Homework 9

Anahi Rodrigueez

Due 11/7/2022

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(leaps)
library(faraway)
```

Exercise 1: Hospital SUPPORT Data [25 points]

For this exercise, we will use the data stored in `hospital.csv` on Canvas. It contains a random sample of 580 seriously ill hospitalized patients from a famous study called "SUPPORT" (Study to Understand Prognoses Preferences Outcomes and Risks of Treatment). As the name suggests, the purpose of the study was to determine what factors affected or predicted outcomes, such as how long a patient remained in the hospital. The variables in the dataset are:

- `Days` - Day to death or hospital discharge
- `Age` - Age on day of hospital admission
- `Sex` - Female or male
- `Comorbidity` - Patient diagnosed with more than one chronic disease
- `EdYears` - Years of education
- `Education` - Education level; high or low
- `Income` - Income level; high or low
- `Charges` - Hospital charges, in dollars

- `Care` - Level of care required; high or low
- `Race` - Non-white or white
- `Pressure` - Blood pressure, in mmHg
- `Blood` - White blood cell count, in gm/dL
- `Rate` - Heart rate, in bpm

part a

For this part, first fit an additive multiple regression model (that is, one without any interaction terms) predicting `Charges` from `Age`, `EdYears`, `Pressure`, `Days`, and `Care`. Report the summary of this model.

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

lm1 = lm(Charges ~ Age + EdYears + Pressure + Days + Care,
         data = hospital)
summary(lm1)
```

```
##
## Call:
## lm(formula = Charges ~ Age + EdYears + Pressure + Days + Care,
##     data = hospital)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -177021  -26315   -5311    5149  435820
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  44960.9    15090.3   2.979  0.00301 **
## Age           -187.4      147.0  -1.275  0.20293
## EdYears       1637.1      623.3   2.627  0.00886 **
## Pressure      -163.8       88.1  -1.860  0.06341 .
## Days          2132.4      106.7  19.990  < 2e-16 ***
## Carelow     -42472.9     4862.7  -8.734  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54750 on 574 degrees of freedom
## Multiple R-squared:  0.5397, Adjusted R-squared:  0.5357
## F-statistic: 134.6 on 5 and 574 DF,  p-value: < 2.2e-16
```

```
summary(lm1$residuals)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -177021  -26315   -5311       0    5149  435820
```

part b

Provide an interpretation for the intercept, the slope coefficient fitted for the `Days` variable, and the slope coefficient fittted for the `Care` variable.

Answer: Intercept: I expect an average estimated charge of 44960.90 dollars when an individual who is 0 years old with no education and a blood pressure of 0 stays in the hospital for 0 days to death or hospital discharge and requires a high level of care.

Days: For each additional day to death or hospital discharge an individual stays, given that they require a high level of care, I expect the estimated charge to increase by 2132.40 dollars on average, holding age, education, and blood pressure constant

Care: I expect individuals who require a low level of care to owe an estimated 42472.90 less than those who require a high level of care, holding age, education, blood pressure, and day to death or hospital discharges in hospital constant

part c

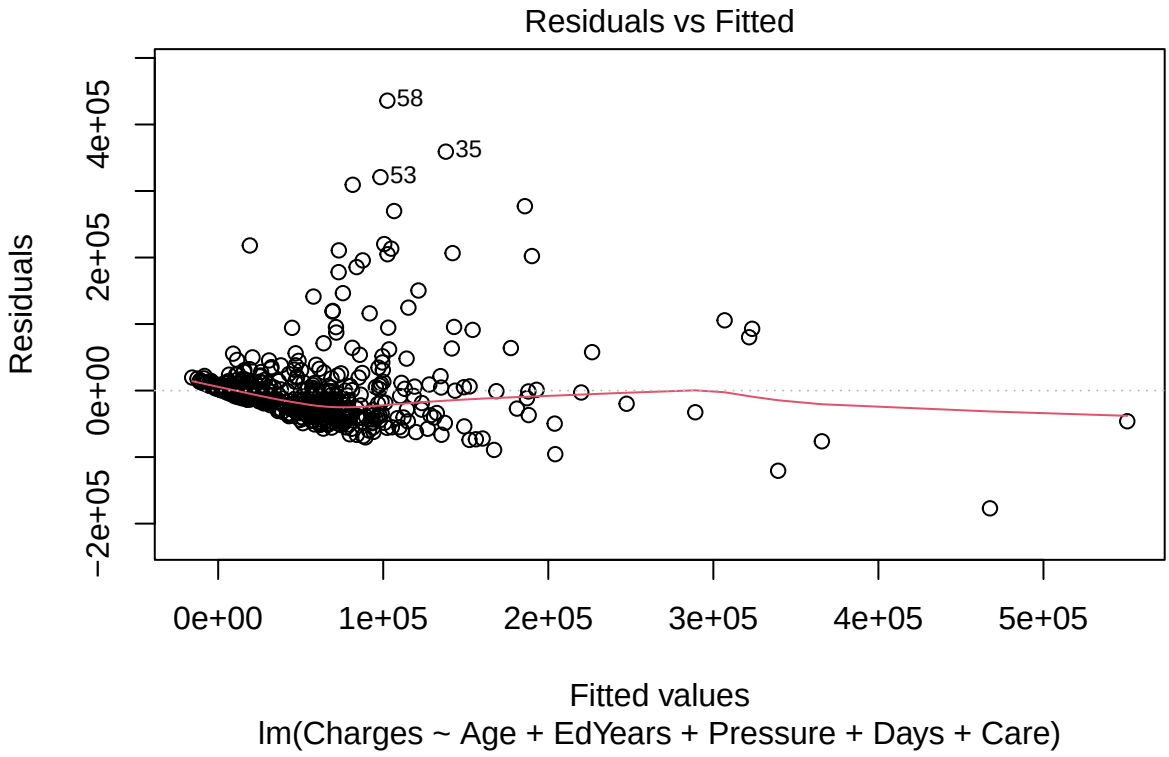
What is the largest and the smallest residual from this model? From the summary of the model, does it seem like the distribution of residuals is roughly symmetric?

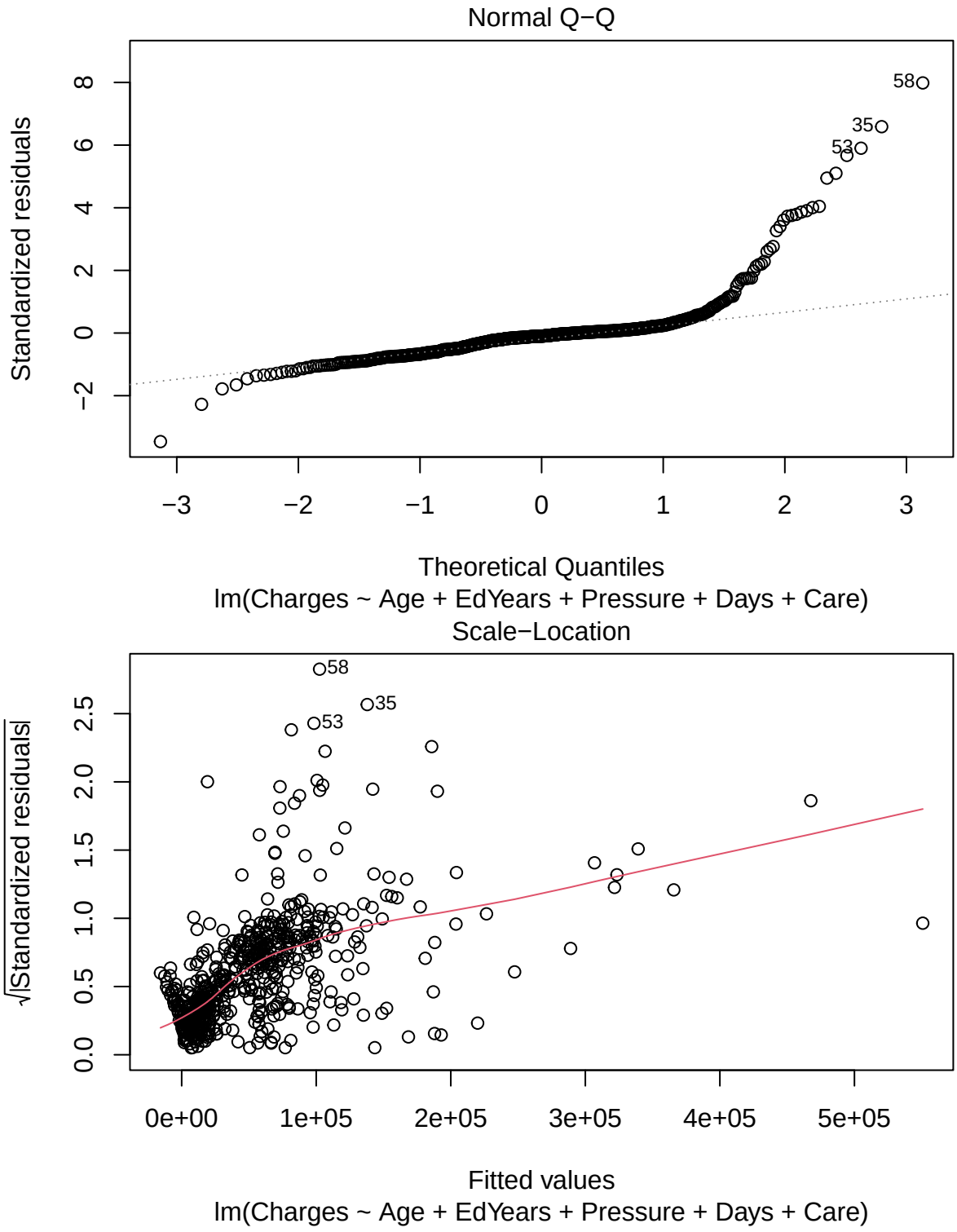
Answer: the smallest residual is -177021 and the largest is 435820. From the output, it does not seem that the distribution of residuals is roughly symetric as many more , and larger, positive residuals exist comapred to negative residuals.

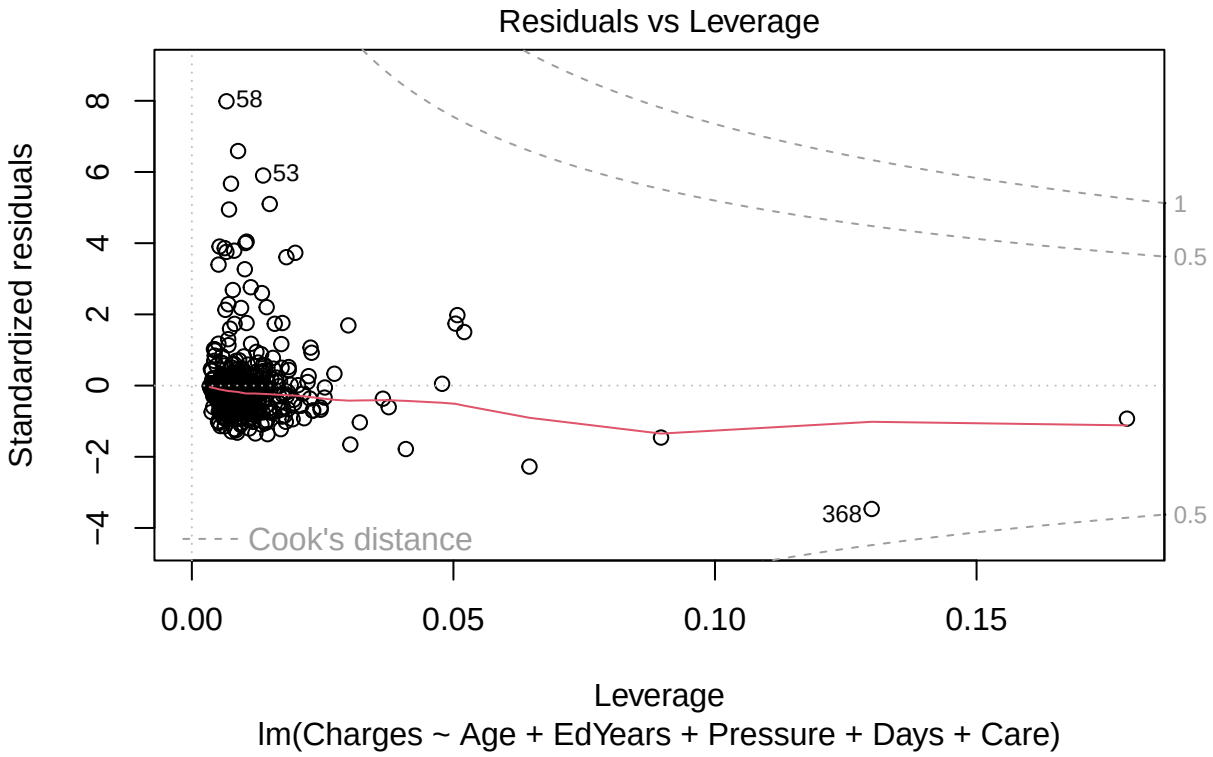
part d

Generate the default plots in R. Then, interpret each of these plots.

```
# Use this code chunk for your answer.
plot(lm1)
```




Normal Q−Q
Standardized residuals
58
35
53
−3
−2
−1
0
1
2
3
−2
0
2
4
6
8
Theoretical Quantiles
lm(Charges ~ Age + EdYears + Pressure + Days + Care)
Scale−Location
√|Standardized residuals|
58
35
53
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
0.0
0.5
1.0
1.5
2.0
2.5
Fitted values
lm(Charges ~ Age + EdYears + Pressure + Days + Care)



Answer: The residuals vs fitted plot tells us that our errors likely do not have constant variance, but that the relationship between y and xs is likely linear. The qq-plot tells us that our errors are likely not distributed normally as we see many observations at either end deviating from the dashed line. The scale-location plot tells us, as the residuals vs fitted values plot did, that our errors are likely not homoskedastic as our red line is not flat but instead has a positive slope. And lastly, the residuals vs leverage plot tells us that we likely have few unusual observations that we should take a look at, specifically 53, 58, and 368.

Exercise 2: Model Selection in Hospital Support Data [25 points]

part a

Fit a multiple regression model with `Charges` as the response. Use the main effects of `Age`, `EdYears`, `Pressure`, and `Days`, along with all second, third, and fourth order interaction terms. For this model, you will not use the `Care` variable that you used in Exercise 1. Report the summary of this model. How many coefficients are included in this model? How many of these coefficients are significantly different from 0?

```
# Use this code chunk for your answer.
lm2 = lm(Charges ~ Age * EdYears * Pressure * Days,
         data = hospital)
```

Answer: There are 16 coefficients included in this model and none of them are statistically diffeent than 0 at even the 10% level

part b

Perform model selection using a forward searching mechanism, BIC as the metric, and the full interaction model from part a as your the scope of the searching. Make sure to define `n` in your code before using the model searching method.

```r
# Use this code chunk for your answer.
null_model = lm(Charges ~ 1, data = hospital)
n = 580
bic_measure = step(null_model, scope = Charges ~ Age * EdYears * Pressure * Days,
                   direction = 'forward', k = log(n))
```

```
## Start:  AIC=13106.64
## Charges ~ 1
##
##             Df  Sum of Sq        RSS   AIC
## + Days       1 1.7141e+12 2.0244e+12 12757
## + Age        1 6.9087e+10 3.6694e+12 13102
## <none>                    3.7385e+12 13107
## + EdYears    1 2.6150e+10 3.7124e+12 13109
## + Pressure   1 4.8299e+08 3.7380e+12 13113
##
## Step:  AIC=12757.22
## Charges ~ Days
##
##             Df  Sum of Sq        RSS   AIC
## + EdYears    1 3.1660e+10 1.9927e+12 12754
## + Pressure   1 2.6739e+10 1.9977e+12 12756
## + Age        1 2.5451e+10 1.9989e+12 12756
## <none>                    2.0244e+12 12757
##
## Step:  AIC=12754.44
## Charges ~ Days + EdYears
##
##                Df  Sum of Sq        RSS   AIC
## + Pressure      1 2.3833e+10 1.9689e+12 12754
## <none>                       1.9927e+12 12754
## + Age           1 1.8243e+10 1.9745e+12 12756
## + EdYears:Days  1 1.5287e+10 1.9775e+12 12756
##
## Step:  AIC=12753.83
## Charges ~ Days + EdYears + Pressure
##
##                    Df  Sum of Sq        RSS   AIC
## + Pressure:Days     1 4.8807e+10 1.9201e+12 12746
## <none>                           1.9689e+12 12754
## + Age               1 1.9278e+10 1.9496e+12 12754
## + EdYears:Days      1 1.5128e+10 1.9538e+12 12756
## + EdYears:Pressure  1 7.5542e+07 1.9688e+12 12760
##
## Step:  AIC=12745.63
## Charges ~ Days + EdYears + Pressure + Days:Pressure
##
##                    Df  Sum of Sq        RSS   AIC
## + Age               1 2.2773e+10 1.8973e+12 12745
## <none>                           1.9201e+12 12746
## + EdYears:Days      1 5.4562e+09 1.9146e+12 12750
## + EdYears:Pressure  1 8.5093e+08 1.9192e+12 12752
##
## Step:  AIC=12745.07
```

```
## Charges ~ Days + EdYears + Pressure + Age + Days:Pressure
##
##                     Df  Sum of Sq        RSS   AIC
## <none>                              1.8973e+12 12745
## + Age:EdYears        1 4035858491 1.8933e+12 12750
## + EdYears:Days       1 3544558406 1.8938e+12 12750
## + Age:Days           1 2889007407 1.8944e+12 12751
## + Age:Pressure       1 1277572323 1.8960e+12 12751
## + EdYears:Pressure   1 1009563696 1.8963e+12 12751
```

```
final_model = lm(Charges ~ Days + EdYears + Pressure + Age + Days:Pressure,
                 data = hospital)
summary(final_model)
```

```
##
## Call:
## lm(formula = Charges ~ Days + EdYears + Pressure + Age + Days:Pressure,
##     data = hospital)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -158287  -23119  -12661    3052  449983
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13191.669  16792.404   0.786   0.4324
## Days           3575.876    309.656  11.548  < 2e-16 ***
## EdYears        1689.961    654.428   2.582   0.0101 *
## Pressure          7.306    112.673   0.065   0.9483
## Age            -401.751    153.059  -2.625   0.0089 **
## Days:Pressure   -11.558      2.906  -3.978 7.84e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57490 on 574 degrees of freedom
## Multiple R-squared:  0.4925, Adjusted R-squared:  0.4881
## F-statistic: 111.4 on 5 and 574 DF,  p-value: < 2.2e-16
```

part c

Report the predictor variables for the model chosen after the first step.

Answer: day to death or hospital discharges

part d

Report the final fitted model chosen by model selection; write out the full fitted model with all coefficients.

Answer: Estimated Charges = 13191.669 + (3575.876 * Days) + (1689.961 * EdYears) + (7.306 * Pressure) + (-401.751 * Age) + (-11.558 * Days * Pressure)

part e

The final model selected in part d should have a total of six coefficients, including the intercept. Interpret the two coefficients pertaining to `Days`. Then, calculate the slopes for `Days` for an individual whose `Pressure` is 115 and again for an individual whose `Pressure` is 67.

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

3575.876 + (-11.558 * 67) + (7.306 * 67)
```

```
## [1] 3290.992
```

```
3575.876 + (-11.558 * 115) + (7.306 * 115)
```

```
## [1] 3086.896
```

Answer: Days interpretation: For each additional day to death or hospital discharge, I expect the estimated charge to increase by 3575.876 dollars on average, holding education, blood pressure, and age constant

Days * Pressure interpretation: For each additional day to death or hospital discharge, holding pressure constant, I expect the estimated charge to decrease by 11.558 dollars on average, holding education and age constant

The slope for days for an individual whos pressure is 67, is 3290.992

The slope for days for an individual whos pressure is 115, is 3086.896

part f

Suppose that we meant to use a backward searching mechanism for part d. Without actually running any code, what would be the variable removed from the model in the first step?

Answer: The interaction between age and days

Exercise 3: Hospital SUPPORT Data: Unusual Observations [20 points]

We'll continue exploring the hospitals dataset in this question. We'll use a model with `Charges` as the response, and with predictors of `Days`, `EdYears`, `Pressure`, `Age`, and the interaction of `Days` with `Pressure`.

part a

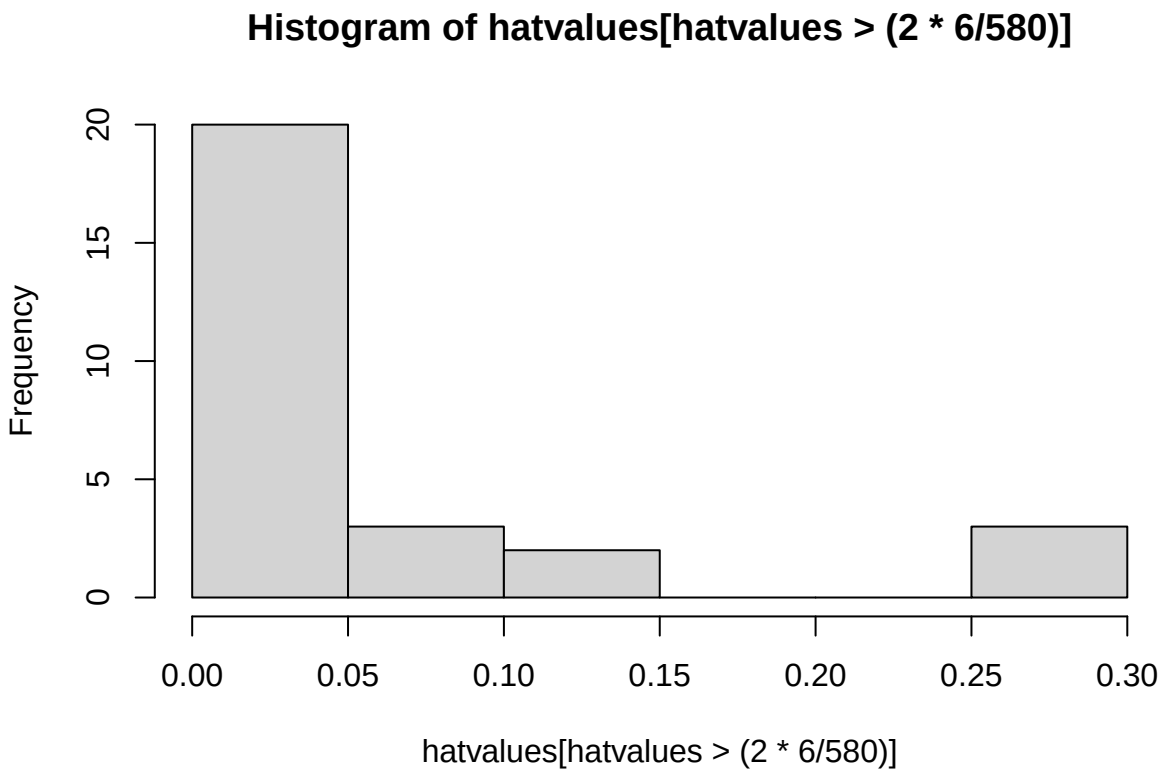
Calculate the leverages for each observation in the dataset. Print only those leverages that are above the threshold defined in the lecture. After looking through these leverages by eye, the leverages for what specific observations, if any, appear to be especially large? Make a histogram of all leverages for the dataset.

```
# Use this code chunk for your answer.
hatvalues = hatvalues(final_model)

hatvalues[hatvalues > (2 * 6/580)]
```

```
##         10         15         26        111        116        118        130
## 0.04634381 0.02640902 0.02772571 0.03715324 0.02587398 0.02102872 0.02762897
##        141        191        198        204        205        207        224
## 0.04295369 0.26713797 0.02441086 0.08683177 0.04311469 0.03044227 0.14264613
##        249        252        257        317        327        368        402
## 0.04864141 0.12447852 0.03974212 0.02069030 0.06980435 0.27460392 0.28258640
##        407        423        443        467        479        499        511
## 0.03130539 0.02375375 0.03807016 0.02449678 0.06011985 0.02118676 0.02147177
```

```
# big leverage = 10, 141, 204, 205, 224, 249, 257, 327, 402, 479
# biggest leverage = 191, 402, 252, 224
```

```
hist(hatvalues[hatvalues > (2 * 6/580)])
```



Answer: The following observation have a large leverage: 10, 141, 204, 205, 224, 249, 257, 327, 402, 479 while observations 191, 402, 252, and 224 have the largest leverages

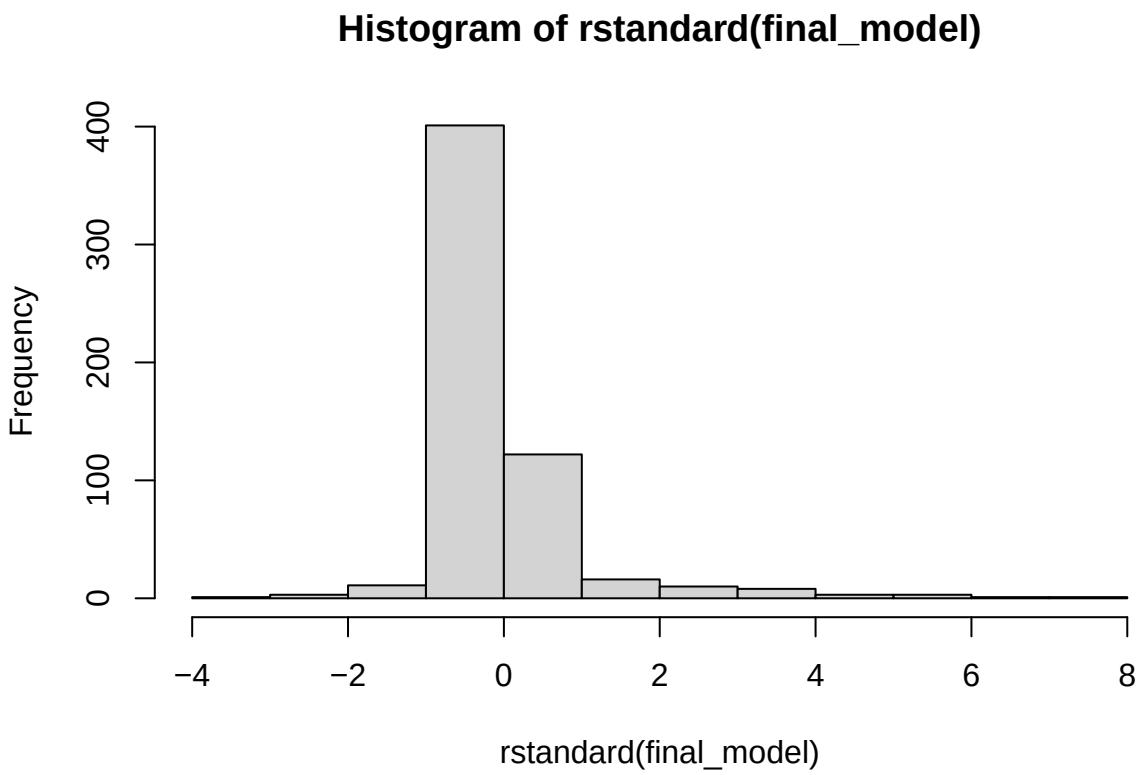
part b

Calculate the standardized residuals for each observation in the dataset. Generate a histogram of all standardized residuals for the dataset. Then, print only those standardized residuals that have a magnitude greater than 2. What observations, if any, were identified as having both a large leverage in part a and as having a high standardized residual in this part? Did you define any of these observations as having an especially large leverage?

```
# Use this code chunk for your answer.
print(rstandard(final_model)[rstandard(final_model) > 2])
```

```
##        2        3        8       14       16       23       24       26
## 3.841903 4.095629 2.391056 3.674166 4.682015 2.075751 4.062163 3.206343
##       34       35       38       39       53       56       58       66
## 2.880794 6.258869 3.781971 5.782725 5.780338 2.993832 7.848393 2.421559
##       67       73       74       75       77       80      204      218
## 5.074611 3.594462 2.871110 3.934032 2.496430 2.372697 2.059669 3.528306
##      333      351
## 2.097318 3.552764
```

```
hist(rstandard(final_model))
```



Answer: observations 26 and 204 were identified as having both a large leverage and a high standard residual

part c

Calculate the Cook's distance for each observation in the dataset. Print only those observations that are above the threshold defined in lecture. After looking through these Cook's distances by eye, the Cook's distance for what specific observations, if any, appear to be especially large? Finally, what is Cook's distance used to measure?

```
# Use this code chunk for your answer.
cooks.distance(final_model)[cooks.distance(final_model) > (4/580)]
```

```
##           2           3          14          16          23          24
## 0.040024463 0.026616539 0.020896820 0.070609267 0.012838702 0.010341465
##          26          34          35          38          39          53
## 0.048861008 0.014618233 0.078228116 0.013923893 0.031448957 0.074117884
##          56          58          66          67          74          75
## 0.009422662 0.056850484 0.007970262 0.023146989 0.015951410 0.011472095
##          77         116         141         191         204         218
## 0.011965780 0.019104409 0.058872958 0.628351971 0.067231322 0.018476456
##         224         230         252         305         327         333
## 0.046861256 0.007125327 0.009141393 0.008505841 0.028897794 0.009041465
##         351         368         479
## 0.014652282 0.386299268 0.038422047
```

Answer: The cooks distance or observation 191 and 368 appear to be especially large. Cook's distnace is used to meausre how influential a point is

part d

In order to assess the fit of this model, calculate the value of the RMSE using leave one out cross validation.

```
# Use this code chunk for your answer.
sqrt(sum(final_model$residuals ^2)/580)
```

```
## [1] 57194.84
```

Exercise 4: Scottish Hill Races [25 points]

For this last exercise, we'll use the `races.table` dataset that includes information on record-winning times for 35 hill races in Scotland, as reported by Atkinson (1986). The additional variables record the overall distance travelled and the height climbed in the race. Below, we are reading in the data from an online source. We do correct one error reported by Atkinson before beginning our analysis.

Source: Atkinson, A. C. (1986). Comment: Aspects of diagnostic regression analysis (discussion of paper by Chatterjee and Hadi). *Statistical Science*, **1**, 397-402.

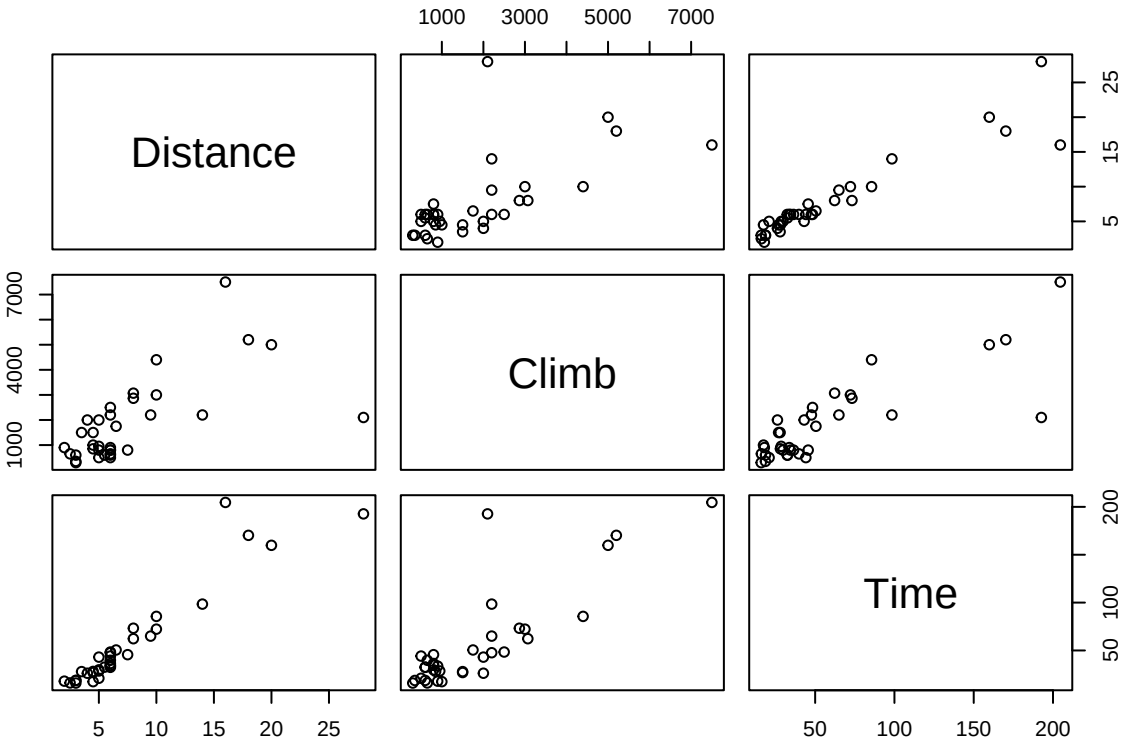
```
url = 'http://www.statsci.org/data/general/hills.txt'
races.table = read.table(url, header=TRUE, sep='\t')
races.table[18,4] = 18.65
head(races.table)
```

```
##          Race Distance Climb   Time
## 1 Greenmantle      2.5   650 16.083
## 2    Carnethy      6.0  2500 48.350
## 3 CraigDunain      6.0   900 33.650
## 4      BenRha      7.5   800 45.600
## 5   BenLomond      8.0  3070 62.267
## 6    Goatfell      8.0  2866 73.217
```

part a

Create a scatterplot matrix of the quantitative variables contained in the `race.table` dataset. Interpret this scatterplot matrix. What variable do you think will be more important in predicting the record time of that race?

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

for_matrix1 = races.table[,2:4]

pairs(for_matrix1)
```



Answer: It appears that climb has a positive relationship with both distane and time and similalrly, distance also has a postive relationship with time. I believe that distance will be more important to predict the record time of a race.

part b

Fit a multiple regression model predicting the record time of a race from the distance travelled, the height climbed, and an interaction of the two variables. Report the summary of the model. What is the R^2 for this model? What does this suggest about the strength of the model?

```
# Use this code chunk for your answer.
lm3 = lm (Time ~ Distance * Climb, data = races.table)
summary(lm3)
```

```
##
## Call:
## lm(formula = Time ~ Distance * Climb, data = races.table)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.3078  -2.8309   0.7048   2.2312  18.9270
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.3532285  3.9121887  -0.090 0.928638
## Distance        4.9290166  0.4750168  10.377 1.32e-11 ***
## Climb           0.0035217  0.0023686   1.487 0.147156
## Distance:Climb  0.0006731  0.0001746   3.856 0.000545 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 7.35 on 31 degrees of freedom
## Multiple R-squared:  0.9806, Adjusted R-squared:  0.9787
## F-statistic: 521.1 on 3 and 31 DF,  p-value: < 2.2e-16
```

Answer: the R-square of this model is 0.9806, indicating that this model is very strong

part c

Identify any influential points as defined in the lecture. Which of these observations, if any, are especially influential based on their values? For these influential points, do they have high leverage, high standardized residual, both, or neither?

```
# Use this code chunk for your answer.
cooks.distance(lm3)[cooks.distance(lm3) > (4/35)]
```

```
##        7       11       35
## 3.758307 2.704165 1.805942
```

```
# 7, 11, 35

hatvalues(lm3)[hatvalues(lm3) > (8/35)]
```

```
##         7        11        33        35
## 0.5207512 0.7182517 0.2379383 0.3261854
```

```
(rstandard(lm3)[rstandard(lm3) > 2])
```

```
##        7       11
## 3.719559 2.059866
```

Answer: Observations 7, 11, and 35 are influential with 7 being especially influential. Observations 7, 11, and 35 all have high leverage. And observatons 7 and 11 have a large standardized residuals.

part d

Refit the model from part b without any points that you identified as influential. Note: this is not something that we should automatically do, but we will do it for now as a demonstration of how much our model may be affected by these points! Print the coefficients for this model. How do they compare to the coefficients from the model in part b?

Hint: Create a subset of your data that only includes those points that are not influential before fitting your data.

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

remove_obs = races.table[-c(7, 11, 35),]

lm4 = lm(Time ~ Distance * Climb, data = remove_obs)
summary(lm4)
```

```
##
## Call:
## lm(formula = Time ~ Distance * Climb, data = remove_obs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1574  -2.7089   0.3387   2.2074  10.3180
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.6141193  3.3490412   0.183 0.855828
## Distance       5.1003107  0.6078802   8.390 3.98e-09 ***
## Climb          0.0018117  0.0017531   1.033 0.310273
## Distance:Climb 0.0007105  0.0001663   4.272 0.000202 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.877 on 28 degrees of freedom
## Multiple R-squared:  0.9778, Adjusted R-squared:  0.9754
## F-statistic: 411.1 on 3 and 28 DF,  p-value: < 2.2e-16
```

Answer: All of the coefficients change a lot compared to part b with the exception of the interaction not changing too much

part e

How much does this updated model affect our actual predictions for the response? Let's create a scatterplot that compares our fitted values from our original model to those from our newer model (influential points removed).

Calculate and save each of the fitted values (for the original model and for the newer model) to their own named object in `R`. Note: If you are using the `predict` function, you can supply as an argument `newdata = races.table` since we will use all of the variables and all of the data.

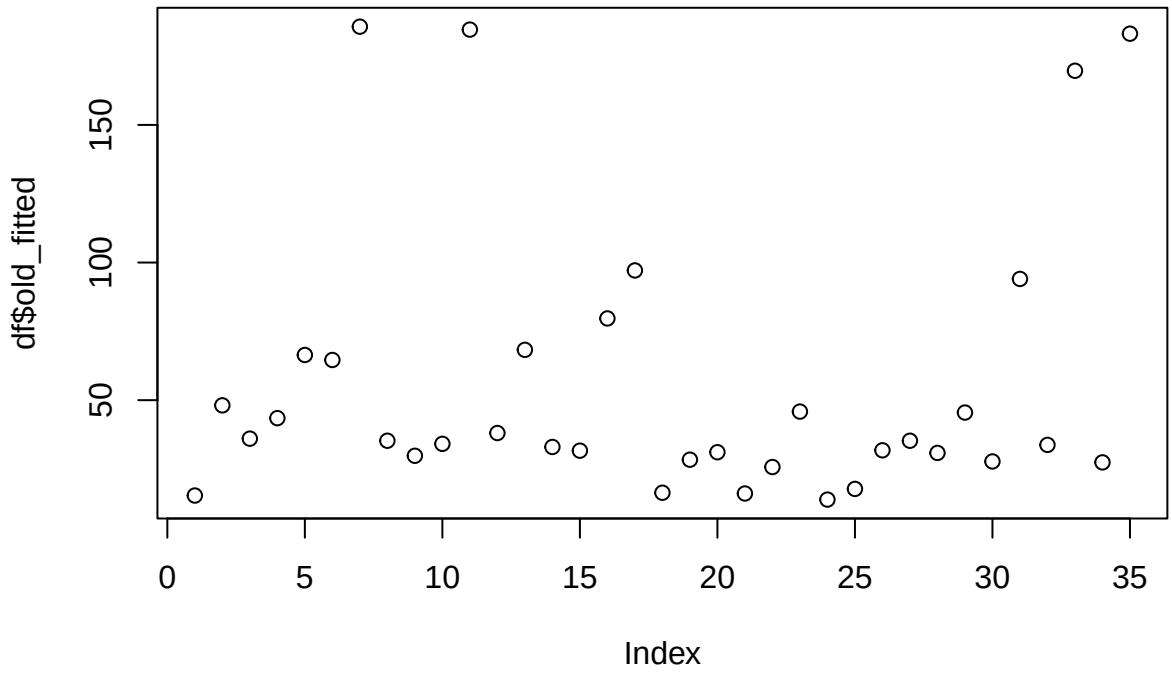
Then, create a dataframe in `R` by providing your two named objects with fitted values as two arguments inside the `data.frame` function, and save the result to a new named object in `R`.

Now, create a scatterplot to compare the fitted values for each model. Include an appropriate title and axes labels. All other formatting is optional and up to you!

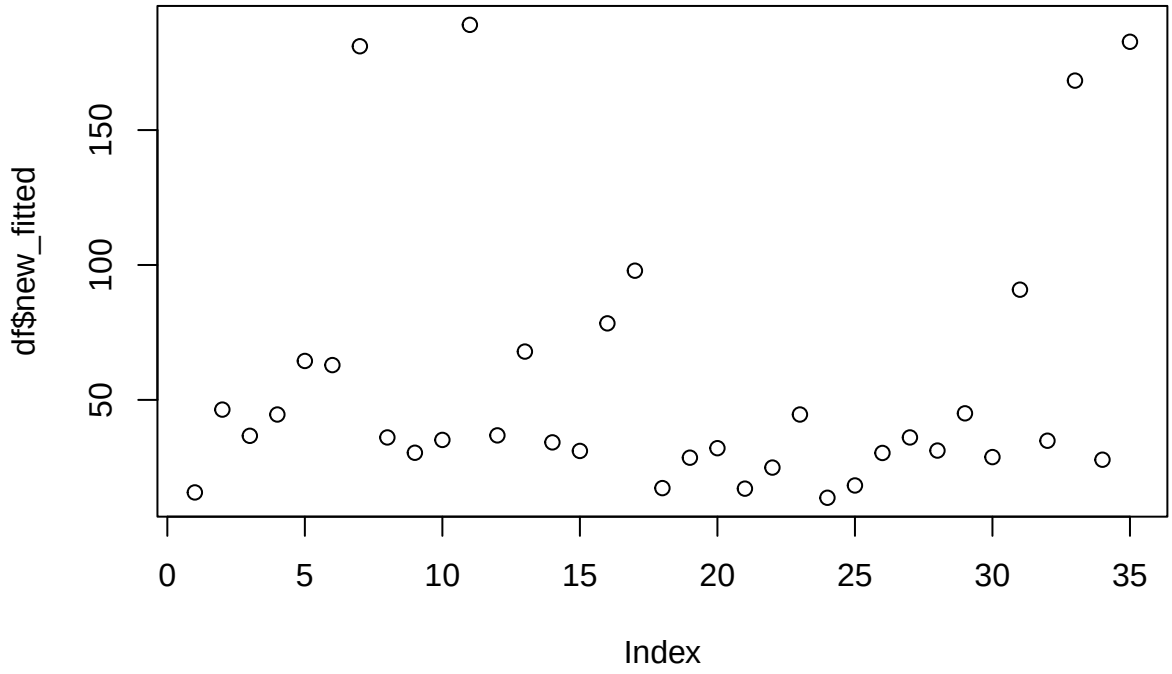
It might be helpful to add a line with intercept 0 and slope 1 to represent what perfect matching would look like.

Finally, briefly comment on what this plot reveals. Would you say there are big differences in the predictions made by each model, or would you say the predictions by each model are quite similar? Is this what you would expect from the results in part c?

```
# Use this code chunk for your answer.
old_fitted = predict(lm3, newdata = races.table)

new_fitted = predict(lm4, newdata = races.table)


df = data.frame(old_fitted, new_fitted)

plot(df$old_fitted)
```




```
plot(df$new_fitted)
```




```
# HOW?
```

Answer:

Exercise 5: 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 5)
- all code is printed and readable for each question
- all output is printed
- generated a pdf file