Determining NBA Salaries

Colin Busby & Hem Charan Bagul Krishnamurthy

STAT 632

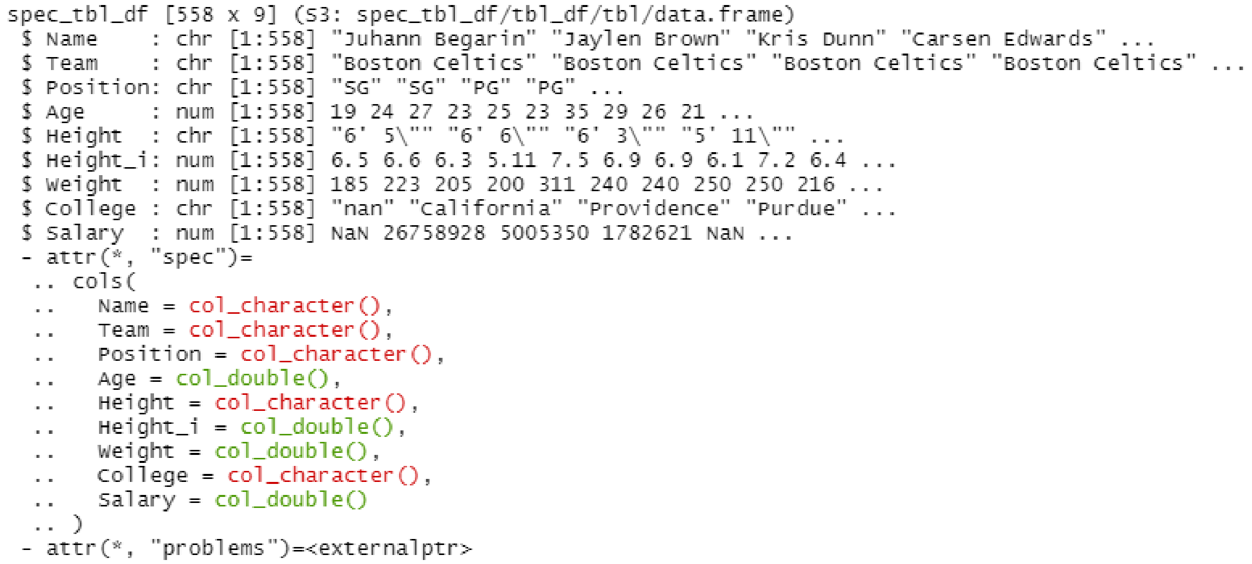
Spring 2022

I. INTRODUCTION

For our final project, Charan and I decided to explore and answer the research question of how much of the variance present in NBA players salaries can be explained by non-skill factors. Sports players are uniquely suited for attempting to answer this question, as virtually every detail of their lives can be and is tracked, from the college they may or may not have attended, to their physical attributes such as their height, weight , and age. Our secondary research question, once we had a working model, was determining how accurate such a model might be at predicting the salary of any given NBA player in the dataset. The primary and secondary questions would be answered with the use of a multiple linear regression model. A tertiary question was how well a logistic regression model could predict whether a player was considered "well-paid" or not.

II. DATA DESCRIPTION

The dataset we decided to use is the 2021-22 NBA Season Active NBA Players dataset from Kaggle, by the user Muhammet Ali Büyüknacar. The dataset is a [558x9] data frame with 9 variables containing data about an NBA player, as seen in Figure 1: *Name*, *Position* (position player plays), *Team*, *Age* (in years), *Height* (height in feet and inches), *Height_i* (height as an integer), *Weight* (lbs), *College* (college attended, if any), and *Salary* (how much a player is paid in dollars). The response variable used in our models is *Salary*, while the chosen predictor variables used are: *Position*, *Team*, *Age*, *Height_i*, *Weight*, and *College*.



```
spec_tbl_df [558 x 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ Name    : chr [1:558] "Juhann Begarin" "Jaylen Brown" "Kris Dunn" "Carsen Edwards" ...
 $ Team    : chr [1:558] "Boston Celtics" "Boston Celtics" "Boston Celtics" "Boston Celtics" ...
 $ Position: chr [1:558] "SG" "SG" "PG" "PG" ...
 $ Age     : num [1:558] 19 24 27 23 25 23 35 29 26 21 ...
 $ Height  : chr [1:558] "6' 5\"" "6' 6\"" "6' 3\"" "5' 11\"" ...
 $ Height_i: num [1:558] 6.5 6.6 6.3 5.11 7.5 6.9 6.9 6.1 7.2 6.4 ...
 $ Weight  : num [1:558] 185 223 205 200 311 240 240 250 250 216 ...
 $ College : chr [1:558] "nan" "California" "Providence" "Purdue" ...
 $ Salary  : num [1:558] NaN 26758928 5005350 1782621 NaN ...
 - attr(*, "spec")=
  .. cols(
  ..   Name = col_character(),
  ..   Team = col_character(),
  ..   Position = col_character(),
  ..   Age = col_double(),
  ..   Height = col_character(),
  ..   Height_i = col_double(),
  ..   Weight = col_double(),
  ..   College = col_character(),
  ..   Salary = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

Figure 1. Original dataframe from Kaggle

Initial tests using Shapiro-Wilks and Breusch-Pagan to test for normality and heteroskedasticity showed that the Shapiro-Wilks test had W = 0.9352 and p-value = 5.455e-13, while the Breusch-Pagan test had a p-value = 0.07744, telling us that the initial data failed the assumption of normality but passed the assumption of homoscedasticity.

The response variable *Salary,* as taken untransformed from the dataset after cleaning, is heavily right skewed, and even after applying a log transformation it is still right skewed (Figure

2). Additionally, we looked at a scatter plot matrix (Figure 3) to determine whether correlation was of concern, which revealed some possible correlation between several of the variables: *Salary* and *Age*, *Age* and *Weight*, and *Weight* and *Height_i*.

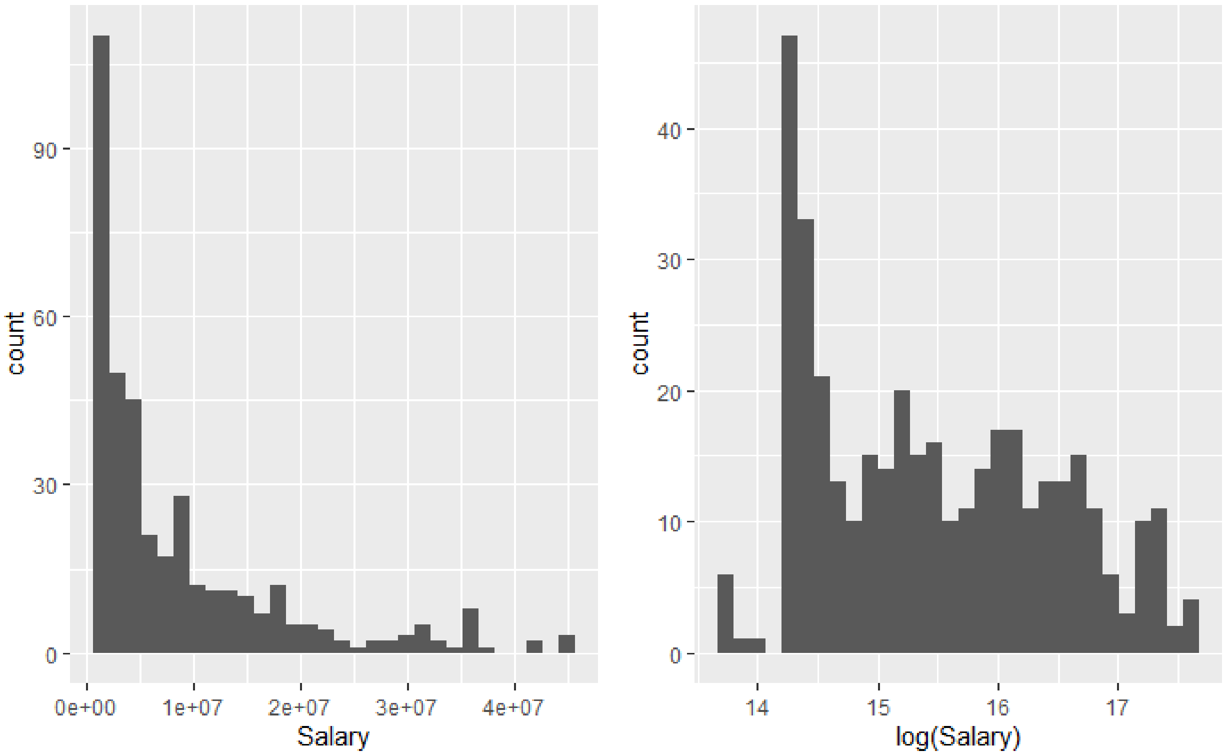


Figure 2. Untransformed and Log-transformed Salary

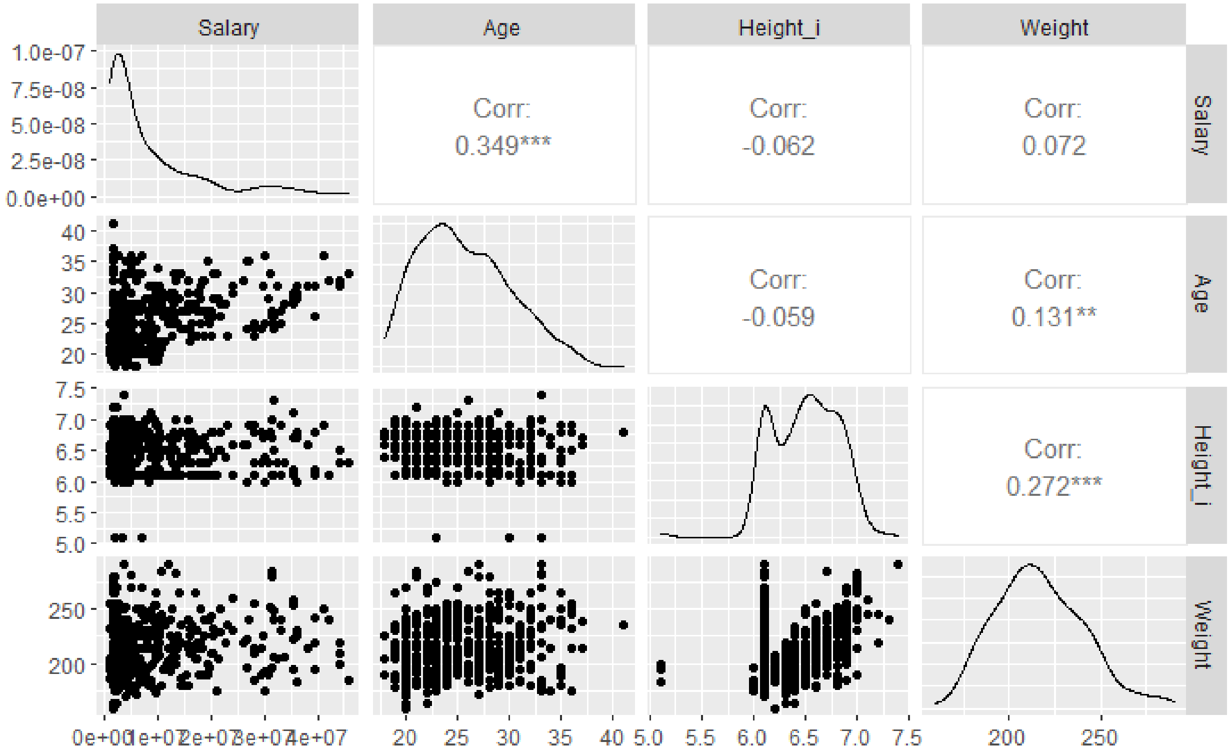
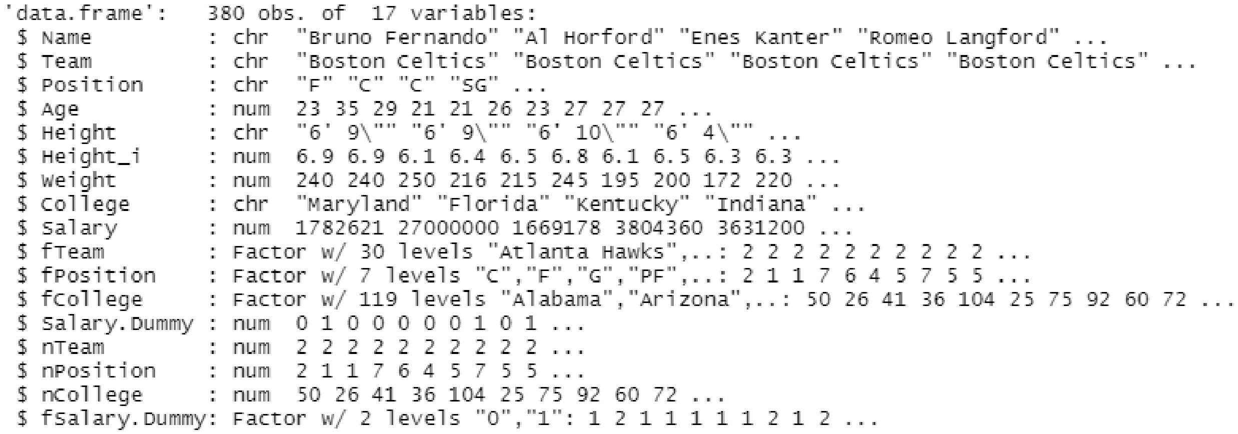


Figure 3. Scatterplot matrix additionally showing correlation in upper right side and probability densities on diagonal.

III. METHODS AND RESULTS

The primary method we used to answer our primary and secondary research questions is multiple linear regression, using both backwards elimination and the Backwards Stepwise Selection (using AIC). Multiple linear regression was chosen as the primary method as we wanted to be able to explain and attribute the variance in salaries to various factors beyond skill. The secondary method we chose was logistic regression, since we wanted to determine if, using an arbitrary cutoff point based on salaries, players could be determined to be "well-paid" or "not" based on non-skill factors.

Before any modeling, we first cleaned the dataset of any NA's that were present in the *Salary* column, as their presence would either imply that the player in question did not ultimately play, or that for whatever reason their salary data was not available. Either way, this would have impacted our model. We then reclassified the character variables *Team*, *Position*, and *College* into factors (*fTeam*, *fPosition*, *fCollege*) in order to make use of categorical data in our models, as well as reclassing *Salary* into a factor (*fSalary*) for later use in the logistic regression model. Additionally, *fTeam*, *fPosition*, and *fCollege* were reclassified again into numeric variables for use in the logistic regression , resulting in a final dataframe of [380 observations x 17 variables] (Figure 4).

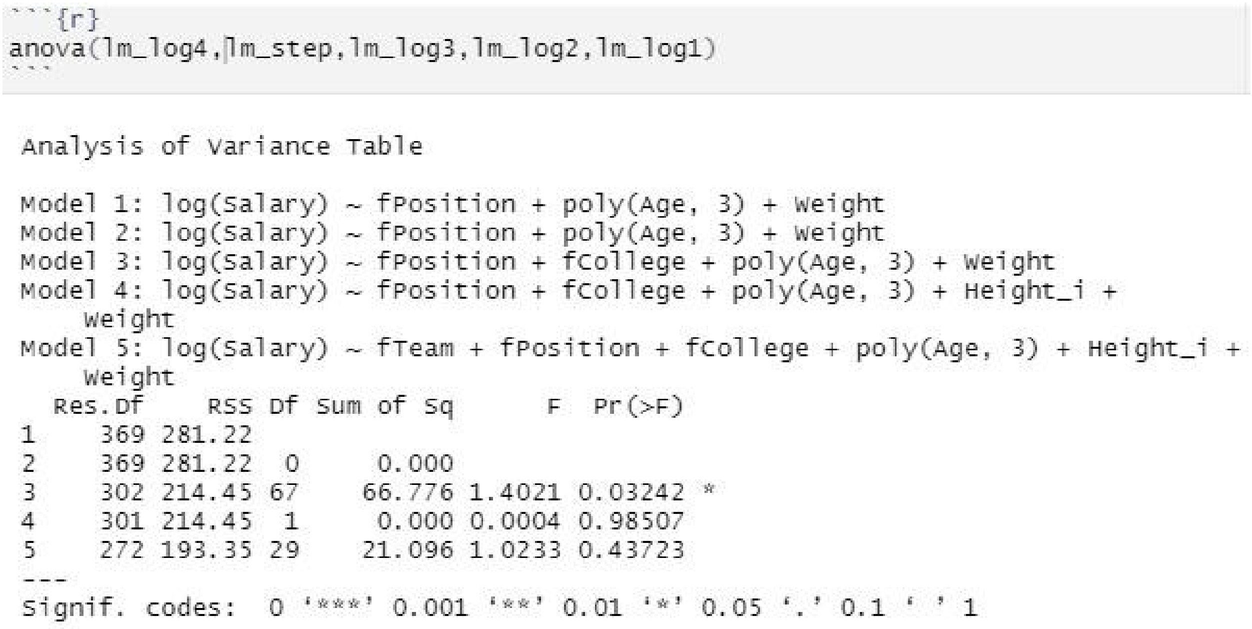


```
'data.frame':   380 obs. of  17 variables:
 $ Name         : chr  "Bruno Fernando" "Al Horford" "Enes Kanter" "Romeo Langford" ...
 $ Team         : chr  "Boston Celtics" "Boston Celtics" "Boston Celtics" "Boston Celtics" ...
 $ Position     : chr  "F" "C" "C" "SG" ...
 $ Age          : num  23 35 29 21 21 26 23 27 27 27 ...
 $ Height       : chr  "6' 9\"" "6' 9\"" "6' 10\"" "6' 4\"" ...
 $ Height_i     : num  6.9 6.9 6.1 6.4 6.5 6.8 6.1 6.5 6.3 6.3 ...
 $ Weight       : num  240 240 250 216 215 245 195 200 172 220 ...
 $ College      : chr  "Maryland" "Florida" "Kentucky" "Indiana" ...
 $ Salary       : num  1782621 27000000 1669178 3804360 3631200 ...
 $ fTeam        : Factor w/ 30 levels "Atlanta Hawks",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ fPosition    : Factor w/ 7 levels "C","F","G","PF",..: 2 1 1 7 6 4 5 7 5 5 ...
 $ fCollege     : Factor w/ 119 levels "Alabama","Arizona",..: 50 26 41 36 104 25 75 92 60 72 ...
 $ Salary.Dummy : num  0 1 0 0 0 0 0 1 0 1 ...
 $ nTeam        : num  2 2 2 2 2 2 2 2 2 2 ...
 $ nPosition    : num  2 1 1 7 6 4 5 7 5 5 ...
 $ nCollege     : num  50 26 41 36 104 25 75 92 60 72 ...
 $ fSalary.Dummy: Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 2 1 2 ...
```

Figure 4. Final data.frame

When creating our MLR model, we first checked the Salary response variable for any possible Box-Cox transformations, revealing a possible need for a log transformation, and then checked for the presence of polynomial terms, which ultimately led us to including the variable Age as an polynomial term to the third power. Third, we checked for the presence of outliers and high leverage points, as the player data included players like Stephen Curry, who we know is extremely well paid due to his skill, and removed those particular points, as they would affect our final model. Finally, we used the Shapiro-Wilks and Breusch-Pagan tests, showing that the final initial model passed the normality assumption, as it had a W of 0.99185, above the 0.95 W threshold for a large sample, telling us the data was “normal enough”, while the Breusch-Pagan test had a p-value of 0.09832, thereby passing the homoscedasticity assumption. This leaves us with a final initial model of *log(Salary) ~ fTeam + fPosition + fCollege + Age + Height_i + Weight*, with an adjusted R^2 of 0.282.

Once we had a final initial model , including the log-transformation and polynomial terms, and the removal of outliers and high leverage points, we started with backwards elimination, by manually removing variables according to their p-value. This resulted in the MLR model *log(Salary) ~ fPosition+fCollege+poly(Age,3)+Height_i+Weight*, with an adjusted R^2 value of 0.2828 (Figure 5). It should be noted that had we removed *fCollege*, the resulting model would be identical to the Backwards Stepwise model discussed below (ANOVA, Figure 6), but we decided to keep *fCollege* in the model, since at least one college was significant and because many colleges and universities have reputations for their sports programs, which may have an effect on the salaries of some players.



~~~
```{r}
anova(lm_log4,lm_step,lm_log3,lm_log2,lm_log1)
```
~~~

```
Analysis of Variance Table

Model 1: log(Salary) ~ fPosition + poly(Age, 3) + Weight
Model 2: log(Salary) ~ fPosition + poly(Age, 3) + Weight
Model 3: log(Salary) ~ fPosition + fCollege + poly(Age, 3) + Weight
Model 4: log(Salary) ~ fPosition + fCollege + poly(Age, 3) + Height_i +
    Weight
Model 5: log(Salary) ~ fTeam + fPosition + fCollege + poly(Age, 3) + Height_i +
    Weight
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    369 281.22
2    369 281.22  0     0.000
3    302 214.45 67    66.776 1.4021 0.03242 *
4    301 214.45  1     0.000 0.0004 0.98507
5    272 193.35 29    21.096 1.0233 0.43723
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6. ANOVA of MLR models

The second multiple linear regression method we used was the Backwards Stepwise selection, an automatic variable selection utilizing Akaike Information Criterion (AIC). Using the same final initial model as the manual backwards elimination process, we determined the final reduced model to be *log(Salary) ~ fPosition + poly(Age, 3) + Weight*, with an AIC score of -92.39 and an adjusted R^2 of 0.2302 (Figure 7).

Running Shapiro-Wilks tests for both models shows that the backwards elimination model (*lm_log3*) had a p-value of 0.01796 but a W of 0.9908, while the stepwise model (*lm_step*) had a p-value of 0.0007663 but a W of 0.98555, fulling the conditions for the assumption of normality for a large sample for both models. However, for the Breusch-Pagan tests, the *lm_log3* model had a p-value of 0.0626, meeting the assumption of homoscedasticity, but the *lm_step* model had a p-value of 0.0001093, rejecting the assumption of homoscedasticity.

A final check of both models using *performance::check_model()* revealed that for *lm_log3*, there were potential collinearity issues between *fCollege* and *fPosition*, with the *vif()* function showing *fPosition* and *fCollege* with variance inflation factors of 11.18 and 7.65 respectively, but that otherwise the two models were fairly similar (Figure 8, Figure 9), showing approximate normality of residuals. Despite the potential collinearity issues, a comparison of the two models using the *predict()* function in R to predict the salaries of several players from the dataset revealed that the *lm_log3* model performed better than the *lm_step* model, achieving a closer result to the actual salaries (Figure 10), though better is relative to the adjusted R^2 values of 0.2828 and 0.2302, respectively.

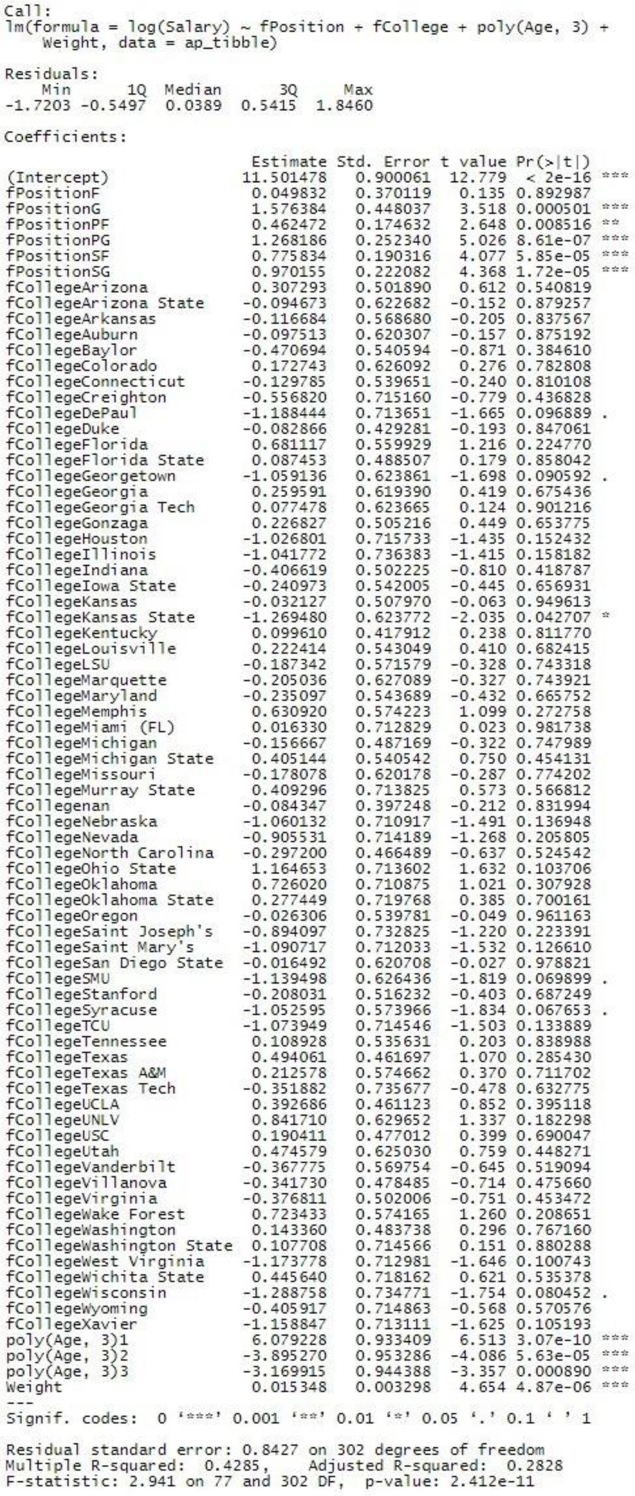
The second regression analysis we performed was a logistic regression, using a 65% cutoff point of the players' salaries as "well-paid" or "not". Using the *train()* function from the caret package for R, we fit both a full model *(fSalary.Dummy ~ nTeam + nPosition + nCollege + poly(Age,3) + Height_i + Weight*) and a reduced model using the argument *method = "glmStepAIC"* (*fSalary.Dummy ~ nPosition + poly(Age,3) + Height_i + Weight*). We then performed a confusion matrix for both, resulting in an average accuracy of 0.713 for the full model and an average accuracy of 0.715 for the reduced model (Figure 11). Both models were then fitted to an ROC curve and AUC, showing that the full model had an AUC of 0.8593 while the reduced model had an AUC of 0.8741 (Figure 12), suggesting that the reduced model was better at predicting which players would be classed as being "well-paid", that is being being paid more than the 65% cutoff.

IV. CONCLUSION

Overall, we found the multiple linear regression model *lm_log3* to be a better fit for interpreting variances in salaries and predicting salaries of current players than the *lm_step* model, though with such low adjusted R^2s (0.2828 and 0.2302, respectively), both models are relative in their usefulness for interpretation and prediction of the salaries of NBA players. The logistic regression reduced model was much better in terms of accuracy with an AUC of 0.8741, but only in regards to classification of the salaries as being above or below 65%. In short, without the inclusion of skill factors in our models, the models are only useful in explaining a small percentage of the variation in NBA salaries, which was something we expected at the beginning. After all, players like Stephen Curry aren't paid millions for their age or height, they're paid millions for their skills.

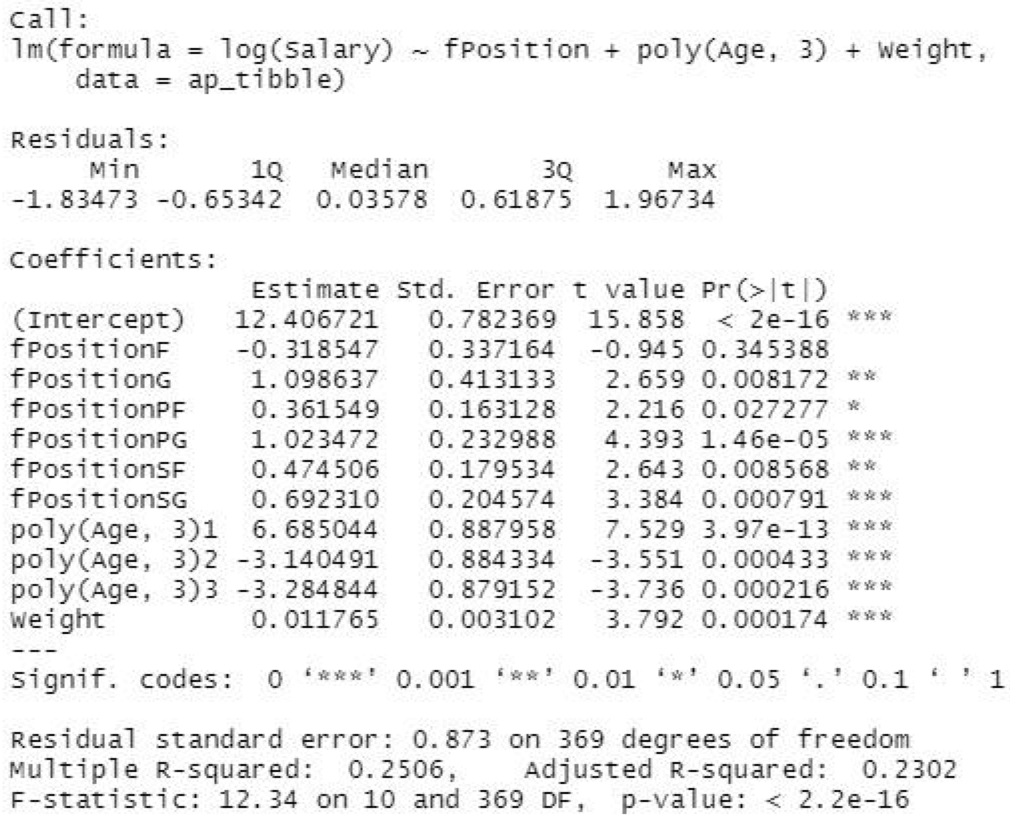
Some ideas for future work would be to include race into the model, to see if it has any effect on salaries, given racial disparities in wages in America. Another idea would be to use the models on data from the Women's NBA, given gender disparities in wages. Or perhaps we could take it from the opposite angle, and create a model to interpret and predict NBA salaries using only skill factors to do so.

Large Figures



```
Call:
lm(formula = log(Salary) ~ fPosition + fCollege + poly(Age, 3) + 
    Weight, data = ap_tibble)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.7203 -0.5497  0.0389  0.5415  1.8460 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              11.501478   0.900061  12.779  < 2e-16 ***
fPositionF                0.049832   0.370119   0.135 0.892987    
fPositionG                1.576384   0.448037   3.518 0.000501 ***
fPositionPF               0.462472   0.174632   2.648 0.008516 ** 
fPositionPG               1.268186   0.252340   5.026 8.61e-07 ***
fPositionSF               0.775834   0.190316   4.077 5.85e-05 ***
fPositionSG               0.970155   0.222082   4.368 1.72e-05 ***
fCollegeArizona           0.307293   0.501890   0.612 0.540819    
fCollegeArizona State    -0.094673   0.622682  -0.152 0.879257    
fCollegeArkansas         -0.116684   0.568680  -0.205 0.837567    
fCollegeAuburn           -0.097513   0.620307  -0.157 0.875192    
fCollegeBaylor           -0.470694   0.540594  -0.871 0.384610    
fCollegeColorado          0.172743   0.626092   0.276 0.782808    
fCollegeConnecticut      -0.129785   0.539651  -0.240 0.810108    
fCollegeCreighton        -0.556820   0.715160  -0.779 0.436828    
fCollegeDePaul           -1.188444   0.713651  -1.665 0.096889 .  
fCollegeDuke             -0.082866   0.429281  -0.193 0.847061    
fCollegeFlorida           0.681117   0.559929   1.216 0.224770    
fCollegeFlorida State     0.087453   0.488507   0.179 0.858042    
fCollegeGeorgetown       -1.059136   0.623861  -1.698 0.090592 .  
fCollegeGeorgia           0.259591   0.619390   0.419 0.675436    
fCollegeGeorgia Tech      0.077478   0.623665   0.124 0.901216    
fCollegeGonzaga           0.226827   0.505216   0.449 0.653775    
fCollegeHouston          -1.026801   0.715733  -1.435 0.152432    
fCollegeIllinois         -1.041772   0.736383  -1.415 0.158182    
fCollegeIndiana          -0.406619   0.502225  -0.810 0.418787    
fCollegeIowa State       -0.240973   0.542005  -0.445 0.656931    
fCollegeKansas           -0.032127   0.507970  -0.063 0.949613    
fCollegeKansas State     -1.269480   0.623772  -2.035 0.042707 *  
fCollegeKentucky          0.099610   0.417912   0.238 0.811770    
fCollegeLouisville        0.222414   0.543049   0.410 0.682415    
fCollegeLSU              -0.187342   0.571579  -0.328 0.743318    
fCollegeMarquette        -0.205036   0.627089  -0.327 0.743921    
fCollegeMaryland         -0.235097   0.543689  -0.432 0.665752    
fCollegeMemphis           0.630920   0.574223   1.099 0.272758    
fCollegeMiami (FL)        0.016330   0.712829   0.023 0.981738    
fCollegeMichigan         -0.156667   0.487169  -0.322 0.747989    
fCollegeMichigan State    0.405144   0.540542   0.750 0.454131    
fCollegeMissouri         -0.178078   0.620178  -0.287 0.774202    
fCollegeMurray State      0.409296   0.713825   0.573 0.566812    
fCollegenan              -0.084347   0.397248  -0.212 0.831994    
fCollegeNebraska         -1.060132   0.710917  -1.491 0.136948    
fCollegeNevada           -0.905531   0.714189  -1.268 0.205805    
fCollegeNorth Carolina   -0.297200   0.466489  -0.637 0.524542    
fCollegeOhio State        1.164653   0.713602   1.632 0.103706    
fCollegeOklahoma          0.726020   0.710875   1.021 0.307928    
fCollegeOklahoma State    0.277449   0.719768   0.385 0.700161    
fCollegeOregon           -0.026306   0.539781  -0.049 0.961163    
fCollegeSaint Joseph's   -0.894097   0.732825  -1.220 0.223391    
fCollegeSaint Mary's     -1.090717   0.712033  -1.532 0.126610    
fCollegeSan Diego State  -0.016492   0.620708  -0.027 0.978821    
fCollegeSMU              -1.139498   0.626436  -1.819 0.069899 .  
fCollegeStanford         -0.208031   0.516232  -0.403 0.687249    
fCollegeSyracuse         -1.052595   0.573966  -1.834 0.067653 .  
fCollegeTCU              -1.073949   0.714546  -1.503 0.133889    
fCollegeTennessee         0.108928   0.535631   0.203 0.838988    
fCollegeTexas             0.494061   0.461697   1.070 0.285430    
fCollegeTexas A&M         0.212578   0.574662   0.370 0.711702    
fCollegeTexas Tech       -0.351882   0.735677  -0.478 0.632775    
fCollegeUCLA              0.392686   0.461123   0.852 0.395118    
fCollegeUNLV              0.841710   0.629652   1.337 0.182298    
fCollegeUSC               0.190411   0.477012   0.399 0.690047    
fCollegeUtah              0.474579   0.625030   0.759 0.448271    
fCollegeVanderbilt       -0.367775   0.569754  -0.645 0.519094    
fCollegeVillanova        -0.341730   0.478485  -0.714 0.475660    
fCollegeVirginia         -0.376811   0.502006  -0.751 0.453472    
fCollegeWake Forest       0.723433   0.574165   1.260 0.208651    
fCollegeWashington        0.143360   0.483738   0.296 0.767160    
fCollegeWashington State  0.107708   0.714566   0.151 0.880288    
fCollegeWest Virginia    -1.173778   0.712981  -1.646 0.100743    
fCollegeWichita State     0.445640   0.718162   0.621 0.535378    
fCollegeWisconsin        -1.288758   0.734771  -1.754 0.080452 .  
fCollegeWyoming          -0.405917   0.714863  -0.568 0.570576    
fCollegeXavier           -1.158847   0.713111  -1.625 0.105193    
poly(Age, 3)1             6.079228   0.933409   6.513 3.07e-10 ***
poly(Age, 3)2            -3.895270   0.953286  -4.086 5.63e-05 ***
poly(Age, 3)3            -3.169915   0.944388  -3.357 0.000890 ***
Weight                    0.015348   0.003298   4.654 4.87e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8427 on 302 degrees of freedom
Multiple R-squared:  0.4285,	Adjusted R-squared:  0.2828 
F-statistic: 2.941 on 77 and 302 DF,  p-value: 2.412e-11
```

Figure 5. MLR Reduced Model



```
Call:
lm(formula = log(Salary) ~ fPosition + poly(Age, 3) + Weight,
    data = ap_tibble)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.83473 -0.65342  0.03578  0.61875  1.96734 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   12.406721   0.782369  15.858  < 2e-16 ***
fPositionF    -0.318547   0.337164  -0.945 0.345388    
fPositionG     1.098637   0.413133   2.659 0.008172 ** 
fPositionPF    0.361549   0.163128   2.216 0.027277 *  
fPositionPG    1.023472   0.232988   4.393 1.46e-05 ***
fPositionSF    0.474506   0.179534   2.643 0.008568 ** 
fPositionSG    0.692310   0.204574   3.384 0.000791 ***
poly(Age, 3)1  6.685044   0.887958   7.529 3.97e-13 ***
poly(Age, 3)2 -3.140491   0.884334  -3.551 0.000433 ***
poly(Age, 3)3 -3.284844   0.879152  -3.736 0.000216 ***
Weight         0.011765   0.003102   3.792 0.000174 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.873 on 369 degrees of freedom
Multiple R-squared:  0.2506,	Adjusted R-squared:  0.2302 
F-statistic: 12.34 on 10 and 369 DF,  p-value: < 2.2e-16
```

Figure 7. MLR Step Model

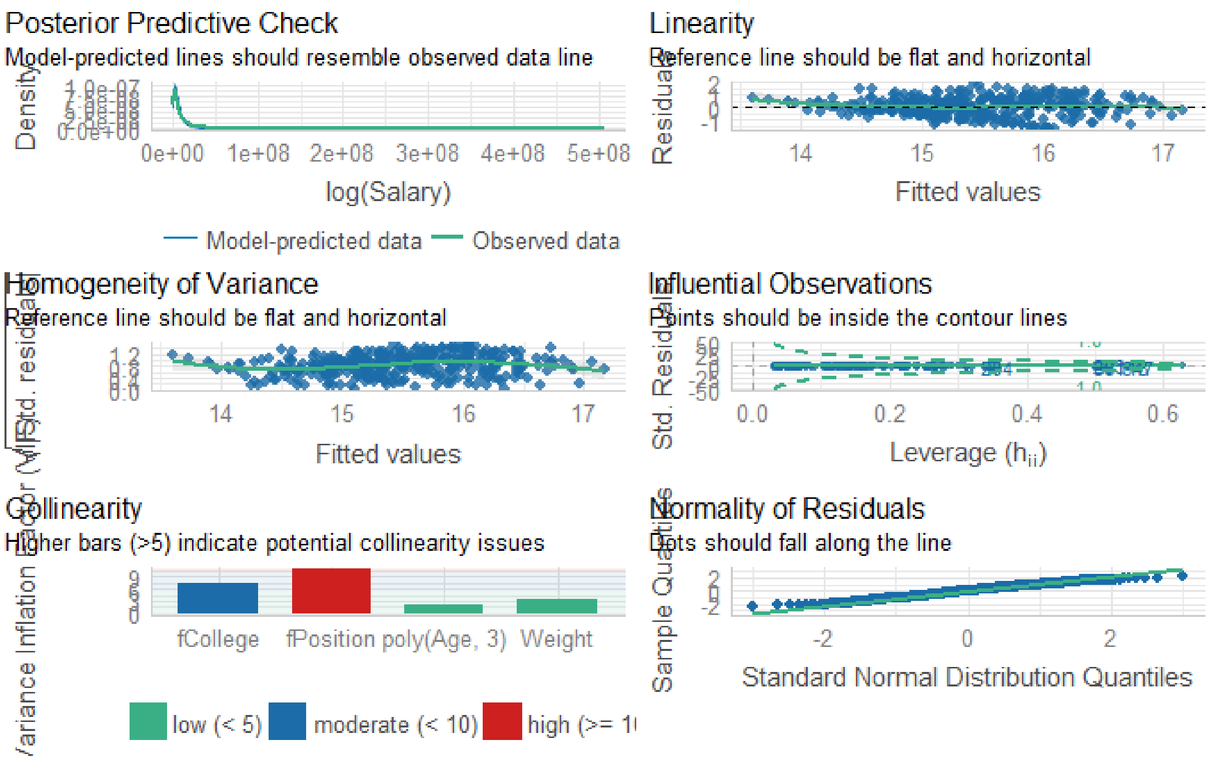


Figure 8. Check_model of backwards elimination model

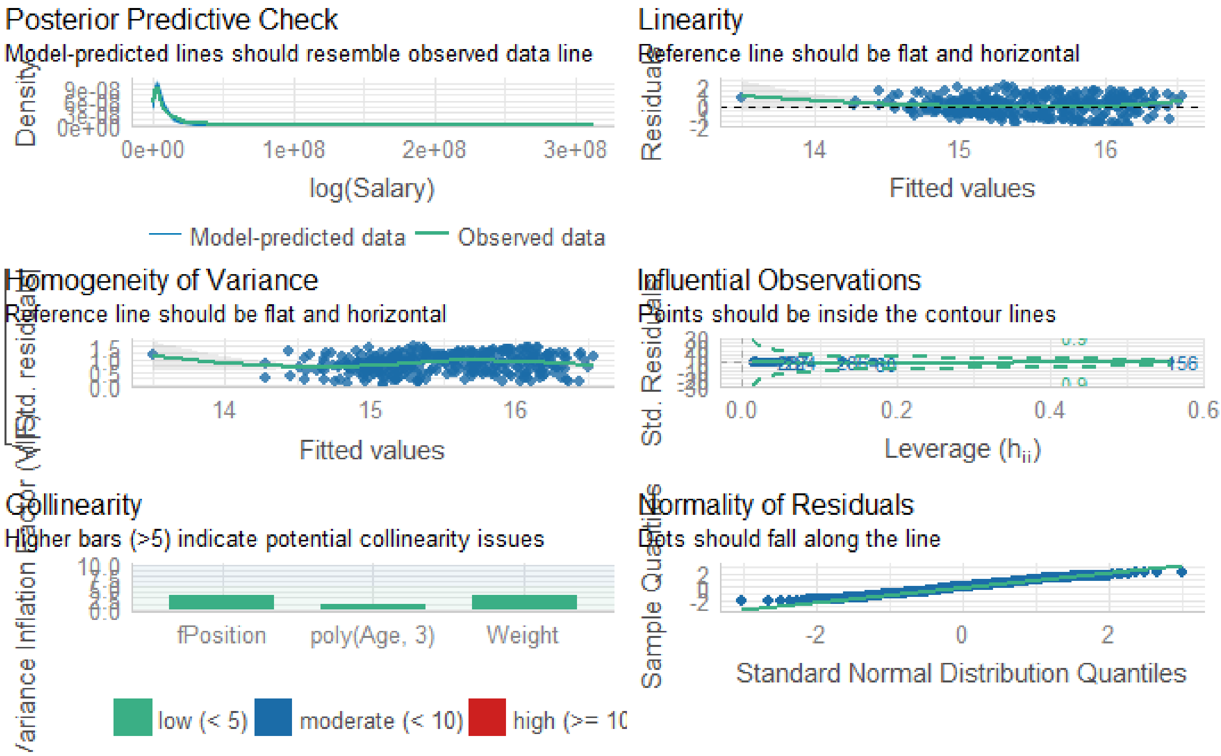
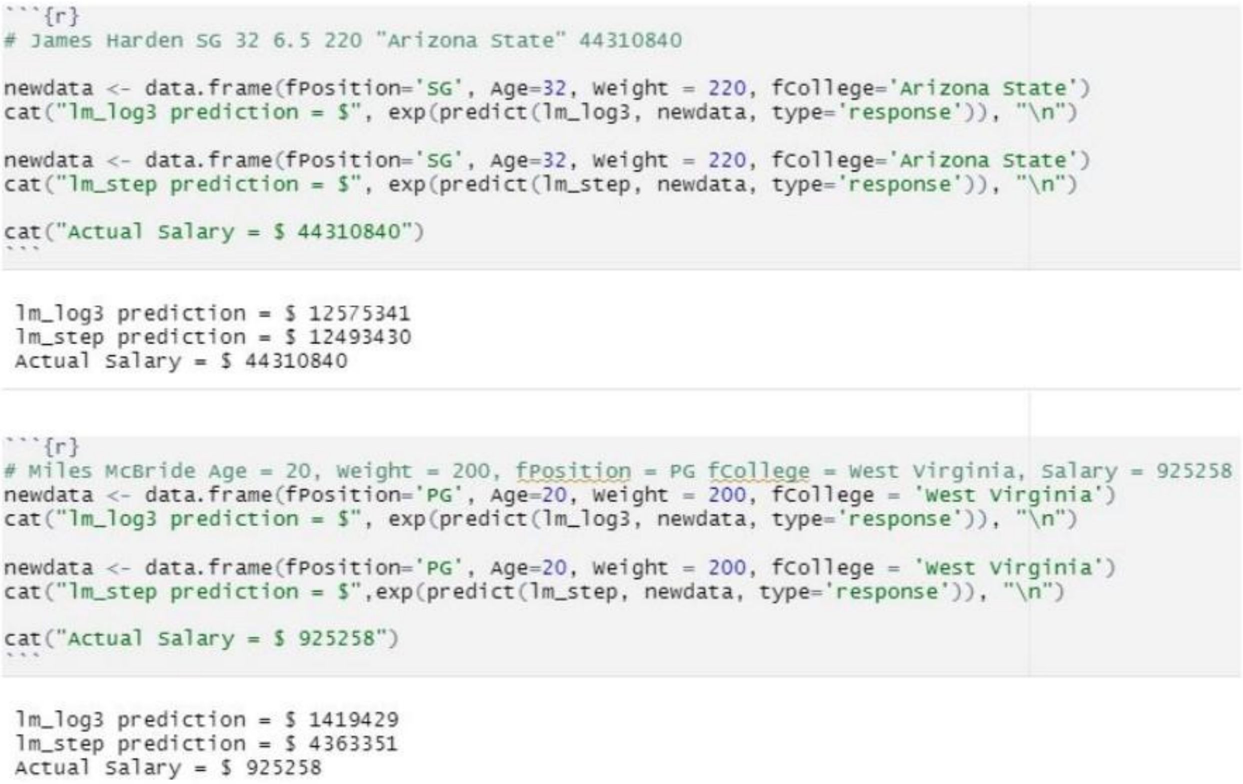


Figure 9. Check_model of stepwise model



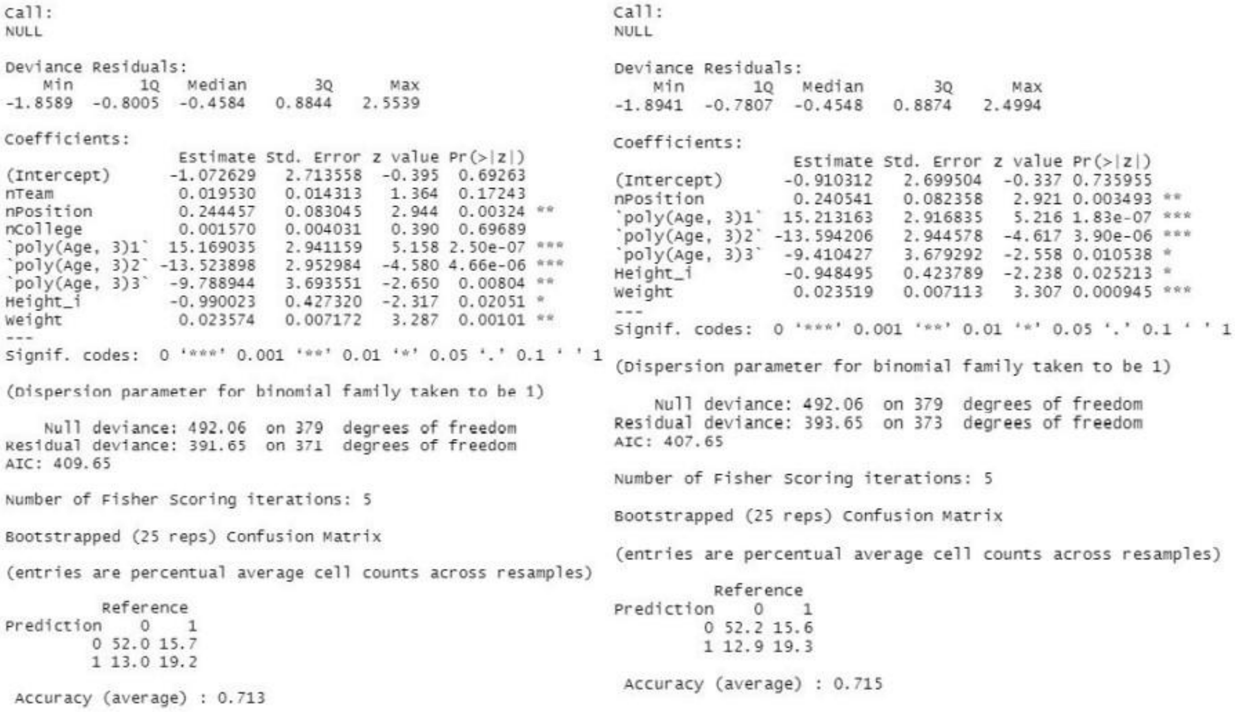
~~~
```{r}
# James Harden SG 32 6.5 220 "Arizona State" 44310840

newdata <- data.frame(fPosition='SG', Age=32, Weight = 220, fCollege='Arizona State')
cat("lm_log3 prediction = $", exp(predict(lm_log3, newdata, type='response')), "\n")

newdata <- data.frame(fPosition='SG', Age=32, Weight = 220, fCollege='Arizona State')
cat("lm_step prediction = $", exp(predict(lm_step, newdata, type='response')), "\n")

cat("Actual Salary = $ 44310840")
```
~~~

```
 lm_log3 prediction = $ 12575341
 lm_step prediction = $ 12493430
 Actual Salary = $ 44310840
```

~~~
```{r}
# Miles McBride Age = 20, Weight = 200, fPosition = PG fCollege = West Virginia, Salary = 925258
newdata <- data.frame(fPosition='PG', Age=20, Weight = 200, fCollege = 'West Virginia')
cat("lm_log3 prediction = $", exp(predict(lm_log3, newdata, type='response')), "\n")

newdata <- data.frame(fPosition='PG', Age=20, Weight = 200, fCollege = 'West Virginia')
cat("lm_step prediction = $",exp(predict(lm_step, newdata, type='response')), "\n")

cat("Actual Salary = $ 925258")
```
~~~

```
 lm_log3 prediction = $ 1419429
 lm_step prediction = $ 4363351
 Actual Salary = $ 925258
```

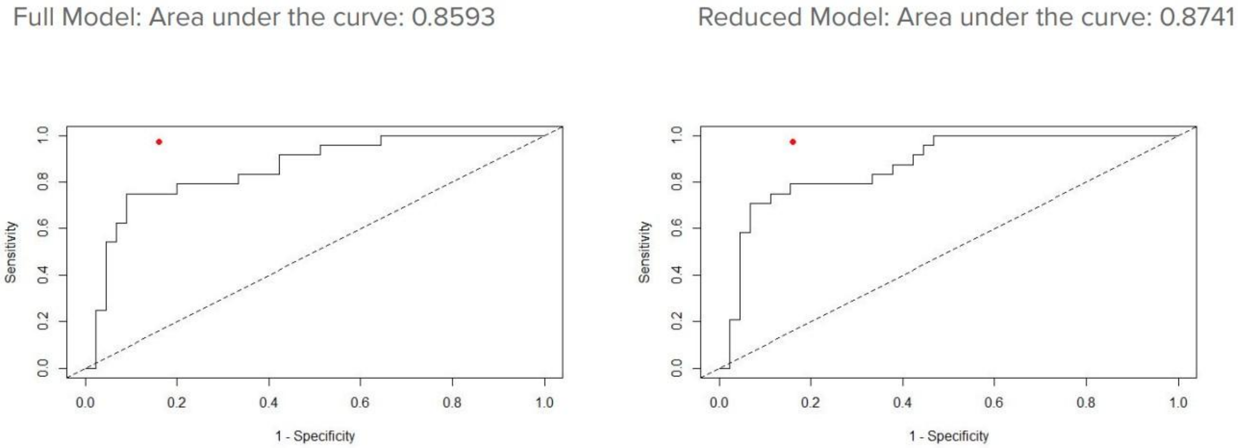
Figure 10. Salary predictions using backwards elimination (lm_log3) and stepwise (lm_step) models



```
Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8589  -0.8005  -0.4584   0.8844   2.5539

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.072629   2.713558  -0.395  0.69263
nTeam             0.019530   0.014313   1.364  0.17243
nPosition         0.244457   0.083045   2.944  0.00324 **
nCollege          0.001570   0.004031   0.390  0.69689
`poly(Age, 3)1`  15.169035   2.941159   5.158 2.50e-07 ***
`poly(Age, 3)2` -13.523898   2.952984  -4.580 4.66e-06 ***
`poly(Age, 3)3`  -9.788944   3.693551  -2.650  0.00804 **
Height_i         -0.990023   0.427320  -2.317  0.02051 *
Weight            0.023574   0.007172   3.287  0.00101 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 492.06  on 379  degrees of freedom
Residual deviance: 391.65  on 371  degrees of freedom
AIC: 409.65

Number of Fisher Scoring iterations: 5

Bootstrapped (25 reps) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0 52.0 15.7
         1 13.0 19.2

 Accuracy (average) : 0.713
```

```
Call:
NULL

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8941  -0.7807  -0.4548   0.8874   2.4994

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.910312   2.699504  -0.337 0.735955
nPosition         0.240541   0.082358   2.921 0.003493 **
`poly(Age, 3)1`  15.213163   2.916835   5.216 1.83e-07 ***
`poly(Age, 3)2` -13.594206   2.944578  -4.617 3.90e-06 ***
`poly(Age, 3)3`  -9.410427   3.679292  -2.558 0.010538 *
Height_i         -0.948495   0.423789  -2.238 0.025213 *
Weight            0.023519   0.007113   3.307 0.000945 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 492.06  on 379  degrees of freedom
Residual deviance: 393.65  on 373  degrees of freedom
AIC: 407.65

Number of Fisher Scoring iterations: 5

Bootstrapped (25 reps) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0 52.2 15.6
         1 12.9 19.3

 Accuracy (average) : 0.715
```

Figure 11. Confusion matrices of the logistic regression full model (left) and the reduced model (right)



Full Model: Area under the curve: 0.8593

Reduced Model: Area under the curve: 0.8741

Figure 12. AUC/ROC of the logistic regression full model (left) and the reduced model (right)

Code Appendix

Github Repository: BusbyCI11/632-Project