Stat 21 Homework 6

Person 1, Person 3, etc

Due: Sunday, March 20th by midnight

Contents

Use this file as the template for your submission. Do not delete anything from this template unless you are prompted to do so (e.g. where to write your name above, where to write your solutions or code below). Make sure you have installed the following packages in your version of RStudio: `tidyverse`, `knitr` **before** you attempt to knit this document.

Your completed assignment should be submitted as a single **PDF** using the link under Week 8 titled "Submit HW 6 to Gradescope". You must use R markdown to write up your solutions. You are allowed to work with your classmates on this homework assignment for all problems *except problem 10* which I recommend you complete on your own. **This homework assignment will be graded for completion rather than for correctness so please pay careful attention to the following additional instructions to make sure you receive credit for your work.**

Additionally, make sure that when you upload your solutions to Gradescope, you select which pages go correspond with which questions. Also, check to make sure that your knitted homework document is not uploaded as an extra-long single page document. Failure to do these things will result in a penalty on your homework grade. Finally, I strongly recommend that you address and resolve any knitting or R coding issues before Saturday as solutions to any R-coding questions that are not knitted properly will not receive any credit.

Consider the 2016 MLB data that we explored in class, the first six rows of which are shown below.

```r
data("MLBStandings2016")
MLBStandings2016 %>% head
```

```
##                    Team League Wins Losses WinPct BattingAverage Runs Hits  HR
## 1 Arizona Diamondbacks     NL   69     93  0.426          0.261  752 1479 190
## 2       Atlanta Braves     NL   68     93  0.422          0.255  649 1404 122
## 3    Baltimore Orioles     AL   89     73  0.549          0.256  744 1413 253
## 4       Boston Red Sox     AL   93     69  0.574          0.282  878 1598 208
## 5         Chicago Cubs     NL  103     58  0.640          0.256  808 1409 199
## 6    Chicago White Sox     AL   78     84  0.481          0.257  686 1428 168
##   Doubles Triples RBI  SB   OBP   SLG   ERA HitsAllowed Walks StrikeOuts Saves
## 1     285      56 709 137 0.320 0.432 5.09        1563   603       1318    31
## 2     295      27 615  75 0.321 0.384 4.51        1414   547       1227    39
## 3     265       6 710  19 0.317 0.443 4.22        1408   545       1248    54
## 4     343      25 836  83 0.348 0.461 4.00        1342   490       1362    43
## 5     293      30 767  66 0.343 0.429 3.15        1125   495       1441    38
## 6     277      33 656  77 0.317 0.410 4.10        1422   521       1270    43
##    WHIP
## 1 1.492
## 2 1.355
## 3 1.364
## 4 1.273
## 5 1.110
## 6 1.343
```

We are going to consider the following variables to create a MLR model and practice visualizing the data for this model.

Predictors: League (categorical), Runs, OBP, ERA

Response: WinPct

1.

Use `select()` function to create new data set called `my_MLB` that contains only the variables we are interested in using in our model (listed above).

2.

Use the space below to create two side-by-side box plots for the numeric response variable over each level of the categorical variable `League`. Briefly interpret these box plots.

[Leave your comments here]

3.

Use the space below and the `filter()` function to create a new data set called `your_MLB` that contains only the data for either the National League or the American League (your choice).

4.

Use the space below to create a scatter plot comparing the numeric predictor `OBP` to another numeric predictor, `ERA`. Briefly interpret this plot.

[Leave your comments here]

5.

Use the space below to fit a MLR model to the data set `my_MLB` with all predictor variables and response `WinPct`. Then, use the `mutate()` function to add two columns to the `my_MLB` data set corresponding to the residuals and the fitted values. Use the `head()` function to print the first six rows of `my_MLB`.

6.

Use the space below to create a histogram of the residuals from your model in problem 5.

7.

Use the `mutate()` function and the `rstudent()` function to add another column onto the `my_MLB` data set that consists of the studentized versions of the residuals from the model in problem 5. Use the `head()` function to print the first six rows of `my_MLB`.

8.

Use the space below to create a histogram of the studentized residuals that you added to the `my_MLB` data set in problem 7. Comment on the difference/similarity between this histogram and the one in problem 7.

[Leave your comments here]

9.

Use the space below to create a Normal quantile plot for the studentized residuals from problem 8. Compare this plot to the histogram in problem 8 and comment on their relationship.

[Leave your comments here]

10.

Improve upon this MLR model in some demonstrable way. Include your code for the new model and the new data based on this model and then justify that your model is an improvement.

[Write about why this is an improved version of the model here.]