STAT 420: Group Project Part 4

Group 10

4/21/2020

Introduction

The goal of this project is to find the best model that can predict the percent body fat of an individual based on body measurements.
We have the data of 252 individuals that measures their density, percent body fat, age (in years), weight (in pounds), height (in inches), as well as their neck, chest, abdomen, hip, thigh, knee, ankle, bicep, forearm, and wrist circumference (in cm).

Methods

First, we are going to use forward, backward, and stepwise method to find the model with the smallest AIC and BIC.

```
fat <- read.csv("Bodyfat.csv")

#Backwards
modelALL = lm(bodyfat ~ ., data = fat)
backwardsAIC = step(modelALL, direction = "backward") #Picks model with predictors Density Age and Ches
N = length(resid(modelALL))
backwardsBIC = step(modelALL, direction = "backward", k = log(N)) #Picks Density and Chest #BIC= 133.3
```

The backward AIC method picks the model with Density, Age, and Chest as the model with the smallest AIC (120.84).
The backward BIC method picks the model with Density and Chest as the as the model with the smallest BIC (133.3).

```
#Forward
modelstart = lm(bodyfat ~ 1, data = fat)
forwardAIC = step(modelstart, scope = bodyfat ~ Density + Age + Weight + Height + Neck + Chest + Abdomen
forwardBIC = step(modelstart, scope = bodyfat ~ Density + Age + Weight + Height + Neck + Chest + Abdomen
```

The forward AIC method picks the model with Density, Abdomen, and Age as the model with the smallest AIC (120.04).
The forward BIC method picks the model with Density and Abdomen as the as the model with the smallest BIC (131.99).

```
#Stepwise
stepAIC = step(modelstart, scope = bodyfat ~ Density + Age + Weight + Height + Neck + Chest + Abdomen +
stepBIC = step(modelstart, scope = bodyfat ~ Density + Age + Weight + Height + Neck + Chest + Abdomen +
```

The stepwise AIC method picks the model with Density, Abdomen, and Age as the model with the smallest AIC (120.04).
The stepwise BIC method picks the model with Density and Abdomen as the as the model with the smallest BIC (131.99).

```
#Exhaustive
library(leaps)
all_fat_mod = summary(regsubsets(bodyfat ~ ., data = fat))
p = length(coef(modelALL))
n = length(resid(modelALL))
fat_mod_aic = n * log(all_fat_mod$rss / n) + 2 * (2:p)
best_fat_ind = which.min(fat_mod_aic)
all_fat_mod$which[best_fat_ind,]
```

```
## (Intercept)     Density         Age      Weight      Height        Neck
##        TRUE        TRUE        TRUE       FALSE       FALSE       FALSE
##       Chest     Abdomen         Hip       Thigh        Knee       Ankle
##       FALSE        TRUE       FALSE       FALSE       FALSE       FALSE
##      Biceps     Forearm       Wrist
##       FALSE       FALSE       FALSE
```

```
fat_mod_best_aic = lm(bodyfat ~ Density + Age + Abdomen, data = fat) #AIC = 120.0427

fat_mod_bic = n * log(all_fat_mod$rss / n) + log(n) * (2:p)
best_fat_bic = which.min(fat_mod_bic)
all_fat_mod$which[best_fat_bic,] #Density, Abdomen
```

```
## (Intercept)     Density         Age      Weight      Height        Neck
##        TRUE        TRUE       FALSE       FALSE       FALSE       FALSE
##       Chest     Abdomen         Hip       Thigh        Knee       Ankle
##       FALSE        TRUE       FALSE       FALSE       FALSE       FALSE
##      Biceps     Forearm       Wrist
##       FALSE       FALSE       FALSE
```

```
fat_mod_best_Bic = lm(bodyfat ~ Density + Abdomen, data = fat)
extractAIC(fat_mod_best_Bic, k = log(n)) #BIC = 131.9893
```

```
## [1]   3.0000 131.9893
```

From the methods that we used, we found that these 4 models have the lowest AIC and BIC:

```
#models to consider
Model_DAC = lm(bodyfat ~ Density + Age + Chest, data = fat) #AIC Selection
Model_DC = lm(bodyfat ~ Density + Chest, data = fat) #BIC selection
Model_DAA = lm(bodyfat ~ Density + Age + Abdomen, data = fat) #AIC Selection
Model_DA = lm(bodyfat ~ Density + Abdomen, data = fat) #BIC Selection
```

```
#BIC - picks smaller models
extractAIC(Model_DAC, k = log(n)) #134.9566 last #AIC Selection
```

```
## [1]   4.0000 134.9566
```

```r
extractAIC(Model_DC, k = log(n)) #133.3018 second #BIC Selection
```

```
## [1]   3.0000 133.3018
```

```r
extractAIC(Model_DAA, k = log(n)) #134.1604 third #AIC Selection
```

```
## [1]   4.0000 134.1604
```

```r
extractAIC(Model_DA, k = log(n)) #131.9893 best #BIC Selection
```

```
## [1]   3.0000 131.9893
```

```r
#AIC - picks larger models
extractAIC(Model_DAC) #120.8389 second #AIC Selection
```

```
## [1]   4.0000 120.8389
```

```r
extractAIC(Model_DC) #122.7135 last #BIC Selection
```

```
## [1]   3.0000 122.7135
```

```r
extractAIC(Model_DAA) #120.0427 best #AIC Selection
```

```
## [1]   4.0000 120.0427
```

```r
extractAIC(Model_DA) #121.401 third #BIC Selection
```

```
## [1]   3.000 121.401
```

Now, we are going to compare the RMSE of these 4 models.

```
##      [,1]              [,2]                [,3]              [,4]
## [1,] "RMSE_DAC"        "RMSE_DAA"          "RMSE_DA"         "RMSE_DC"
## [2,] "1.2509293329801" "1.24895473652963" "1.2573046485178" "1.2605831014381"
```

Using the ANOVA test, we can see that Age is not significant at 0.05 but significant ar 0.07.

```r
#anova
anova(Model_DA,Model_DAA) #not significant at 5 percent - significant at 7
```

```
## Analysis of Variance Table
##
## Model 1: bodyfat ~ Density + Abdomen
## Model 2: bodyfat ~ Density + Age + Abdomen
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    249 398.37
## 2    248 393.09  1    5.2736 3.3271 0.06935 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, let's check whether or not our models meet the Normality and Constant Variance Assumptions. Normality Assumption Test

```
## [1] 1.223469e-28
```

```
## [1] 6.339676e-29
```

Constant Variance Assumption Test

```
bptest(Model_DAA)$p.value
```

```
##           BP
## 0.0001323568
```

```
bptest(Model_DA)$p.value
```

```
##           BP
## 7.293614e-05
```

In appears that both models don't meet the Normality and Constant Variance assumptions (p-value is very small).

Results

Model_DAA RSE = 1.259 RMSE = 1.248955 R^2 = .9776 Adjusted R^2 = .9774

Model_DAC RSE = 1.261 RMSE = 1.250929 R^2 = .9776 Adjusted R^2 = .9773

Model_DC RSE = 1.268 RMSE = 1.260583 R^2 = .9772 Adjusted R^2 = .9770

Model_DA RSE = 1.265 RMSE = 1.257305 R^2 = .9773 Adjusted R^2 = .9772

Given the intent of backward, forward, stepwise, and exhaustive selection procedures seek to find models with the smallest respective AIC and BIC values, we will omit the higher AIC and BIC models.

This leaves us with `Model_DAA` (AIC) and `Model_DA`(BIC) to consider. In terms of testing, both models don't meet the Normality and Constant Variance assumption making them inadequate in being explanatory. In terms of anova testing, there is not a significant difference between models at a .05 significance level but there is a significant difference at .10. In terms of T tests for individual predictors of a model, the predictor Age in `Model_DAA` would reject the null (proving linearity) at a .10 significance but accepts the null at .05. Based off the rather dynamic nature of test results from change of significance, we will base our final decision on measures of error and variance of each model.

Conclusion

The best model to predict is `Model_DAA`. It has less error associated with it due to lower RSE and RMSE values It also has higher R^2 and adjusted R^2 values than `Model_DA`, meaning 97.76% of variance observed in the explanatory variable of selected model is described by the model.