Model 1 Selection Report

Elias Cheleuitte, Micaleb Johnson, Morgan Williams

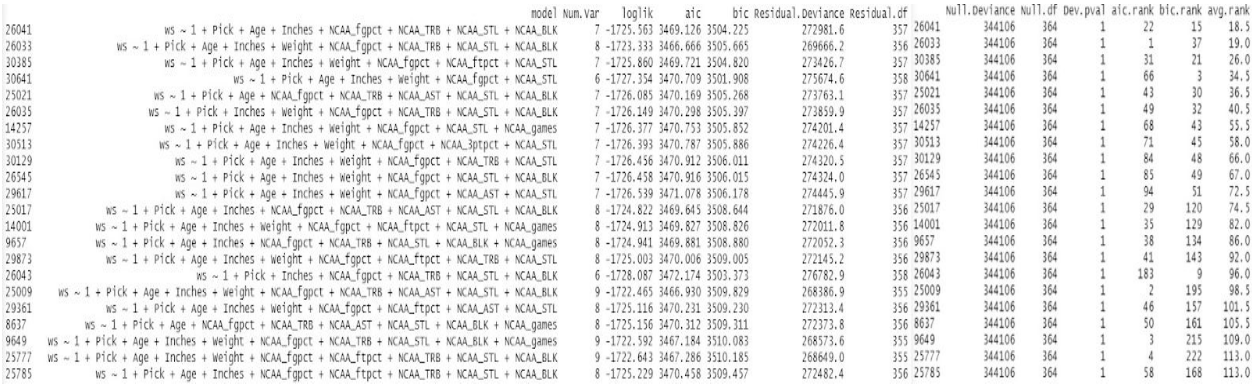
Data Cleaning:

As we first began trying to model players' career win shares, we ran into an issue of a number of missing values. All but one of the highschool players and roughly a third of the international players were missing information for turnovers per game prior to being drafted, because of this we decided to remove the turnover variable from any potential models. After removing turnovers from our model selection process we still were unable to properly model our data due to missing data for 9 players spread between three variables: assists per game, steals per game, and blocks per game. Seven of these were high school players, while the other two were international players. Due to our large original sample size of 468 we concluded that removing the 9 players from our data was best. Furthermore, we removed two more players from our data Qyntel Woods and Kedrick Brown because, while they played in college in the United States, they played at Junior Colleges and not the NCAA like the rest of our college players.

Model Selection:

Once our data issues were corrected, we randomly split our original dataset by rows into train and test sets, with 80% of our data being put into the training set and the remaining 20% into the testing set. Now working only with the training dataset, we automated the selection of the optimal regression model based on the AIC and BIC for all possible model subsets starting from a full model with all 15 covariates down to an intercept only model. As a result, a six variable model consisting of Pick, Age, Height (inches), Weight, pre NBA draft field goal percentage (The number of shots a player makes / The number of shots a player attempts), and pre NBA draft steals per game, looked to be the best model based on our metrics. All of these covariates had p-values less than the significance level of .05 and thus all are likely meaningful to the model. Another model with seven variables including the same covariates as the previous six variable model plus pre NBA free throw percentage also looked promising, with all but one of the variables appearing to be significant by their p-values. We also looked for any interaction effects to include but none appeared to drastically improve the models performance without greatly increasing the complexity of the model thus none were added.

Figure 1: Best models sorted by average AIC and BIC rank



```
                                                                        model Num.Var    loglik      aic      bic Residual.Deviance Residual.df       Null.Deviance Null.df Dev.pval aic.rank bic.rank avg.rank
26041               ws ~ 1 + Pick + Age + Inches + NCAA_fgpct + NCAA_TRB + NCAA_STL + NCAA_BLK       7 -1725.563 3469.126 3504.225          272981.6         357 26041        344106     364        1       22       15     18.5
26033      ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_TRB + NCAA_STL + NCAA_BLK       8 -1723.333 3466.666 3505.665          269666.2         356 26033        344106     364        1        1       37     19.0
30385             ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_ftpct + NCAA_STL       7 -1725.860 3469.721 3504.820          273426.7         357 30385        344106     364        1       31       21     26.0
30641                           ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_STL       6 -1727.354 3470.709 3501.908          275674.6         358 30641        344106     364        1       66        3     34.5
25021            ws ~ 1 + Pick + Age + NCAA_fgpct + NCAA_TRB + NCAA_AST + NCAA_STL + NCAA_BLK       7 -1726.085 3470.169 3505.268          273763.1         357 25021        344106     364        1       43       30     36.5
26035            ws ~ 1 + Pick + Inches + Weight + NCAA_fgpct + NCAA_TRB + NCAA_STL + NCAA_BLK       7 -1726.149 3470.298 3505.397          273859.9         357 26035        344106     364        1       49       32     40.5
14257              ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_STL + NCAA_games       7 -1726.377 3470.753 3505.852          274201.4         357 14257        344106     364        1       68       43     55.5
30513            ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_3ptpct + NCAA_STL       7 -1726.393 3470.787 3505.886          274226.4         357 30513        344106     364        1       71       45     58.0
30129               ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_TRB + NCAA_STL       7 -1726.456 3470.912 3506.011          274320.5         357 30129        344106     364        1       84       48     66.0
26545               ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_STL + NCAA_BLK       7 -1726.458 3470.916 3506.015          274324.0         357 26545        344106     364        1       85       49     67.0
29617               ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_AST + NCAA_STL       7 -1726.539 3471.078 3506.178          274445.9         357 29617        344106     364        1       94       51     72.5
25017   ws ~ 1 + Pick + Age + Inches + NCAA_fgpct + NCAA_TRB + NCAA_AST + NCAA_STL + NCAA_BLK       8 -1724.822 3469.645 3508.644          271876.0         356 25017        344106     364        1       29      120     74.5
14001 ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_ftpct + NCAA_STL + NCAA_games       8 -1724.913 3469.827 3508.826          272011.8         356 14001        344106     364        1       35      129     82.0
9657  ws ~ 1 + Pick + Age + Inches + NCAA_fgpct + NCAA_TRB + NCAA_STL + NCAA_BLK + NCAA_games       8 -1724.941 3469.881 3508.880          272052.3         356 9657         344106     364        1       38      134     86.0
29873   ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_ftpct + NCAA_TRB + NCAA_STL       8 -1725.003 3470.006 3509.005          272145.2         356 29873        344106     364        1       41      143     92.0
26043                     ws ~ 1 + Pick + Inches + NCAA_fgpct + NCAA_TRB + NCAA_STL + NCAA_BLK       6 -1728.087 3472.174 3503.373          276782.9         358 26043        344106     364        1      183        9     96.0
25009 ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_TRB + NCAA_AST + NCAA_STL + NCAA_BLK       9 -1722.465 3466.930 3509.829          268386.9         355 25009        344106     364        1        2      195     98.5
29361   ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_ftpct + NCAA_AST + NCAA_STL       8 -1725.116 3470.231 3509.230          272313.4         356 29361        344106     364        1       46      157    101.5
8637    ws ~ 1 + Pick + Age + NCAA_fgpct + NCAA_TRB + NCAA_AST + NCAA_STL + NCAA_BLK + NCAA_games       8 -1725.156 3470.312 3509.311          272373.8         356 8637         344106     364        1       50      161    105.5
9649  ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_TRB + NCAA_STL + NCAA_BLK + NCAA_games       9 -1722.592 3467.184 3510.083          268573.6         355 9649         344106     364        1        3      215    109.0
25777 ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_ftpct + NCAA_TRB + NCAA_STL + NCAA_BLK       9 -1722.643 3467.286 3510.185          268649.0         355 25777        344106     364        1        4      222    113.0
25785   ws ~ 1 + Pick + Age + Inches + NCAA_fgpct + NCAA_ftpct + NCAA_TRB + NCAA_STL + NCAA_BLK       8 -1725.229 3470.458 3509.457          272482.4         356 25785        344106     364        1       58      168    113.0
```

Before going forward in determining which of the two models would be better for predicting NBA career win shares, we checked the necessary plots of both models to insure that they met the assumptions of a linear regression model. As you can see below the residual plots for both models do not appear to have any sort of pattern, neither have any highly influential points, nor are either drastically non normal and thus we concluded that both models met the assumptions.

Figure 2: Plots of the six variable model

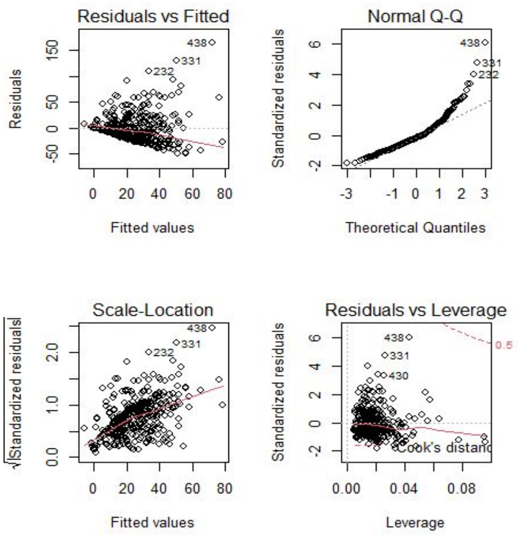


Figure 3: Plots of the seven variable model

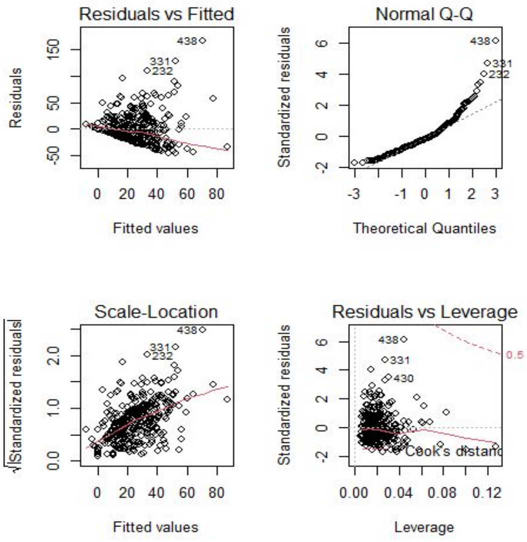
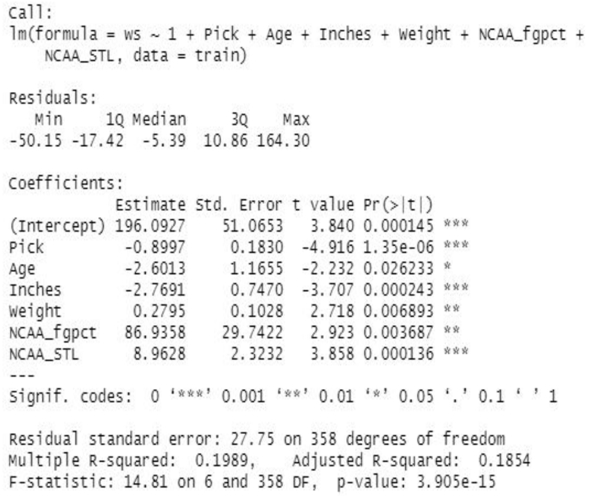
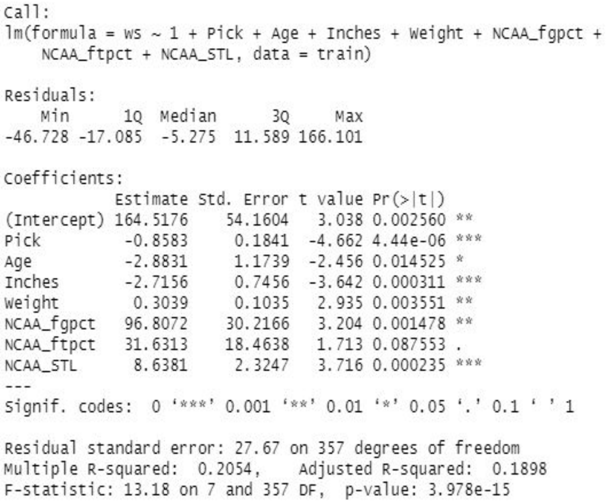


Figure 4: Six variable model



```
Call:
lm(formula = ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct +
    NCAA_STL, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-50.15 -17.42  -5.39  10.86 164.30

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 196.0927    51.0653   3.840 0.000145 ***
Pick         -0.8997     0.1830  -4.916 1.35e-06 ***
Age          -2.6013     1.1655  -2.232 0.026233 *
Inches       -2.7691     0.7470  -3.707 0.000243 ***
Weight        0.2795     0.1028   2.718 0.006893 **
NCAA_fgpct   86.9358    29.7422   2.923 0.003687 **
NCAA_STL      8.9628     2.3232   3.858 0.000136 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.75 on 358 degrees of freedom
Multiple R-squared:  0.1989,	Adjusted R-squared:  0.1854
F-statistic: 14.81 on 6 and 358 DF,  p-value: 3.905e-15
```

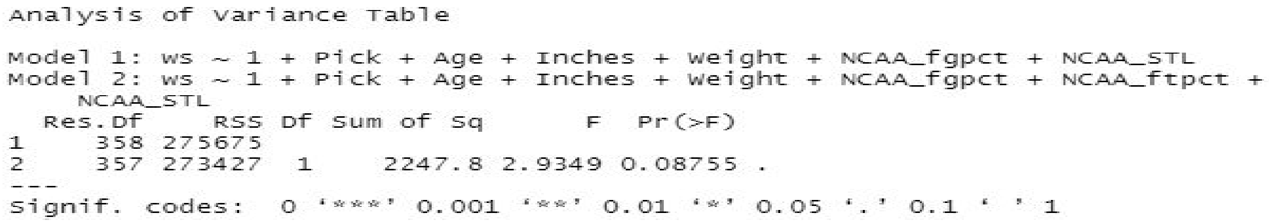
Figure 5: Seven variable model



```
Call:
lm(formula = ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct +
    NCAA_ftpct + NCAA_STL, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-46.728 -17.085  -5.275  11.589 166.101

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 164.5176    54.1604   3.038 0.002560 **
Pick         -0.8583     0.1841  -4.662 4.44e-06 ***
Age          -2.8831     1.1739  -2.456 0.014525 *
Inches       -2.7156     0.7456  -3.642 0.000311 ***
Weight        0.3039     0.1035   2.935 0.003551 **
NCAA_fgpct   96.8072    30.2166   3.204 0.001478 **
NCAA_ftpct   31.6313    18.4638   1.713 0.087553 .
NCAA_STL      8.6381     2.3247   3.716 0.000235 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.67 on 357 degrees of freedom
Multiple R-squared:  0.2054,	Adjusted R-squared:  0.1898
F-statistic: 13.18 on 7 and 357 DF,  p-value: 3.978e-15
```

When we looked at the adjusted r^2 values for both models we saw that there was little difference between them. The six variable model had an adjusted r^2 of .1854 whereas the seven variable model's was only slightly higher at .1898, while these values seem low it is to be expected when analyzing a problem with as many complicating factors as we are. We wanted to verify that there was not a significant difference between the two models. So we performed an ANOVA test between the two models on our training data. Our ANOVA test between the two models provided p value 0.08755. Based on this result we failed to reject Ho, so we concluded that the more complex model wasn't a significantly better fit on the train data than the less complex six variable model, thus we should use the six variable model.

Figure 6: Anova test on six and seven variable models



```
Analysis of Variance Table

Model 1: ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_STL
Model 2: ws ~ 1 + Pick + Age + Inches + Weight + NCAA_fgpct + NCAA_ftpct +
    NCAA_STL
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    358 275675
2    357 273427  1    2247.8 2.9349 0.08755 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Performance:

In order to correctly assess how effective our model is in general, we needed to test its performance on unseen data. Thus, we applied the models to the test dataset that was set aside before our model selection process began.

Figure 7: Plots of six variable model on test data

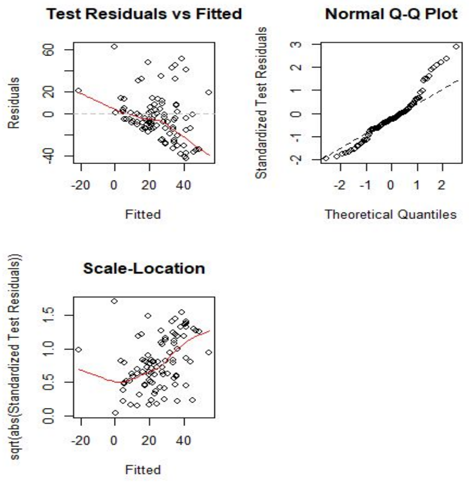
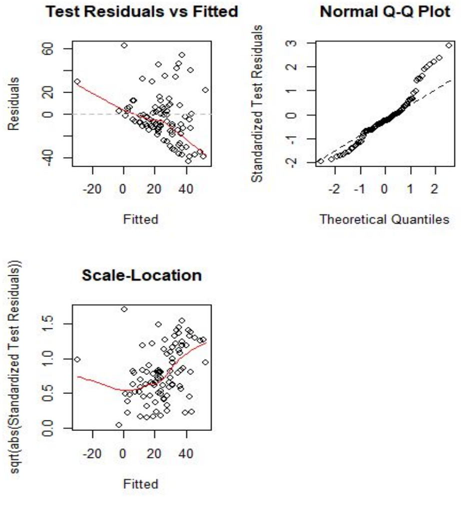


Figure 8: Plots of seven variable model on test data



The residual plots for the six and seven variable models with our test data are very similar and only vary slightly. We need to look further into why these plots are not completely satisfying the assumptions for regression, as the residual plots appear to have some pattern and the normality quantile plot splits from the line in the upper quantile. In order to evaluate which model fit the test data better we used Mean Squared Prediction Error. For our six variable model we got a mean squared prediction error of 471.6991 and for the seven variable model we got a mean squared prediction error of 476.1226 and since a lower value indicates a better fitting model, the six variable model is better.

Overall the six variable model appears to be nearly as accurate as the seven variable model without the added complexity of an additional variable and it is the better model for projecting the career win shares of NBA players.

References:

Leaps, RSQLite, kader, rvest, tidyverse, knitr, goftest, ggplot2, ggiraphExtra, glmulti