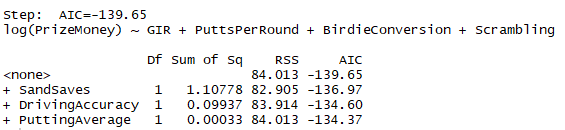


Figure 7 Forwards subset using BIC

These models chosen in the backwards and forwards method of subset selection don’t give you the same models in that minimize the AIC or BIC in Table 1. This is because forwards and backwards subsets method matters about the order in which the algorithm adds each predictor. Putting one in before the other could change the best value you can get for AIC and BIC.

After viewing all the “best” models, I would probably choose the model that uses 5 predictors.

**Downer Cows**

In this exercise we will try to predict whether or not a downer cow will survive given some results of a blood test and the number of days recumbent. First let’s take a look at the variables and how to compare to the response variable: Outcome.

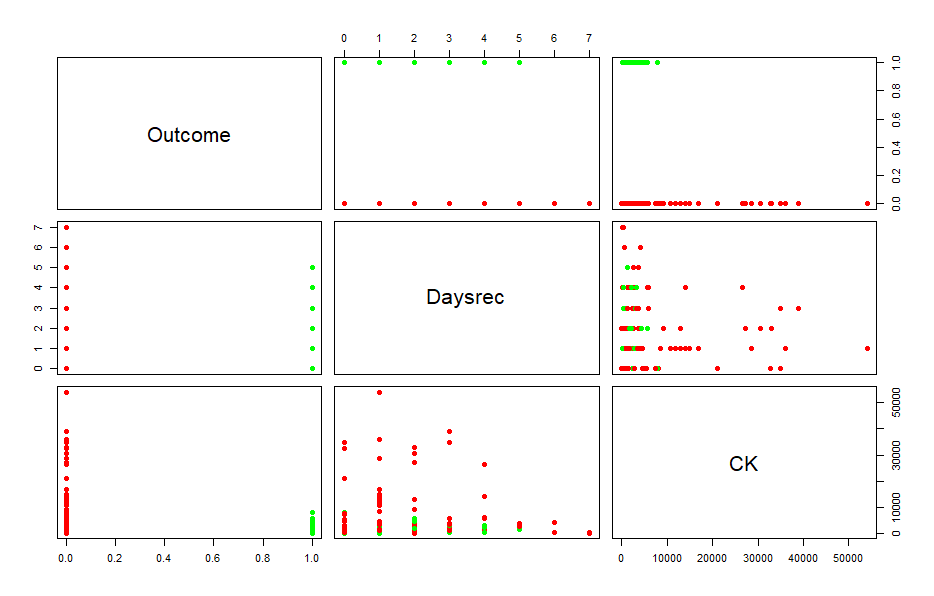


Figure 8 Pairs plot of Daysrec and CK against Outcome

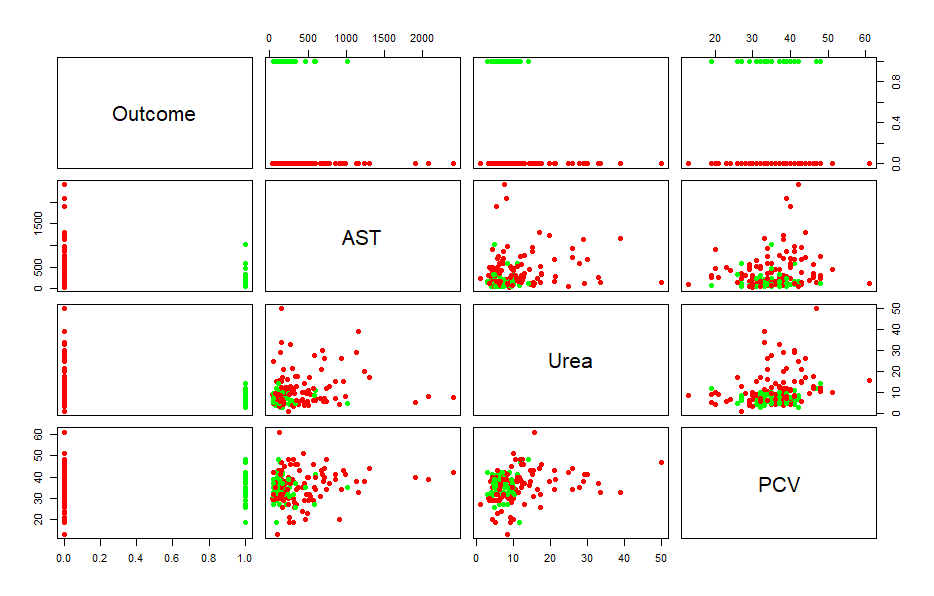


Figure 9 Pairs plot of AST, Urea, and PCV against Outcome

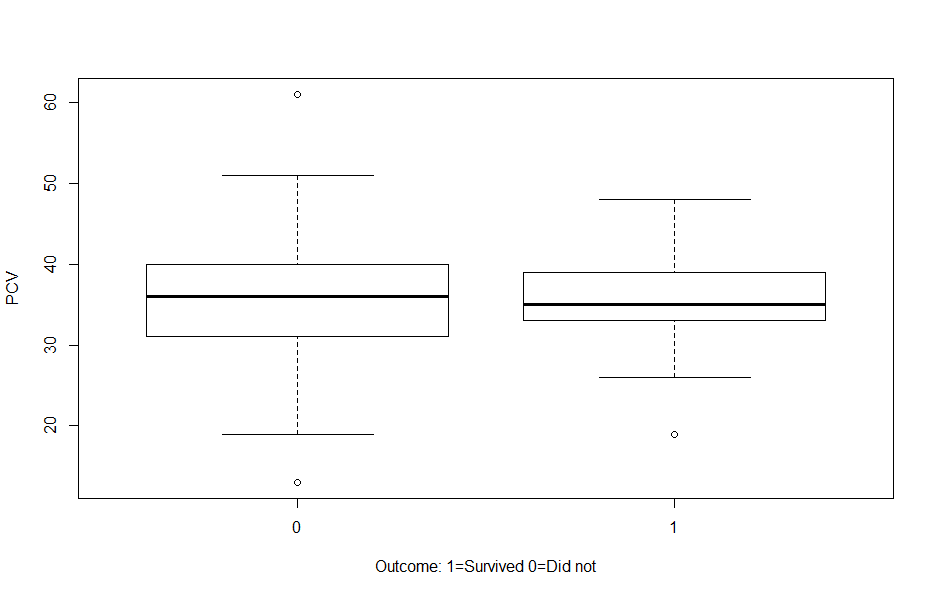


Figure 10 Box plot of PCV ~ Outcome

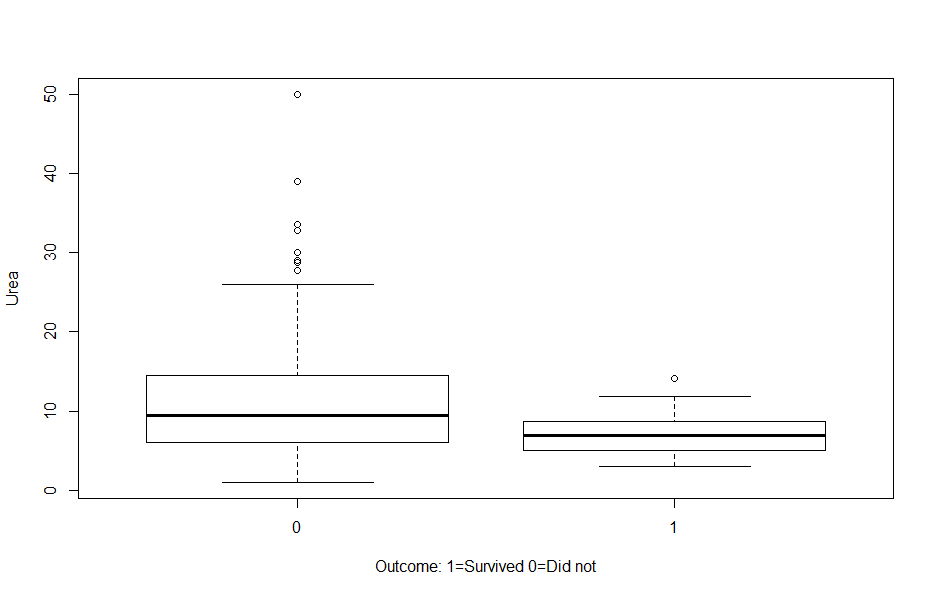


Figure 11 Box plot of Urea ~ Outcome

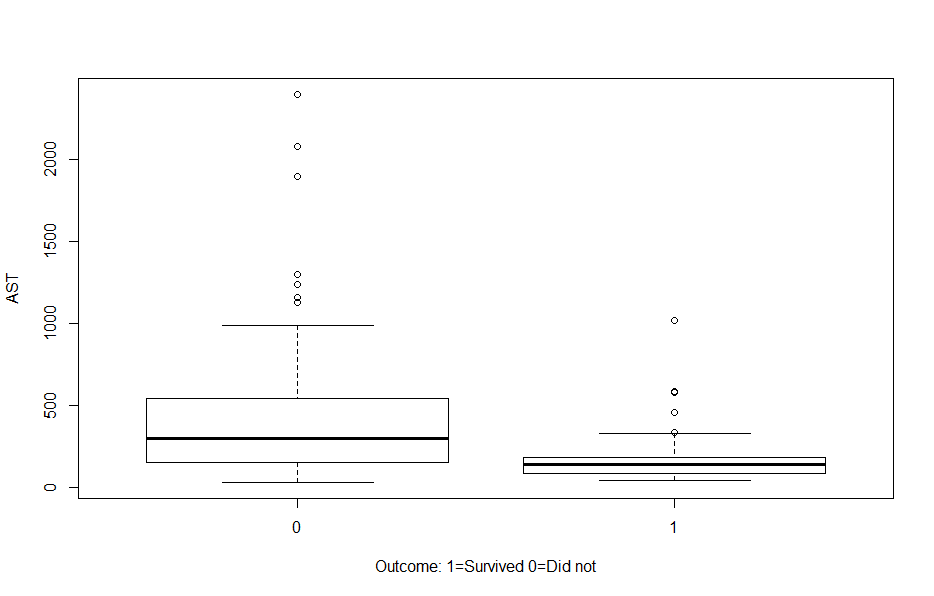


Figure 12 Box plot of AST ~ Outcome

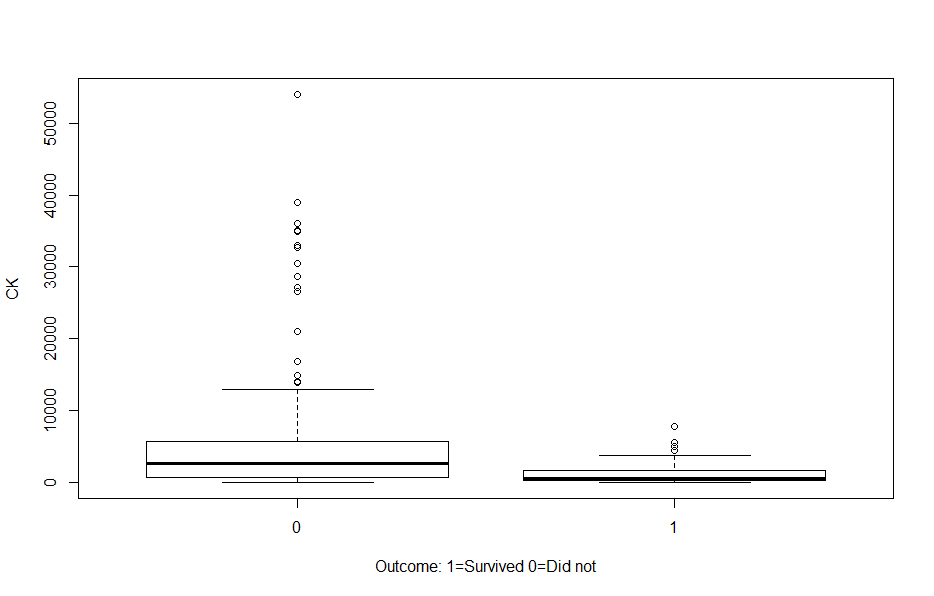


Figure 13 Box plot of CK ~ Outcome

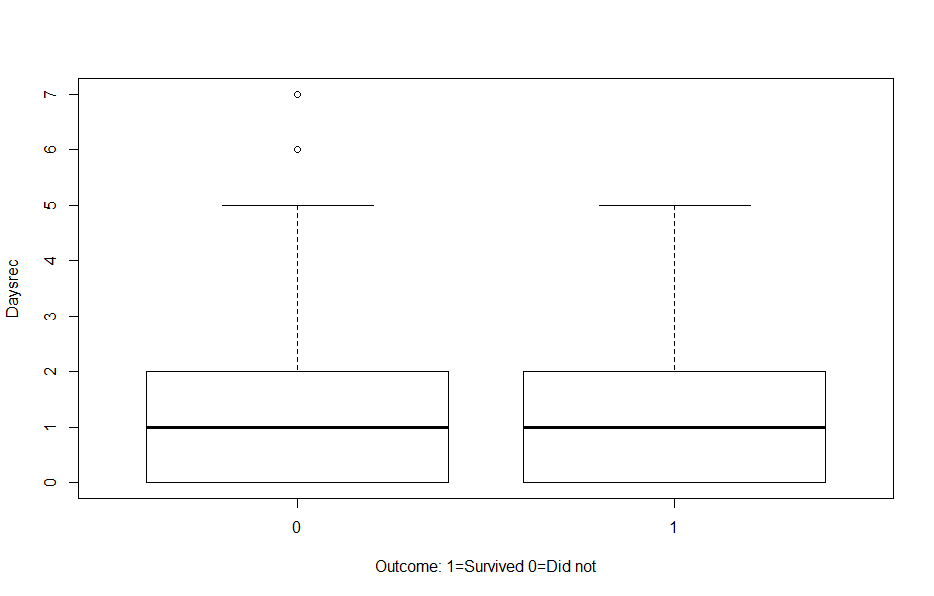


Figure 14 Box plot of Daysrec ~ Outcome

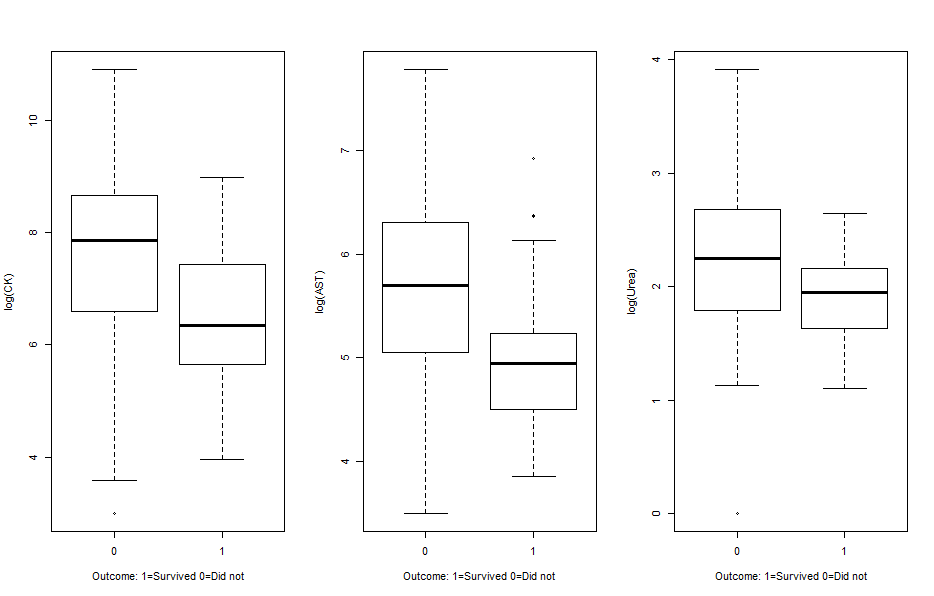


Figure 15 Box plot of CK, AST, and Urea after a log transformation ~ Outcome

Looking at the box plots, you can see that there was some right skewness of AST, CK, and Urea. Taking the log transform of these variables reduced the variance nicely.

**Rich Model**

First fit a “rich” model. This would be fitting Outcome ~ Daysrec + log(CK) + log(AST) + log(Urea) + PCV.

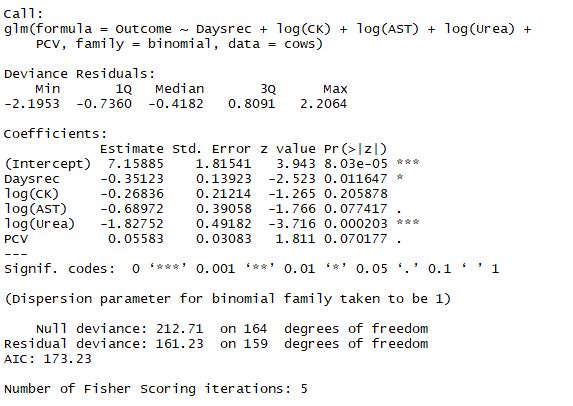


Figure 16 Summary of rich model.

Let’s see if we can reduce our model a little bit using forward and backwards step.

*Backward Subset*

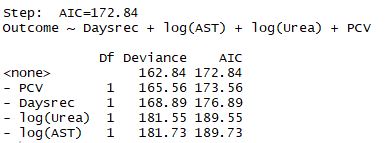


Figure 17 Backward subset of Downer using AIC

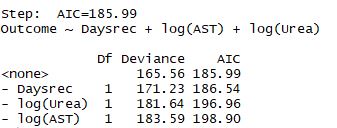


Figure 18 Backward subset of Downer using BIC

*Forward Subset*

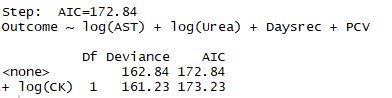


Figure 19 Forward subset using AIC

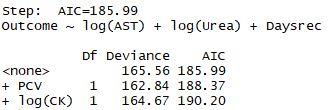


Figure 20 Forward subset using BIC

It’s interesting to see that both AIC criteria methods chose the same model: Outcome ~ log(AST) + log(Urea) + Daysrec + PCV. And both BIC criteria methods chose a smaller model: Outcome ~ log(AST) + log(Urea) + Daysrec. Both criteria contain the same 3 variables: AST, Urea, and Daysrec. The AIC method adds in PCV. We can see the summaries of the two reduced models below.

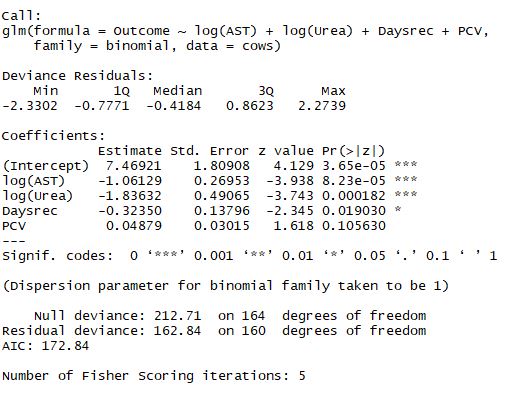


Figure 21 Summary of subset size 4 model

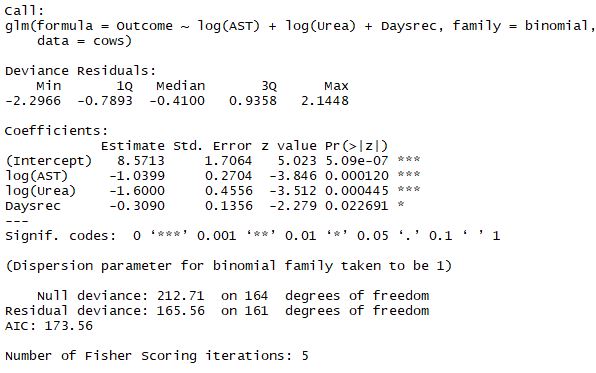


Figure 22 Summary of subset size 3 model

After looking at the model, PCV doesn’t look to be statistically significant, so I’m going to remove it and chose the smaller model as the one to use.

However, looking at the marginal model plots, there seems to be something a little off.

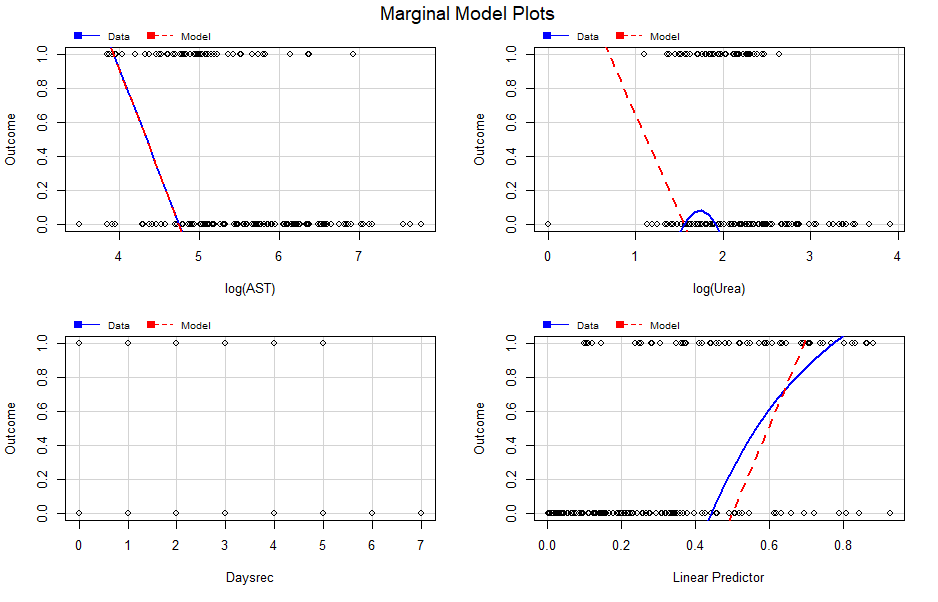


Figure 23 Marginal model plots of chosen model

It’s not fitting log(Urea), Daysrec, or the linear predictor very well. It does the same thing in the larger model too. So I chose to add the non-log transformed columns of AST and Urea into the model and get the following model:

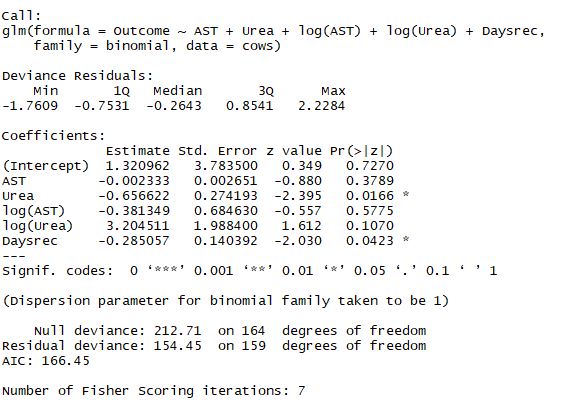


Figure 24 Summary of adding non-log transformed versions of variables into model

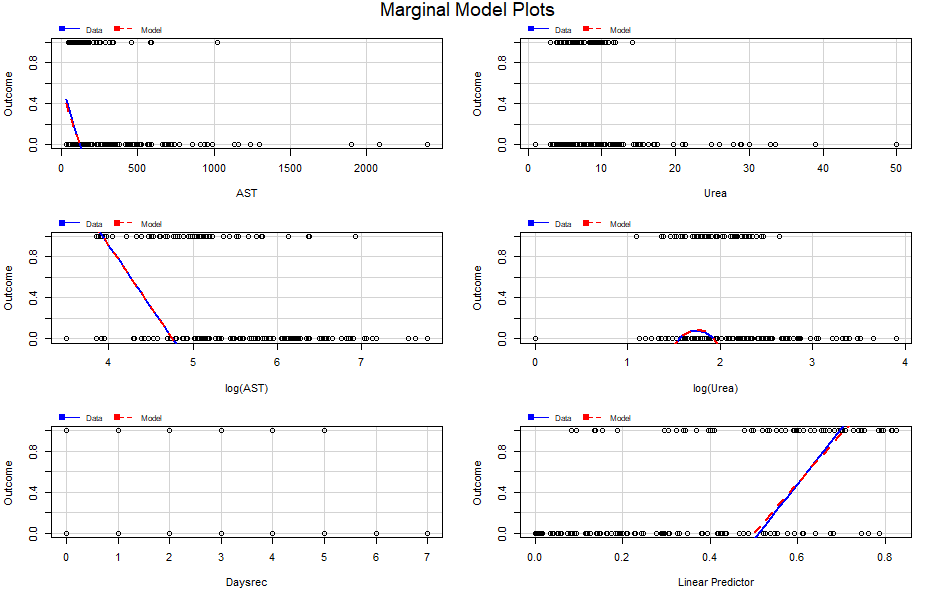


Figure 25 Marginal model plots of larger model

It’s fitting the log transformed variables and the linear predictor much better. However, I’m a little worried about the fitting of the Daysrec, and Urea variables. I can’t get any line to show up no matter what I do. I ran an anova to see if adding the variables was a good idea.

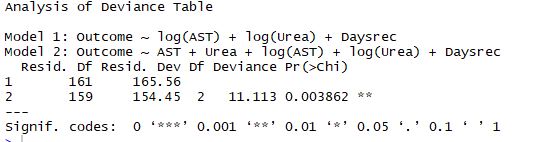


Figure 26 Anova table of size 3 models

Adding in the variables doesn’t seem like a bad idea.