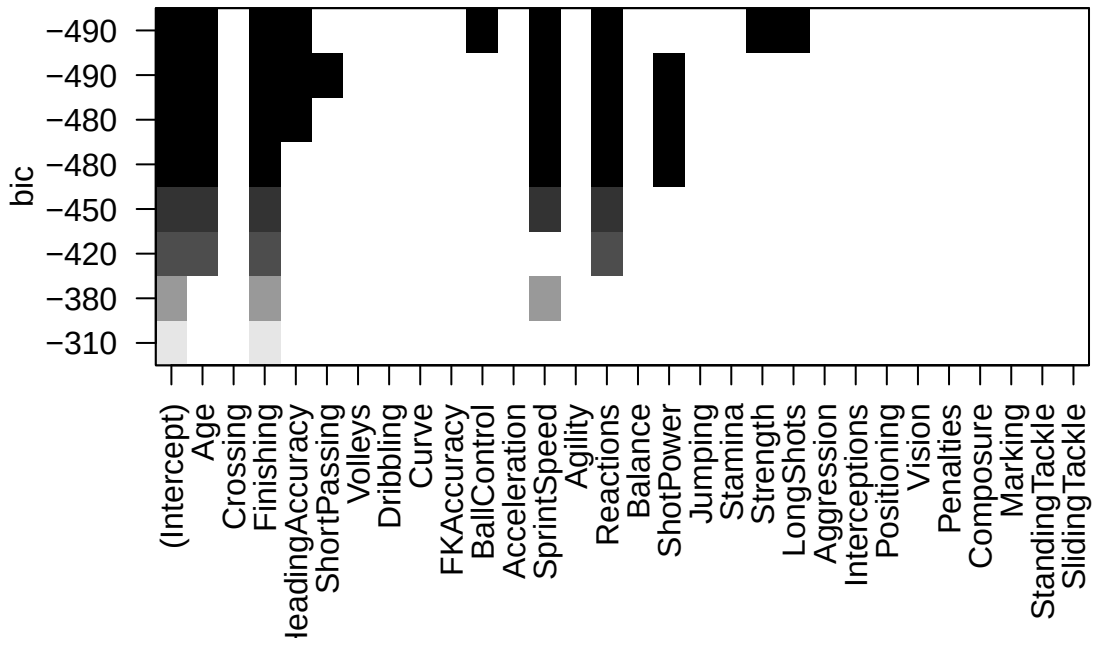
Multiple Linear Regression: Choosing Which Predictors to Include in a Model

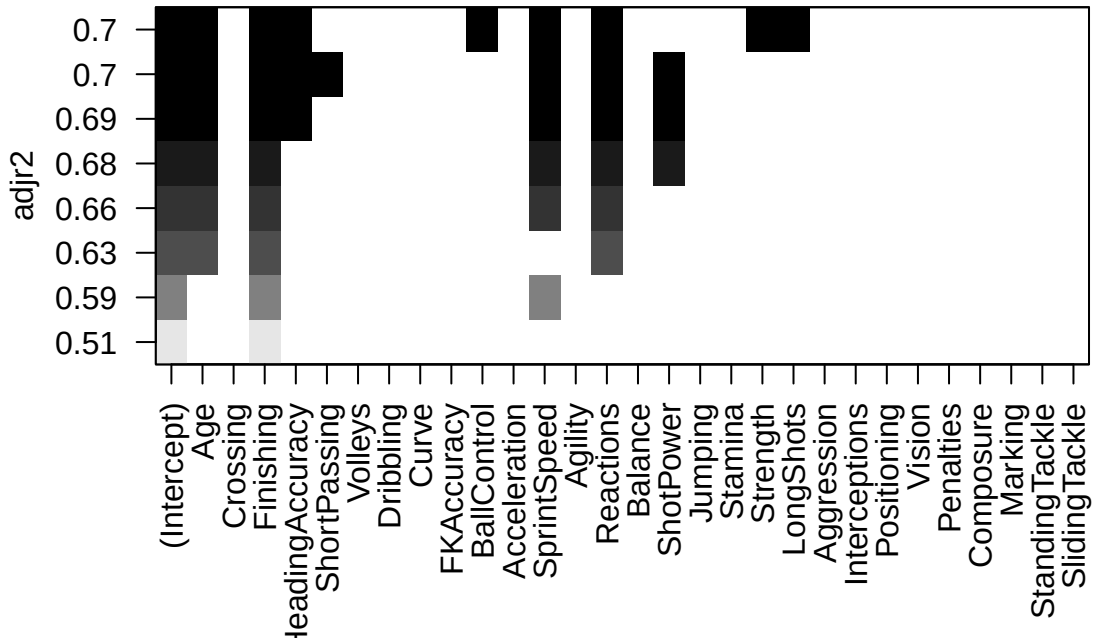
Andrew Levine

November 2021

```
library(leaps)
FIFA19B <- read.csv(file="/Users/andrewlevine/Downloads/Statistics II/FIFA19_VersionB.csv")
FIFA19B2 <- na.omit(FIFA19B)
best_subsetsFIFA19B <- regsubsets(Value ~ ., data=FIFA19B2)
plot(best_subsetsFIFA19B, scale="bic")
```




```
plot(best_subsetsFIFA19B, scale="adjr2")
```




```
lm_info1_FIFA19B <- lm(Value ~ Age + Finishing + HeadingAccuracy + BallControl + SprintSpeed
                       + Reactions + Strength + LongShots, data = FIFA19B2)
lm_info2_FIFA19B <- lm(Value ~ Age + Finishing + HeadingAccuracy + ShortPassing + SprintSpeed
                       + Reactions + ShotPower, data = FIFA19B2)
summary(lm_info1_FIFA19B)
```

```
##
## Call:
## lm(formula = Value ~ Age + Finishing + HeadingAccuracy + BallControl +
##     SprintSpeed + Reactions + Strength + LongShots, data = FIFA19B2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.008  -3.249  -0.730   2.187  43.675
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -115.19503    4.83047 -23.848  < 2e-16 ***
## Age               -0.85573    0.09077  -9.427  < 2e-16 ***
## Finishing          0.65333    0.08752   7.465 4.55e-13 ***
## HeadingAccuracy    0.16530    0.04920   3.360 0.000849 ***
## BallControl        0.30099    0.07858   3.830 0.000147 ***
## SprintSpeed        0.23655    0.03089   7.658 1.22e-13 ***
## Reactions          0.34386    0.06232   5.517 5.90e-08 ***
## Strength           0.13700    0.03540   3.869 0.000126 ***
## LongShots          0.17624    0.05207   3.385 0.000777 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.828 on 438 degrees of freedom
## Multiple R-squared:  0.705,  Adjusted R-squared:  0.6996
## F-statistic: 130.9 on 8 and 438 DF,  p-value: < 2.2e-16
```

```
summary(lm_info2_FIFA19B)
```

```
##
## Call:
## lm(formula = Value ~ Age + Finishing + HeadingAccuracy + ShortPassing +
##     SprintSpeed + Reactions + ShotPower, data = FIFA19B2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.188  -3.137  -0.920   2.188  44.358
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -110.39532    4.53407 -24.348  < 2e-16 ***
## Age               -0.82964    0.09030  -9.187  < 2e-16 ***
## Finishing          0.71728    0.08026   8.937  < 2e-16 ***
## HeadingAccuracy    0.19374    0.04389   4.414 1.28e-05 ***
## ShortPassing       0.20992    0.05949   3.528 0.000462 ***
## SprintSpeed        0.23145    0.03097   7.472 4.30e-13 ***
## Reactions          0.31074    0.06335   4.905 1.32e-06 ***
## ShotPower          0.27069    0.06158   4.395 1.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.864 on 439 degrees of freedom
## Multiple R-squared:  0.7007, Adjusted R-squared:  0.696
## F-statistic: 146.8 on 7 and 439 DF,  p-value: < 2.2e-16
```

Using best subsets regression with the BIC criterion, two models seem to fit the data better than the rest, as they both have the lowest BIC in the chart with all the potential models. The two models that have the lowest BIC of -490 contain the information as follows:

Model 1:
Predictors: Age, Finishing, Heading Accuracy, Ball Control, Sprint Speed, Reactions, Strength, Long Shots
Statistics: $R^2 = 0.705$, adjusted $R^2 = 0.6996$, p-value $< 2.2 * 10^{-16}$, $s = 5.828$
Model Equation: $\hat{Value}_i$ = -115.195 - $0.8557(Age_i)$ + $0.6533(Finishing_i)$ + $0.301(BallControl_i)$ + $0.2366(SprintSpeed_i)$ + $0.3439(Reactions_i)$ + $0.137(Strength_i)$ + $0.1762(LongShots_i)$

Model 2:
Predictors: Age, Finishing, Heading Accuracy, Short Passing, Sprint Speed, Reactions, Shot Power
Statistics: $R^2 = 0.7007$, adjusted $R^2 = 0.696$, p-value $< 2.2 * 10^{-16}$, $s = 5.864$
Model Equation: $\hat{Value}_i$ = -110.3953 - $0.8296(Age_i)$ + $0.7173(Finishing_i)$ + $0.301(HeadingAccuracy_i)$ + $0.2099(ShortPassing_i)$ + $0.2315(SprintSpeed_i)$ + $0.3107(Reactions_i)$ + $0.2707(ShotPower_i)$

Using best subsets regression with the adjusted R^2 criterion, two models seem to fit the data better than the rest, as they both have the highest adjusted R^2 in the chart with all the potential models. The two models that have the highest adjusted R^2 of 0.7 contain the information as follows:

Model 1:
Predictors: Age, Finishing, Heading Accuracy, Ball Control, Sprint Speed, Reactions, Strength, Long Shots
Statistics: $R^2 = 0.705$, adjusted $R^2 = 0.6996$, p-value $< 2.2 * 10^{-16}$, $s = 5.828$

Model Equation: $\hat{Value}_i$ = -115.195 - 0.8557(Age_i) + 0.6533($Finishing_i$) + 0.301($BallControl_i$) + 0.2366($SprintSpeed_i$) + 0.3439($Reactions_i$) + 0.137($Strength_i$) + 0.1762($LongShots_i$)

Model 2:
Predictors: Age, Finishing, Heading Accuracy, Short Passing, Sprint Speed, Reactions, Shot Power
Statistics: $R^2 = 0.7007$, adjusted $R^2 = 0.696$, p-value $< 2.2 * 10^{-16}$, $s = 5.864$
Model Equation: $\hat{Value}_i$ = -110.3953 - 0.8296(Age_i) + 0.7173($Finishing_i$) + 0.301($HeadingAccuracy_i$) + 0.2099($ShortPassing_i$) + 0.2315($SprintSpeed_i$) + 0.3107($Reactions_i$) + 0.2707($ShotPower_i$)

As you can see, the BIC criterion and adjusted R^2 yield the same two best fitting models to predict player value as each other. When taking model parsimony into account, we would consider using the second model due to the fact it contains seven predictors, as compared to the first model, which contains eight. This is due to the fact that when attempting to find the most parsimonious model, we try to find the one with the least amount of explanatory variables, if the BIC or adjusted R^2 values are extremely close to one another (which they are in this example).