Homework 7

Anahi Rodriguez

Due 10/26/2022

Homework Instructions

Make sure to add your name to the header of the document. When submitting the assignment on Gradescope, be sure to assign the appropriate pages of your submission to each Exercise.

The point value for each exercise is noted in the exercise title.

For questions that require code, please create or use the code chunk directly below the question and type your code there. Your knitted pdf will then show both the code and the output, so that we can assess your understanding and award any partial credit.

For written questions, please provide your answer after the indicated *Answer* prompt.

You are encouraged to knit your file as you work, to check that your coding and formatting are done so appropriately. This will also help you identify and locate any errors more easily.

Homework Setup

We'll use the following packages for this homework assignment. We'll also read in data from a csv file. To access the data, you'll want to download the dataset from Canvas, and place it in the same folder as this R Markdown document. You'll then be able to use the following code to load in the data.

```
library(ggplot2)
library(MASS)
```

Exercise 1: Hockey Goalies [20 points]

We will use the data stored in `goalies.csv`, which contains career data for 462 players in the National Hockey League who played goaltender at some point up to and including the 2014-2015 season. The variables in this dataset are:

- `W` - Wins
- `GA` - Goals Against
- `SA` - Shots Against
- `SV` - Saves
- `SV_PCT` - Save Percentage
- `GAA` - Goals Against Average
- `SO` - Shutouts
- `MIN` - Minutes
- `PIM` - Penalties in Minutes

part a

Read in the data. Then fit the following multiple linear regression model in `R`. Save the model to a name and run a summary of the model.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

.

Here,

- Y_i is `W` (Wins)
- x_{i1} is `GAA` (Goals Against Average)
- x_{i2} is `SV_PCT` (Save Percentage)
- x_{i3} is `MIN` (Minutes)

```
# Use this code chunk for your answer.
setwd("~/Desktop/data")
goalies = read.csv("goalies.csv")

lm1 = lm(W ~ GAA + SV_PCT + MIN, data = goalies)
summary(lm1)
```

```
##
## Call:
## lm(formula = W ~ GAA + SV_PCT + MIN, data = goalies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.527  -4.948   1.923   4.831  98.938
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.358e+01  2.206e+01  -0.616    0.538
## GAA         -5.822e-01  6.384e-01  -0.912    0.362
## SV_PCT       1.269e+01  2.313e+01   0.549    0.584
## MIN          7.998e-03  6.113e-05 130.844   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.67 on 458 degrees of freedom
## Multiple R-squared:  0.9746, Adjusted R-squared:  0.9744
## F-statistic:  5855 on 3 and 458 DF,  p-value: < 2.2e-16
```

part b

Use an F-test to test the significance of the regression.

Report the following:

- The null and alternative hypotheses
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.01$

```
# Use this code chunk for you answer, as needed.
intercept_model = lm(W ~ 1, data = goalies)
anova(intercept_model, lm1)
```

```
## Analysis of Variance Table
##
## Model 1: W ~ 1
## Model 2: W ~ GAA + SV_PCT + MIN
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    461 5008654
## 2    458  127285  3   4881368 5854.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: H_0 : $\beta_{GGA} = \beta_{SVPCT} = \beta_{MIN} = 0$ for the model wins $= \beta_0 + \beta_{GGA} + \beta_{SVPCT} + \beta_{MIN}$

H_a : at least one of β_{GGA}, β_{SVPCT}, or β_{MIN} does not equal 0 for the model mentioned in null

Test statistic: 5854.7

P-value: <2.2 * 10^-16

I would reject the null at the 1% level because there is sufficient evidence to suggest that at least one of the slope coefficients is not equal to 0 as (2.2 * 10^-16) < 0.01

part c

Consider this statement: "Since the F-test result gives a very low p-value, then we can conclude that knowing the goals against average, save percentage, and minutes of an NHL goalie allows you to make a highly accurate prediction of that goalie's wins." Do you think this is a good conclusion to draw, or not? Explain your answer.

Answer:

I do believe that this is a good conclusion because from our anova test, we know that there is sufficient evidence that at least some of these variables are important for making a prediction, so we choose the larger model over the intercept only model. Additionally, the multiple R^2 value for the larger model is quite high at 97.46%. Although it seems that none of these variables is statistically significant on their own from the p-value associated with the t-test for each slope coefficient, collectively, these variables are good predictors.

part d

Use your model to predict the number of `Wins` for famous NHL goalie Tony Esposito, who has 2.93 Goals Against Average, 0.906 Save Percentage, and 52476 Minutes.

```
# Use this code chunk for your answer.
predict(lm1, data.frame('GAA' = 2.93, 'SV_PCT' = 0.906,
                        'MIN' = 52476))
```

```
##        1
## 415.9203
```

Answer: 415.9203

part e

Point estimates may have some error, so let's instead create an interval for wins that should contain the true wins of a goalie with these stats 90% of the time.

Create (and print) an interval to estimate the wins of a goalie with Tony Esposito's stats with 90% confidence.

```
# Use this code chunk for your answer.
predict(lm1, data.frame('GAA' = 2.93, 'SV_PCT' = 0.906,
                        'MIN' = 52476),
```

```
        interval = 'prediction',
        level = 0.9)
```

```
##        fit      lwr      upr
## 1 415.9203 388.0814 443.7591
```

part f

Calculate the standard deviation s_y for the observed values of the Wins variable. Report the value of s_e from your multiple regression model.

Briefly interpret what each measure represents.

Do these two measures together communicate anything about the strength of this model? *Hint: think about how each of these values is related to our SS terms from the semester.*

```
# Use this code chunk for your answer.
s_y = sd(goalies$W)
s_y
```

```
## [1] 104.2342
```

```
sst = (s_y)^2 * (dim(goalies)[1]) # 463 = 464 - 1

summary(lm1)
```

```
##
## Call:
## lm(formula = W ~ GAA + SV_PCT + MIN, data = goalies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.527  -4.948   1.923   4.831  98.938
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.358e+01  2.206e+01  -0.616    0.538
## GAA         -5.822e-01  6.384e-01  -0.912    0.362
## SV_PCT       1.269e+01  2.313e+01   0.549    0.584
## MIN          7.998e-03  6.113e-05 130.844   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.67 on 458 degrees of freedom
## Multiple R-squared:  0.9746, Adjusted R-squared:  0.9744
## F-statistic:  5855 on 3 and 458 DF,  p-value: < 2.2e-16
```

```
s_e = 16.67
s_e
```

```
## [1] 16.67
```

```
mse = s_e^2
mse
```

```
## [1] 277.8889
```

```
sse = mse * 458 # 458 = n - p
sse
```

```
## [1] 127273.1
```

```
1- sse/sst
```

```
## [1] 0.9746444
```

Answer: The standard deviation of y tells us that typically, people will win between 104.2342 more or less games than the games won average value and the standard error of the regression itself is a estimated measure of the amount of standard deviation in our true errors. These two measures can be used to calculate the values of sst and sse which we can then use to calculate R^2 which attests to the models strength.

Exercise 2: Hockey Goalies, Testing [15 points]

We will consider four models, each with Wins as the response. The predictors for these models are:

- Model 1: Goals Against, Saves
- Model 2: Shots Against, Minutes, Shutouts
- Model 3: Goals Against, Saves, Shots Against, Minutes, Shutouts
- Model 4: All Available Variables

part a

An F-test allows us to compare two models. An F-test will not provide interpretable results for one set of two models. Which set is it?

Answer: Model 1 and Model 2 becuase neither is a nested model of the other

part b

Use an F-test to compare Models 2 and 3. Report the following:

- The null hypothesis (you can write this in words or symbols)
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.01$
- Your model preference (given this test result).

```
# Use this code chunk for your answer.
model_2 = lm(data = goalies,
             W ~ SA + MIN + SO)
model_3 = lm(data = goalies,
             W ~ SA + MIN + SO + SV + GA)

anova(model_2, model_3)
```

```
## Analysis of Variance Table
##
## Model 1: W ~ SA + MIN + SO
## Model 2: W ~ SA + MIN + SO + SV + GA
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    458 84129
## 2    456 72899  2     11230 35.124 6.496e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: $H_0 : \beta_{SV} = \beta_{GA} = 0$ for the model wins $= \beta_0 + \beta_{SA} + \beta_{MIN} + \beta_{SO} + \beta_{SV} + \beta_{GA}$

H_a : at least one of β_{SV} or β_{GA} does not equal 0 for the model mentioned in null

test statistic: 35.124

p-value: 6.496 * 10^-15

I would reject the null at the 1% level because the p-value is less than 0.01. I would prefer the larger model, model 3, given this test result.

part c

Use a t-test to test if the variable Minutes (`MIN`) has a linear relationship with Wins after accounting for all other predictors in the dataset. In other words, test $H_0 : \beta_{`MIN`} = 0$ vs. $H_1 : \beta_{`MIN`} \neq 0$ for a specific model (which model is it?). Report the following:

- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.05$

```
# Use this code chunk for your answer.
model_t_test = lm(data = goalies,
                  W ~ GA + SA + SV + SV_PCT + GAA + SO + MIN + PIM)
summary(model_t_test)
```

```
##
## Call:
## lm(formula = W ~ GA + SA + SV + SV_PCT + GAA + SO + MIN + PIM,
##     data = goalies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.204  -3.126   0.935   2.835  64.078
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.2651619 16.8181423   0.313 0.754376
## GA          -0.1132805  0.0148085  -7.650 1.22e-13 ***
## SA           0.0516385  0.0135565   3.809 0.000159 ***
## SV          -0.0582151  0.0150905  -3.858 0.000131 ***
## SV_PCT      -8.0475191 17.6600154  -0.456 0.648830
## GAA         -0.0496006  0.4821957  -0.103 0.918116
## SO           0.4599359  0.1989567   2.312 0.021240 *
## MIN          0.0131790  0.0009504  13.867  < 2e-16 ***
## PIM          0.0468422  0.0136373   3.435 0.000647 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.52 on 453 degrees of freedom
## Multiple R-squared:  0.9858, Adjusted R-squared:  0.9856
## F-statistic:  3938 on 8 and 453 DF,  p-value: < 2.2e-16
```

Answer: test stat: 13.867 p-value: $<$ 2e-16 I would reject the null at the 1% level because "$<$ 2e-16" is less than 0.01 and conclude that there is enough evidence to suggest that the minute variable is useful to predict the number of wins

---

Exercise 3: Model Selection by Hand [10 points]

Using the goalies dataset, we'll perform model selection by hand. We would like to choose a model to predict the number of wins from the other variables in the dataset.

part a

We'll perform model selection in this exercise "by hand". That means you should not use the `step` function in `R` for this exercise; if you do, you will not receive credit. We will use a backward searching process and will use the coefficient p-values to determine which variables to remove from the model, with an α of **0.01**.

Show the starting model and any subsequent models fit during your searching process here.

```
# Use this code chunk for your answer.
summary(model_t_test)
```

```
##
## Call:
## lm(formula = W ~ GA + SA + SV + SV_PCT + GAA + SO + MIN + PIM,
##     data = goalies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.204  -3.126   0.935   2.835  64.078
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.2651619 16.8181423   0.313 0.754376
## GA          -0.1132805  0.0148085  -7.650 1.22e-13 ***
## SA           0.0516385  0.0135565   3.809 0.000159 ***
## SV          -0.0582151  0.0150905  -3.858 0.000131 ***
## SV_PCT      -8.0475191 17.6600154  -0.456 0.648830
## GAA         -0.0496006  0.4821957  -0.103 0.918116
## SO           0.4599359  0.1989567   2.312 0.021240 *
## MIN          0.0131790  0.0009504  13.867  < 2e-16 ***
## PIM          0.0468422  0.0136373   3.435 0.000647 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.52 on 453 degrees of freedom
## Multiple R-squared:  0.9858, Adjusted R-squared:  0.9856
## F-statistic:  3938 on 8 and 453 DF,  p-value: < 2.2e-16
```

```
step1 = lm(data = goalies,
           W ~ GA + SA + SV + SO + MIN + PIM + SV_PCT)
summary(step1)
```

```
##
## Call:
## lm(formula = W ~ GA + SA + SV + SO + MIN + PIM + SV_PCT, data = goalies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.201  -3.110   0.936   2.796  64.078
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.958596  11.011238   0.360 0.719384
## GA          -0.113379   0.014761  -7.681  9.8e-14 ***
## SA           0.051681   0.013535   3.818 0.000153 ***
## SV          -0.058266   0.015066  -3.867 0.000126 ***
## SO           0.459474   0.198689   2.313 0.021195 *
## MIN          0.013186   0.000947  13.924  < 2e-16 ***
## PIM          0.046831   0.013622   3.438 0.000640 ***
## SV_PCT      -6.759465  12.439442  -0.543 0.587128
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.51 on 454 degrees of freedom
## Multiple R-squared:  0.9858, Adjusted R-squared:  0.9856
## F-statistic:  4511 on 7 and 454 DF,  p-value: < 2.2e-16
```

```
step2 = lm(data = goalies,
           W ~ GA + SA + SV + SO + MIN + PIM)
summary(step2)
```

```
##
## Call:
## lm(formula = W ~ GA + SA + SV + SO + MIN + PIM, data = goalies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.206  -3.067   1.187   2.696  64.059
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.0111736  0.7420675  -2.710 0.006978 **
## GA          -0.1129541  0.0147289  -7.669 1.06e-13 ***
## SA           0.0520814  0.0135049   3.856 0.000132 ***
## SV          -0.0587246  0.0150306  -3.907 0.000108 ***
## SO           0.4655961  0.1982159   2.349 0.019254 *
## MIN          0.0131616  0.0009452  13.925  < 2e-16 ***
## PIM          0.0469398  0.0136100   3.449 0.000615 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.5 on 455 degrees of freedom
## Multiple R-squared:  0.9858, Adjusted R-squared:  0.9856
## F-statistic:  5271 on 6 and 455 DF,  p-value: < 2.2e-16
```

```
step3 = lm(data = goalies,
           W ~ GA + SA + SV + MIN + PIM)
summary(step3)
```

```
##
## Call:
## lm(formula = W ~ GA + SA + SV + MIN + PIM, data = goalies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.922  -3.546   1.294   2.737  63.656
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.0198636  0.7457250  -2.709 0.007011 **
## GA          -0.1359994  0.0110400 -12.319  < 2e-16 ***
## SA           0.0512308  0.0135668   3.776 0.000180 ***
## SV          -0.0581577  0.0151029  -3.851 0.000135 ***
## MIN          0.0148741  0.0006045  24.607  < 2e-16 ***
## PIM          0.0426871  0.0135557   3.149 0.001746 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.56 on 456 degrees of freedom
## Multiple R-squared:  0.9856, Adjusted R-squared:  0.9855
## F-statistic:  6262 on 5 and 456 DF,  p-value: < 2.2e-16
```

part b

Report the predictor variables included in your selected model from part a.

Answer: GA, SA, SV, MIN, PIM

part c

Report the fitted model for your selected model.

Answer: estimated wins = -2.0198636 - (0.1359994 * GA) + (0.0512308 * SA) - (0.0581577 * SV) + (0.0148741 * MIN) + (0.0426871 + PIM)

Exercise 4: Chick-fil-A Searching Methods [25 points]

For this exercise, we'll analyze the nutritional value of menu items from Chick-fil-A, a fast food restaurant specializing in chicken sandwiches. This data is contained in the chickfila.csv file on Canvas.

We'll be interested in fitting a model to predict the Calories in a menu item from the other nutritional characteristics of that menu item.

part a

Read in the chickfila.csv data file. How many models predicting the number of Calories in a menu item are possible from this dataset? (Consider only first-order terms, which means include all of the variables once and exactly as they appear in the dataset.)

```
# Use this code chunk for your answer, as needed.
setwd("~/Desktop/data")
cfa = read.csv("chickfila.csv")

2^10
```

```
## [1] 1024
```

Answer: 1024

part b

Perform model selection, using BIC as the metric and backward searching.

Report the predictor variables selected for the final model. No need to report the fitted coefficients.

```
# Use this code chunk for your answer.

start_model = lm(Calories ~ Fat + SatFat + TransFat + Cholesterol +
                   Sodium + Carbs + Fiber + Sugar + Protein + Serving,
                 data = cfa)

step(data = cfa,
     object = start_model ,
     direction = 'backward', k = log(290))
```

```
## Start:  AIC=1465.63
## Calories ~ Fat + SatFat + TransFat + Cholesterol + Sodium + Carbs +
##     Fiber + Sugar + Protein + Serving
##
##               Df Sum of Sq     RSS    AIC
## - TransFat     1        34   36667 1460.2
## - Cholesterol  1       207   36841 1461.6
## - Sugar        1       293   36926 1462.3
## <none>                       36634 1465.6
## - Sodium       1       954   37588 1467.4
## - SatFat       1      1630   38263 1472.6
## - Serving      1      2241   38875 1477.2
## - Fiber        1      2904   39538 1482.1
## - Protein      1    173882  210515 1967.0
## - Carbs        1    465541  502175 2219.2
## - Fat          1   2334536 2371170 2669.3
##
## Step:  AIC=1460.23
## Calories ~ Fat + SatFat + Cholesterol + Sodium + Carbs + Fiber +
##     Sugar + Protein + Serving
##
##               Df Sum of Sq     RSS    AIC
## - Cholesterol  1       201   36868 1456.1
## - Sugar        1       357   37024 1457.4
## <none>                       36667 1460.2
## - Sodium       1       935   37602 1461.9
## - SatFat       1      2017   38684 1470.1
## - Serving      1      2262   38929 1471.9
## - Fiber        1      2871   39538 1476.4
## - Protein      1    180272  216940 1970.1
## - Carbs        1    492671  529339 2228.8
## - Fat          1   2380741 2417409 2669.2
##
## Step:  AIC=1456.14
## Calories ~ Fat + SatFat + Sodium + Carbs + Fiber + Sugar + Protein +
##     Serving
##
##           Df Sum of Sq     RSS    AIC
## - Sugar    1       325   37193 1453.0
## <none>                   36868 1456.1
```

```
## - Sodium   1       808   37676 1456.8
## - SatFat   1      1982   38850 1465.7
## - Serving  1      2301   39169 1468.0
## - Fiber    1      2694   39562 1470.9
## - Protein  1    233337  270205 2028.1
## - Carbs    1    493324  530192 2223.6
## - Fat      1   2562021 2598889 2684.6
##
## Step:  AIC=1453.02
## Calories ~ Fat + SatFat + Sodium + Carbs + Fiber + Protein +
##     Serving
##
##           Df Sum of Sq     RSS    AIC
## <none>                   37193 1453.0
## - Sodium   1      1131   38324 1456.0
## - Serving  1      2850   40043 1468.8
## - SatFat   1      3579   40772 1474.0
## - Fiber    1      4934   42127 1483.5
## - Protein  1    237877  275070 2027.6
## - Fat      1   2677776 2714969 2691.6
## - Carbs    1   8554823 8592016 3025.7
##
## Call:
## lm(formula = Calories ~ Fat + SatFat + Sodium + Carbs + Fiber +
##     Protein + Serving, data = cfa)
##
## Coefficients:
## (Intercept)          Fat       SatFat       Sodium        Carbs        Fiber
##    2.002934     8.664515     0.580707     0.006094     3.799963     0.857368
##     Protein      Serving
##    3.872489    -0.006290
```

Answer: Fat, SatFat, Sodium, Carbs, Fiber, Protein, Serving

part c

Perform model selection, using BIC as the metric and forward searching.

Report the predictor variables selected for the model after the first step and for the final model. No need to report the fitted coefficients.

```
# Use this code chunk for your answer.
step(data = cfa,
     object = lm(Calories ~ 1, data = cfa),
     scope = Calories ~ Fat  + SatFat + TransFat + Cholesterol +
             Sodium + Carbs + Fiber + Sugar + Protein + Serving,
     direction = 'forward',
     k = log(290))
```

```
## Start:  AIC=3856.55
## Calories ~ 1
##
##               Df Sum of Sq       RSS    AIC
## + Fat          1 146048680  23521053 3289.4
## + SatFat       1 144469314  25100419 3308.2
```

```
## + Sodium       1 137298325  32271408 3381.1
## + Protein      1 130049765  39519968 3439.8
## + TransFat     1 127307413  42262320 3459.3
## + Carbs        1  96177539  73392194 3619.4
## + Cholesterol  1  94557536  75012197 3625.7
## + Fiber        1  86783514  82786219 3654.3
## + Serving      1  50847203 118722530 3758.8
## + Sugar        1  21764608 147805125 3822.4
## <none>                     169569733 3856.5
##
## Step:  AIC=3289.36
## Calories ~ Fat
##
##               Df Sum of Sq      RSS    AIC
## + Carbs        1  22510439  1010614 2382.3
## + Sugar        1  20202810  3318243 2727.1
## + Serving      1  13606617  9914436 3044.5
## + Fiber        1   2670597 20850456 3260.1
## + SatFat       1   1911124 21609929 3270.5
## + Protein      1    695114 22825939 3286.3
## + Sodium       1    652095 22868958 3286.9
## + TransFat     1    515617 23005436 3288.6
## <none>                     23521053 3289.4
## + Cholesterol  1     11977 23509076 3294.9
##
## Step:  AIC=2382.3
## Calories ~ Fat + Carbs
##
##               Df Sum of Sq     RSS    AIC
## + Protein      1    961520   49093 1510.8
## + Sodium       1    699414  311199 2046.4
## + Sugar        1    353311  657302 2263.2
## + Cholesterol  1    190365  820249 2327.4
## + SatFat       1    148774  861839 2341.8
## + Fiber        1     30032  980581 2379.2
## <none>                     1010614 2382.3
## + TransFat     1      6510 1004104 2386.1
## + Serving      1       405 1010209 2387.9
##
## Step:  AIC=1510.84
## Calories ~ Fat + Carbs + Protein
##
##               Df Sum of Sq   RSS    AIC
## + Sugar        1    6704.3 42389 1473.9
## + Fiber        1    3792.2 45301 1493.2
## + SatFat       1    3665.2 45428 1494.0
## + Serving      1    2249.8 46843 1502.9
## + Sodium       1    1170.3 47923 1509.5
## <none>                     49093 1510.8
## + TransFat     1     465.3 48628 1513.8
## + Cholesterol  1     154.8 48939 1515.6
##
## Step:  AIC=1473.93
## Calories ~ Fat + Carbs + Protein + Sugar
```

```
##
##               Df Sum of Sq   RSS    AIC
## + Serving      1   1324.19 41065 1470.4
## + Fiber        1   1069.50 41319 1472.2
## + SatFat       1    853.52 41535 1473.7
## <none>                     42389 1473.9
## + Sodium       1    291.89 42097 1477.6
## + TransFat     1    256.37 42133 1477.8
## + Cholesterol  1     20.31 42369 1479.5
##
## Step:  AIC=1470.4
## Calories ~ Fat + Carbs + Protein + Sugar + Serving
##
##               Df Sum of Sq   RSS    AIC
## + SatFat       1   1215.61 39849 1467.3
## + Fiber        1   1132.19 39933 1468.0
## <none>                     41065 1470.4
## + Sodium       1    513.40 40551 1472.4
## + TransFat     1    425.02 40640 1473.0
## + Cholesterol  1      4.60 41060 1476.0
##
## Step:  AIC=1467.35
## Calories ~ Fat + Carbs + Protein + Sugar + Serving + SatFat
##
##               Df Sum of Sq   RSS    AIC
## + Fiber        1   2172.80 37676 1456.8
## <none>                     39849 1467.3
## + Sodium       1    286.74 39562 1470.9
## + Cholesterol  1      5.53 39844 1473.0
## + TransFat     1      1.06 39848 1473.0
##
## Step:  AIC=1456.76
## Calories ~ Fat + Carbs + Protein + Sugar + Serving + SatFat +
##     Fiber
##
##               Df Sum of Sq   RSS    AIC
## + Sodium       1    808.28 36868 1456.1
## <none>                     37676 1456.8
## + Cholesterol  1     73.97 37602 1461.9
## + TransFat     1     12.54 37664 1462.3
##
## Step:  AIC=1456.14
## Calories ~ Fat + Carbs + Protein + Sugar + Serving + SatFat +
##     Fiber + Sodium
##
##               Df Sum of Sq   RSS    AIC
## <none>                     36868 1456.1
## + Cholesterol  1   200.785 36667 1460.2
## + TransFat     1    27.329 36841 1461.6
##
## Call:
## lm(formula = Calories ~ Fat + Carbs + Protein + Sugar + Serving +
##     SatFat + Fiber + Sodium, data = cfa)
```

```
##
## Coefficients:
## (Intercept)          Fat        Carbs      Protein        Sugar      Serving
##    1.712288     8.685929     3.897188     3.857196    -0.108422    -0.005800
##      SatFat        Fiber       Sodium
##    0.488492     0.730861     0.005302
```

Answer: After the first step, the added predictor is Fat and after the last step, the predictors are Fat, Carbs, Protein, Sugar, Serving, SatFat, Fiber, and Sodium

part d

Perform model selection, using BIC as the metric and stepwise searching.

Report the predictor variables selected for the final model. No need to report the fitted coefficients. Do you select the same models using backward, forward, and stepwise searching?

```
# Use this code chunk for your answer.
step(data = cfa,
     object = lm(Calories ~ 1, data = cfa),
     scope = Calories ~ Fat  + SatFat + TransFat + Cholesterol +
            Sodium + Carbs + Fiber + Sugar + Protein + Serving,
     direction = 'both',
     k = log(290))
```

```
## Start:  AIC=3856.55
## Calories ~ 1
##
##               Df Sum of Sq       RSS    AIC
## + Fat          1 146048680  23521053 3289.4
## + SatFat       1 144469314  25100419 3308.2
## + Sodium       1 137298325  32271408 3381.1
## + Protein      1 130049765  39519968 3439.8
## + TransFat     1 127307413  42262320 3459.3
## + Carbs        1  96177539  73392194 3619.4
## + Cholesterol  1  94557536  75012197 3625.7
## + Fiber        1  86783514  82786219 3654.3
## + Serving      1  50847203 118722530 3758.8
## + Sugar        1  21764608 147805125 3822.4
## <none>                     169569733 3856.5
##
## Step:  AIC=3289.36
## Calories ~ Fat
##
##               Df Sum of Sq       RSS    AIC
## + Carbs        1  22510439   1010614 2382.3
## + Sugar        1  20202810   3318243 2727.1
## + Serving      1  13606617   9914436 3044.5
## + Fiber        1   2670597  20850456 3260.1
## + SatFat       1   1911124  21609929 3270.5
## + Protein      1    695114  22825939 3286.3
## + Sodium       1    652095  22868958 3286.9
## + TransFat     1    515617  23005436 3288.6
## <none>                     23521053 3289.4
## + Cholesterol  1     11977  23509076 3294.9
## - Fat          1 146048680 169569733 3856.5
```

```
##
## Step:  AIC=2382.3
## Calories ~ Fat + Carbs
##
##               Df Sum of Sq      RSS    AIC
## + Protein      1    961520    49093 1510.8
## + Sodium       1    699414   311199 2046.4
## + Sugar        1    353311   657302 2263.2
## + Cholesterol  1    190365   820249 2327.4
## + SatFat       1    148774   861839 2341.8
## + Fiber        1     30032   980581 2379.2
## <none>                      1010614 2382.3
## + TransFat     1      6510  1004104 2386.1
## + Serving      1       405  1010209 2387.9
## - Carbs        1  22510439 23521053 3289.4
## - Fat          1  72381581 73392194 3619.4
##
## Step:  AIC=1510.84
## Calories ~ Fat + Carbs + Protein
##
##               Df Sum of Sq      RSS    AIC
## + Sugar        1      6704    42389 1473.9
## + Fiber        1      3792    45301 1493.2
## + SatFat       1      3665    45428 1494.0
## + Serving      1      2250    46843 1502.9
## + Sodium       1      1170    47923 1509.5
## <none>                        49093 1510.8
## + TransFat     1       465    48628 1513.8
## + Cholesterol  1       155    48939 1515.6
## - Protein      1    961520  1010614 2382.3
## - Fat          1   8224852  8273945 2992.0
## - Carbs        1  22776846 22825939 3286.3
##
## Step:  AIC=1473.93
## Calories ~ Fat + Carbs + Protein + Sugar
##
##               Df Sum of Sq     RSS    AIC
## + Serving      1      1324   41065 1470.4
## + Fiber        1      1070   41319 1472.2
## + SatFat       1       854   41535 1473.7
## <none>                       42389 1473.9
## + Sodium       1       292   42097 1477.6
## + TransFat     1       256   42133 1477.8
## + Cholesterol  1        20   42369 1479.5
## - Sugar        1      6704   49093 1510.8
## - Protein      1    614913  657302 2263.2
## - Carbs        1    967691 1010080 2387.8
## - Fat          1   6582934 6625323 2933.3
##
## Step:  AIC=1470.4
## Calories ~ Fat + Carbs + Protein + Sugar + Serving
##
##               Df Sum of Sq     RSS    AIC
## + SatFat       1      1216   39849 1467.3
```

```
## + Fiber        1      1132   39933 1468.0
## <none>                       41065 1470.4
## + Sodium       1       513   40551 1472.4
## + TransFat     1       425   40640 1473.0
## - Serving      1      1324   42389 1473.9
## + Cholesterol  1         5   41060 1476.0
## - Sugar        1      5779   46843 1502.9
## - Protein      1    611594  652659 2266.8
## - Carbs        1    965556 1006621 2392.5
## - Fat          1   6557438 6598503 2937.8
##
## Step:  AIC=1467.35
## Calories ~ Fat + Carbs + Protein + Sugar + Serving + SatFat
##
##               Df Sum of Sq     RSS    AIC
## + Fiber        1      2173   37676 1456.8
## <none>                       39849 1467.3
## - SatFat       1      1216   41065 1470.4
## + Sodium       1       287   39562 1470.9
## + Cholesterol  1         6   39844 1473.0
## + TransFat     1         1   39848 1473.0
## - Serving      1      1686   41535 1473.7
## - Sugar        1      2876   42725 1481.9
## - Protein      1    605272  645122 2269.1
## - Carbs        1    754075  793925 2329.3
## - Fat          1   3200272 3240121 2737.2
##
## Step:  AIC=1456.76
## Calories ~ Fat + Carbs + Protein + Sugar + Serving + SatFat +
##     Fiber
##
##               Df Sum of Sq     RSS    AIC
## - Sugar        1       647   38324 1456.0
## + Sodium       1       808   36868 1456.1
## <none>                       37676 1456.8
## + Cholesterol  1        74   37602 1461.9
## + TransFat     1        13   37664 1462.3
## - Serving      1      1969   39645 1465.9
## - Fiber        1      2173   39849 1467.3
## - SatFat       1      2256   39933 1468.0
## - Carbs        1    526584  564261 2236.0
## - Protein      1    606297  643973 2274.3
## - Fat          1   2977463 3015139 2722.0
##
## Step:  AIC=1456.03
## Calories ~ Fat + Carbs + Protein + Serving + SatFat + Fiber
##
##               Df Sum of Sq     RSS    AIC
## + Sodium       1      1131   37193 1453.0
## <none>                       38324 1456.0
## + Sugar        1       647   37676 1456.8
## + TransFat     1        82   38242 1461.1
## + Cholesterol  1        33   38291 1461.5
## - Serving      1      2544   40868 1469.0
```

```
## - Fiber        1      4401   42725 1481.9
## - SatFat       1      4897   43221 1485.2
## - Protein      1    786333  824657 2340.3
## - Fat          1   3031000 3069323 2721.5
## - Carbs        1   8572105 8610429 3020.6
##
## Step:  AIC=1453.02
## Calories ~ Fat + Carbs + Protein + Serving + SatFat + Fiber +
##     Sodium
##
##               Df Sum of Sq     RSS    AIC
## <none>                       37193 1453.0
## - Sodium       1      1131   38324 1456.0
## + Sugar        1       325   36868 1456.1
## + Cholesterol  1       169   37024 1457.4
## + TransFat     1        84   37109 1458.0
## - Serving      1      2850   40043 1468.8
## - SatFat       1      3579   40772 1474.0
## - Fiber        1      4934   42127 1483.5
## - Protein      1    237877  275070 2027.6
## - Fat          1   2677776 2714969 2691.6
## - Carbs        1   8554823 8592016 3025.7
```

```
##
## Call:
## lm(formula = Calories ~ Fat + Carbs + Protein + Serving + SatFat +
##     Fiber + Sodium, data = cfa)
##
## Coefficients:
## (Intercept)          Fat        Carbs      Protein      Serving       SatFat
##    2.002934     8.664515     3.799963     3.872489    -0.006290     0.580707
##       Fiber       Sodium
##    0.857368     0.006094
```

Answer: The final model contains the following predictor variables: Fat, Carbs, Protein, Serving, SatFat, Fiber, and Sodium. The stepwise search and the backward search result in the same model whereas the forward search results in a model with one additional predictor variable, Sugar

part e

Report the BIC for the final model(s) selected with the three searching methods. Based on the BIC, which model would you select overall?

```
# Use this code chunk for your answer, if needed.
```

Answer: backward: 1453.02, forward: 1456.14, stepwise: 1453.02,

I would choose the backward/stepwise model as they are the same with the overall lower BIC value compared to the forward search final model

Exercise 5: Comparing Chick-Fil-A Model Metrics [25 points]

For this exercise, we'll continue analyzing the chickfila dataset but now using an exhaustive searching method to identify our optimal model.

part a

First, run the exhaustive searching function. What variables are included in the optimal model with 3 predictor variables? What metric is used to determine the optimal model at each p? Do the optimal models at each p result in nested models for the chickfila data?

```
# Use this code chunk for your ansswer.
library(leaps)
all_calories_model = summary(regsubsets(Calories ~ Fat + SatFat + TransFat + Cholesterol + Sodium
                    + Carbs + Fiber + Sugar + Protein + Serving, data = cfa))
all_calories_model
```

```
## Subset selection object
## Call: regsubsets.formula(Calories ~ Fat + SatFat + TransFat + Cholesterol +
##     Sodium + Carbs + Fiber + Sugar + Protein + Serving, data = cfa)
## 10 Variables  (and intercept)
##             Forced in Forced out
## Fat             FALSE      FALSE
## SatFat          FALSE      FALSE
## TransFat        FALSE      FALSE
## Cholesterol     FALSE      FALSE
## Sodium          FALSE      FALSE
## Carbs           FALSE      FALSE
## Fiber           FALSE      FALSE
## Sugar           FALSE      FALSE
## Protein         FALSE      FALSE
## Serving         FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          Fat SatFat TransFat Cholesterol Sodium Carbs Fiber Sugar Protein
## 1  ( 1 ) "*" " "    " "      " "         " "    " "   " "   " "   " "
## 2  ( 1 ) "*" " "    " "      " "         " "    "*"   " "   " "   " "
## 3  ( 1 ) "*" " "    " "      " "         " "    "*"   " "   " "   "*"
## 4  ( 1 ) "*" " "    " "      " "         " "    "*"   " "   "*"   "*"
## 5  ( 1 ) "*" "*"    " "      " "         " "    "*"   "*"   " "   "*"
## 6  ( 1 ) "*" "*"    " "      " "         " "    "*"   "*"   " "   "*"
## 7  ( 1 ) "*" "*"    " "      " "         "*"    "*"   "*"   " "   "*"
## 8  ( 1 ) "*" "*"    " "      " "         "*"    "*"   "*"   "*"   "*"
##          Serving
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
```

Answer: The optimal model with 3 predictors contains Fat, Carbs, and protein

R^2 is used to determine between models

and yes these could be nested models of a model using all the variables to predict calories

part b

Calculate the AIC for each of the models selected in part a. Based on AIC, which predictor variables should be included in the optimal model?

```
# Use this code chunk for your answer.
model1 = lm(data = cfa,
            Calories ~ Fat)
model2 = lm(data = cfa,
            Calories ~ Fat + Carbs)
model3 = lm(data = cfa,
            Calories ~ Fat + Carbs + Protein)
model4 = lm(data = cfa,
            Calories ~ Fat + Carbs + Sugar + Protein)
model5 = lm(data = cfa,
            Calories ~ Fat + SatFat + Carbs + Fiber + Protein)
model6 = lm(data = cfa,
            Calories ~ Fat + SatFat + Carbs + Fiber + Protein + Serving)
model7 = lm(data = cfa, Calories ~ Fat + SatFat  +
                Sodium + Carbs + Fiber + Protein + Serving)
model8 = lm(data = cfa,
            Calories ~ Fat + SatFat + Sodium + Carbs + Fiber + Sugar + Protein + Serving)


extractAIC(model1)
```

```
## [1]    2.000 3282.022
```

```
extractAIC(model2)
```

```
## [1]    3.000 2371.294
```

```
extractAIC(model3)
```

```
## [1]    4.000 1496.163
```

```
extractAIC(model4)
```

```
## [1]    5.000 1455.581
```

```
extractAIC(model5)
```

```
## [1]    6.000 1446.985
```

```
extractAIC(model6)
```

```
## [1]    7.000 1430.345
```

```
extractAIC(model7)
```

```
## [1]    8.000 1423.658
```

```
extractAIC(model8)
```

```
## [1]    9.000 1423.114
```

Answer: Fat, SatFat, Sodium, Carbs, Fiber, Sugar, Protein, Serving

part c

Calculate the BIC for each of the models selected in part a. Based on BIC, which predictor variables should be included in the optimal model? Does this match any of the models selected in Exercise 4?

```
# Use this code chunk for your answer.
extractAIC(model1, k = log(290))
```

```
## [1]    2.000 3289.362
```

```
extractAIC(model2, k = log(290))
```

```
## [1]    3.000 2382.304
```

```
extractAIC(model3, k = log(290))
```

```
## [1]    4.000 1510.843
```

```
extractAIC(model4, k = log(290))
```

```
## [1]    5.000 1473.931
```

```
extractAIC(model5, k = log(290))
```

```
## [1]    6.000 1469.005
```

```
extractAIC(model6, k = log(290))
```

```
## [1]    7.000 1456.034
```

```
extractAIC(model7, k = log(290))
```

```
## [1]    8.000 1453.017
```

```
extractAIC(model8, k = log(290))
```

```
## [1]    9.000 1456.143
```

Answer: Fat, SatFat, Sodium, Carbs, Fiber, Protein, Serving. This is the same as the stepwise and backward search selected models

part d

Calculate the adjusted R^2 for each of the models selected in part a. Based on the adjusted R^2, which predictor variables should be included in the optimal model?

```
#Use this code chunk for your answer.
all_calories_model$adjr2
```

```
## [1] 0.8608082 0.9939986 0.9997074 0.9997465 0.9997547 0.9997692 0.9997752
## [8] 0.9997764
```

Answer: Fat, SatFat, Sodium, Carbs, Fiber, Sugar, Protein, Serving

part e

Calculate the RMSE for each of the models selected in part a. Based on the RMSE, which predictor variables should be included in the optimal model?

```
# Use this code chunk for your answer.
sqrt((1/290) * all_calories_model$rss)
```

```
## [1] 284.79305  59.03282  13.01105  12.09004  11.87117  11.49571  11.32482
## [8]  11.27526
```

Answer: Fat, SatFat, Sodium, Carbs, Fiber, Sugar, Protein, Serving

part f

Are the same models selected for each of parts b through e? How many different models are selected from the different metrics but with the same exhaustive searching method?

Answer: No, there are two models selected fro parts b through e. Model 8 is selected when using AIC, Adjusted R^2, and RMSE while model 7 is selected when using BIC

part g

For which of the metrics used in parts b through e is the comparison of models unfair? In other words, which metric would you not want to use in this situation?

Answer: RMSE

Exercise 6: Formatting [5 points]

The last five points of the assignment will be earned for properly formatting your final document. Check that you have:

- included your name on the document
- properly assigned pages to exercises on Gradescope
- selected **page 1 (with your name)** and this page for this exercise (Exercise 6)
- all code is printed and readable for each question
- all output is printed
- generated a pdf file