Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 2

Phat Doan, Ted Pham

June 22, 2017

The Lab:

The purpose of this lab is to practice using Poission Regression models, conduct residual diagnostics, and examine goodness of fit.

Data

Agresti (2007) provides data on the social behavior of horseshoe crabs. These data are contained in the HorseshoeCrabs.csv file available on our website. Each observation corresponds to one female crab. The response variable is Sat, the number of "satellite" males in her vicinity. Physical measurements of the female-Color (4-level ordinal), Spine (3-level ordinal), Width (cm), and Weight (kg)-are explanatory variables.

```
suppressMessages(library(car))
suppressMessages(library(Hmisc))
```

```
df = read.csv("HorseshoeCrabs.csv")
```

Exploratory Data Analysis

Summary of Descriptive Statistical Analysis

```
describe(df)
```

```
## df
##
##  5  Variables      173  Observations
## ---------------------------------------------------------------------------
## Color
##        n  missing distinct     Info     Mean      Gmd
##      173        0        4    0.816    2.439   0.8278
##
## Value          1     2     3     4
## Frequency     12    95    44    22
## Proportion 0.069 0.549 0.254 0.127
## ---------------------------------------------------------------------------
## Spine
##        n  missing distinct     Info     Mean      Gmd
##      173        0        3    0.647    2.486   0.7611
##
## Value          1     2     3
## Frequency     37    15   121
```

```
## Proportion 0.214 0.087 0.699
## ---------------------------------------------------------------------------
## Width
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      173        0       66    0.999     26.3    2.379    23.00    23.70
##      .25      .50      .75      .90      .95
##    24.90    26.10    27.70    29.00    29.88
##
## lowest : 21.0 22.0 22.5 22.9 23.0, highest: 30.3 30.5 31.7 31.9 33.5
## ---------------------------------------------------------------------------
## Weight
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      173        0       58    0.999    2.437   0.6411     1.60     1.80
##      .25      .50      .75      .90      .95
##     2.00     2.35     2.85     3.19     3.28
##
## lowest : 1.20 1.30 1.40 1.47 1.55, highest: 3.50 3.60 3.73 3.85 5.20
## ---------------------------------------------------------------------------
## Sat
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      173        0       15    0.949    2.919    3.354        0        0
##      .25      .50      .75      .90      .95
##        0        2        5        7        9
##
## Value          0     1     2     3     4     5     6     7     8     9
## Frequency     62    16     9    19    19    15    13     4     6     3
## Proportion 0.358 0.092 0.052 0.110 0.110 0.087 0.075 0.023 0.035 0.017
##
## Value         10    11    12    14    15
## Frequency      3     1     1     1     1
## Proportion 0.017 0.006 0.006 0.006 0.006
## ---------------------------------------------------------------------------
```

- There are 173 observations, each corresponding to a horseshoe crab and its characteristics. In total 5 variables are available:
  - Color : Physical measurements of the female-Color (4-level ordinal)
  - Spine : 3-level ordinal
  - Width : in cm
  - Weight : in kg
  - Sat : the number of "satellite" males in her vicinity
- There are no missing values for all records and variables.
- We are interested in predicting number of satellite males in a female crab vicinity with poisson regression models.
  - The response variable is "Sat" to indicate number of satellite males in.
  - The explanatory variable can be "Color", "Spine", "Width", and "Weight".

Univariate Analysis

Response Variable: Sat

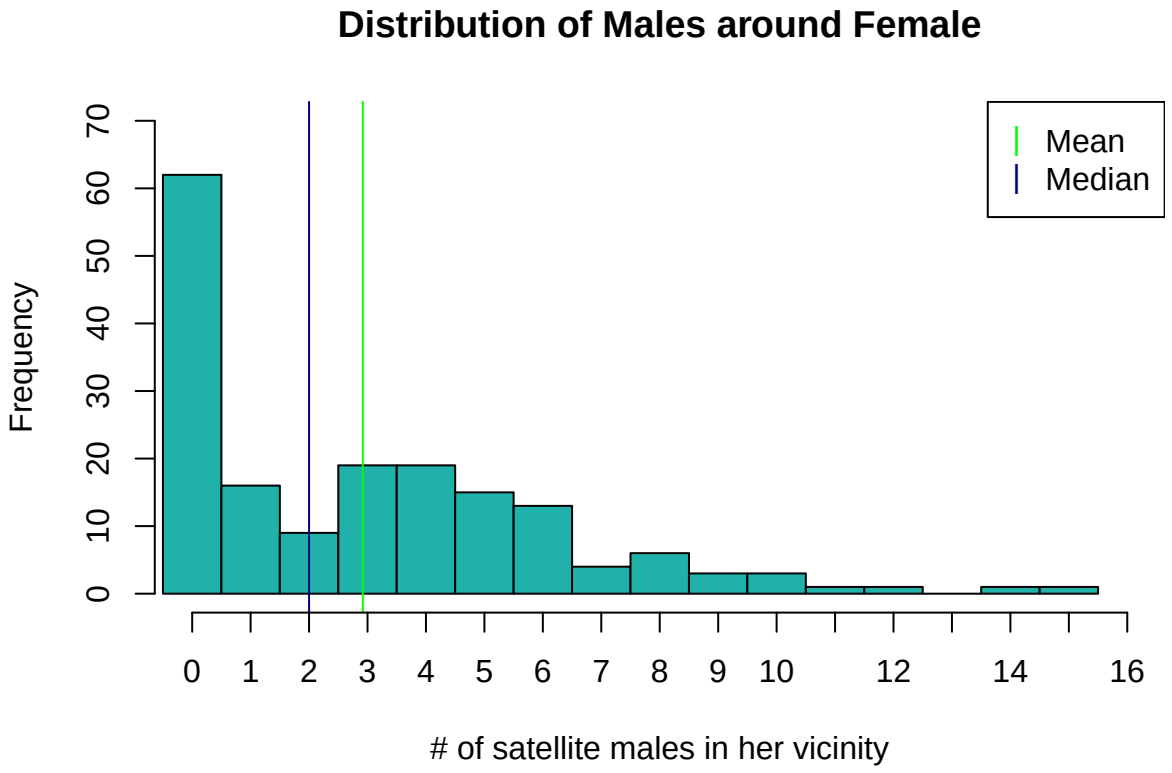
```
#sort(df$Sat)

hist(df$Sat, breaks = 0:16 - 0.5
    ,xlim = range(0:16)
    ,ylim = c(0, 70)
    ,col = "lightseagreen"
    ,xaxt = "n" #remove default x axis marker
    ,main = "Distribution of Males around Female"
    ,xlab = "# of satellite males in her vicinity"
    )

abline(v = mean(df$Sat), col = "green")
abline(v = median(df$Sat), col = "navy")

#add marker for x axis to be interger only
axis(1, at = 0:16)

legend("topright", c("Mean", "Median"), col = c("green", "navy"), pch = "|")
```



Response Variable: Color

```
#summary(df$Color)
```

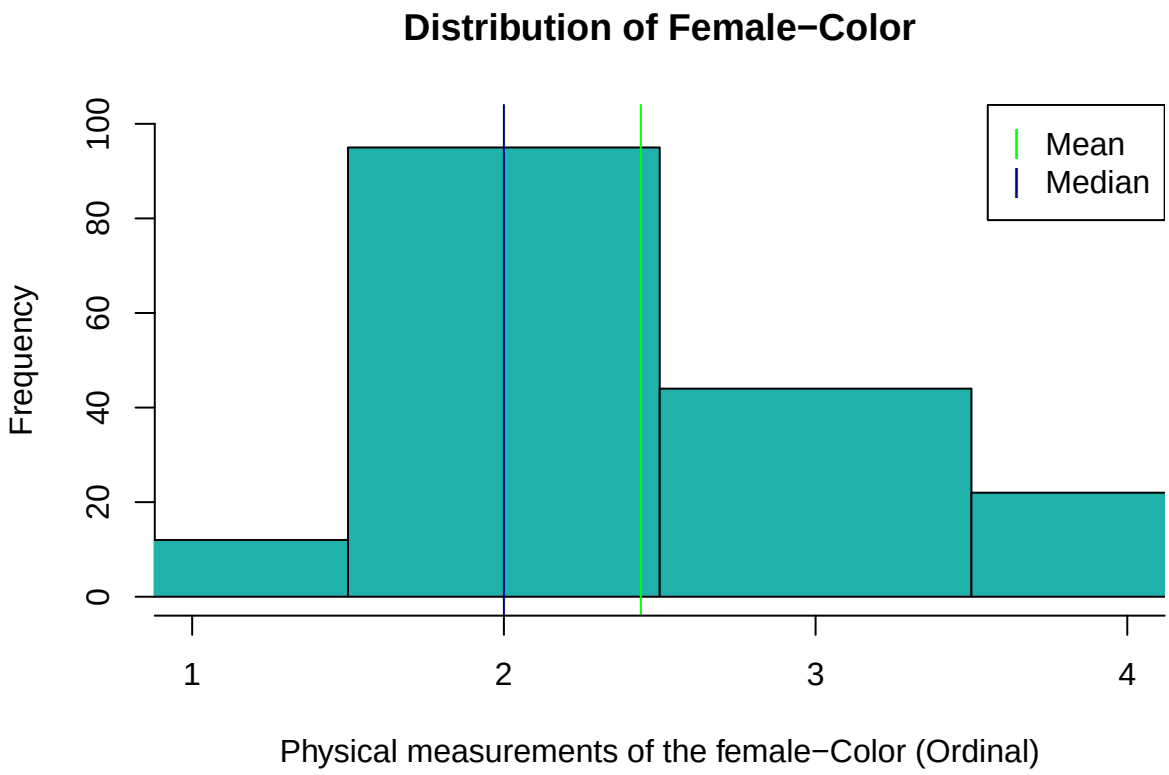
```r
#sort(df$Color)

hist(df$Color, breaks = 0:5 - 0.5
    ,xlim = range(1:4)
    ,ylim = range(0:100)
    ,col = "lightseagreen"
    ,xaxt = "n" #remove x lab
    ,main = "Distribution of Female-Color"
    ,xlab = "Physical measurements of the female-Color (Ordinal)"
    )

abline(v = mean(df$Color), col = "green")
abline(v = median(df$Color), col = "navy")

#add x lab to only contain 1 to 4
axis(1, at = 0:5)

legend("topright", c("Mean", "Median"), col = c("green", "navy"), pch = "|")
```



Response Variable: Spine

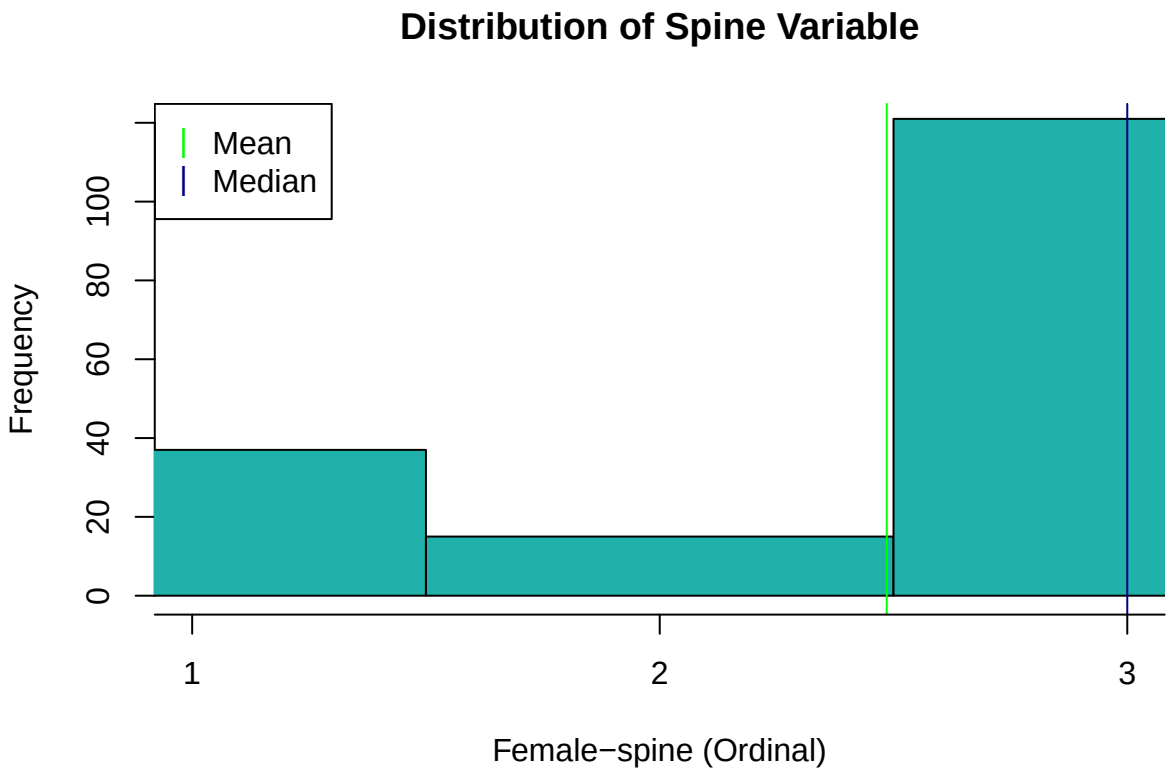
```r
hist(df$Spine, breaks = 0:4 - 0.5
    ,xlim = range(1:3)
    ,ylim = range(0:120)
    ,col = "lightseagreen"
    ,xaxt = "n" #remove x lab
    ,main = "Distribution of Spine Variable"
```

```
    ,xlab = "Female-spine (Ordinal)"
    )

abline(v = mean(df$Spine), col = "green")
abline(v = median(df$Spine), col = "navy")

#add x lab to only contain 1 to 4
axis(1, at = 0:4)

legend("topleft", c("Mean", "Median"), col = c("green", "navy"), pch = "|")
```

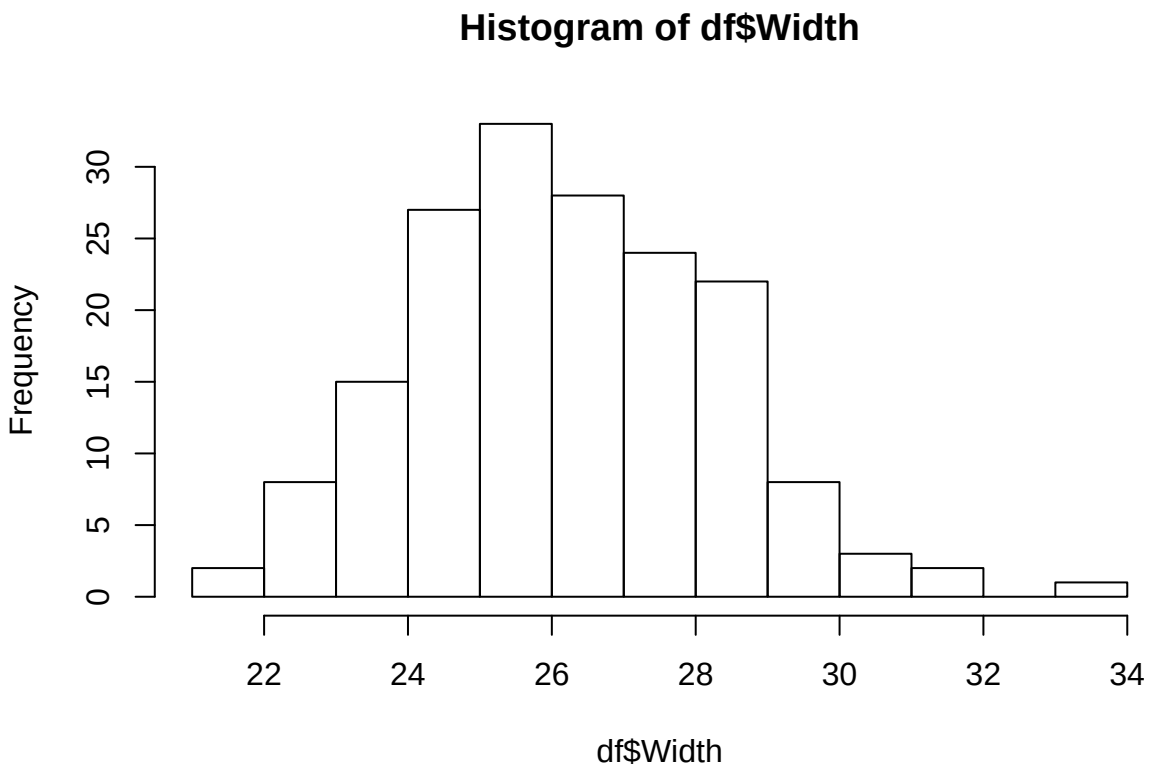


Response Variable: Width

```
describe(df$Width)
```

```
## df$Width
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      173        0       66    0.999     26.3    2.379    23.00    23.70
##      .25      .50      .75      .90      .95
##    24.90    26.10    27.70    29.00    29.88
##
## lowest : 21.0 22.0 22.5 22.9 23.0, highest: 30.3 30.5 31.7 31.9 33.5
```

```
hist(df$Width)
```



Response Variable: Width

Questions:

23.

(a) Fit a Poisson regression model with a log link using all four explanatory variables in a linear form. Test their significance and summarize results.

```
mod.pois <- glm(formula = Sat ~ Color + Spine + Width + Weight, family =
poisson(link = "log"), data = df)
summary(mod.pois)
```

```
##
## Call:
## glm(formula = Sat ~ Color + Spine + Width + Weight, family = poisson(link = "log"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0127  -1.8844  -0.5401   0.9449   4.9605
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.52381    0.94909  -0.552  0.58101
```

```
## Color       -0.18503    0.06652  -2.781  0.00541 **
## Spine        0.04007    0.05681   0.705  0.48061
## Width        0.02728    0.04796   0.569  0.56954
## Weight       0.47319    0.16493   2.869  0.00412 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 551.83  on 168  degrees of freedom
## AIC: 917.13
##
## Number of Fisher Scoring iterations: 6
```

```
Anova(mod.pois)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Sat
##        LR Chisq Df Pr(>Chisq)
## Color    7.9690  1   0.004759 **
## Spine    0.5009  1   0.479111
## Width    0.3221  1   0.570374
## Weight   8.3329  1   0.003893 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(mod.pois)
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %      97.5 %
## (Intercept) -2.37766007  1.34224959
## Color       -0.31683684 -0.05599568
## Spine       -0.07026323  0.15255892
## Width       -0.06741930  0.12054802
## Weight       0.15118202  0.79735713
```

(b) Compute deviance/df and interpret its value.

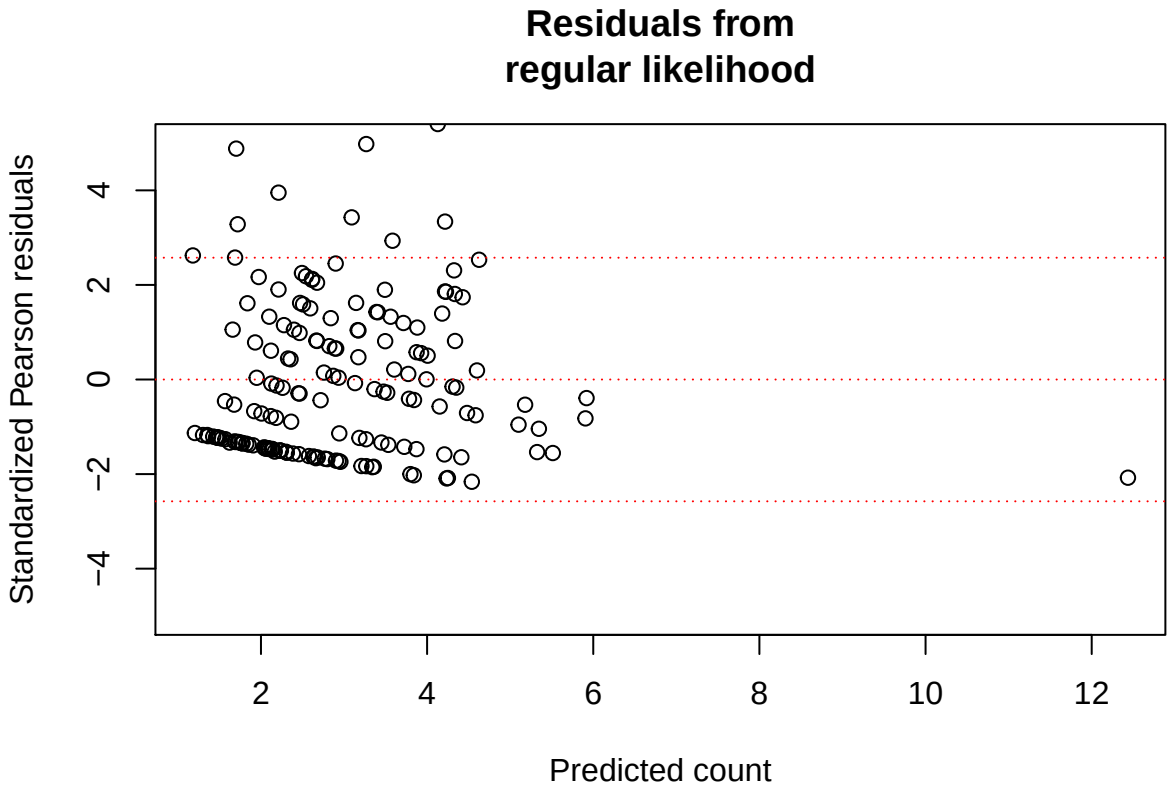
$\text{deviance/df} = 551.83/168 = 3.285 > 1.3273 (= 1 + 3\sqrt{\frac{2}{168}})$

deviance/df is much larger than the suggested guidelines above indicating that the model is a poor fit

(c) Examine residual diagnostics and identify any potential problems with the model.

```
pred <- predict(mod.pois, type = "response")
# Standardized Pearson residuals
stand.resid <- rstandard(model = mod.pois, type = "pearson")
plot(x = pred, y = stand.resid, xlab = "Predicted count", ylab
= "Standardized Pearson residuals", main = "Residuals from
regular likelihood", ylim = c(-5,5))
```

```
abline(h = c(qnorm(0.995), 0, qnorm(0.005)), lty = "dotted",
col = "red")
```



8 of the 173 residuals lie beyond ±3

(d) Carry out the GOF test described on page 296 using the PostFitGOFTest() function available in the PostFitGOFTest.R program of the same name from our website. Use the default number of groups. (M/5 when M 100)

i. State the hypotheses, test statistic and p-value, and interpret the results.

ii. Plot the Pearson residuals for the groups against the interval centers (available in the pear.res and centers components, respectively, of the list returned by the function). Use this plot and the residual plots from part (c) to explain the result

24. Conduct an influence analysis. Interpret the results.

25. Notice that there is one crab with a weight that is substantially different from the rest. This can be seen, for example, in a histogram of the weights. Remove this crab from the data and repeat the steps from Exercise 23. Has this fixed any problems with the model? Are there any other problems with the model, and what could be done to solve these problems?

Instructions:

- **Due Date: 7/2/2017 (11:59 p.m. PST, Sunday)**
- Submission:
  - Submit your own assignment via ISVC
  - Submit 2 files:
    1. A pdf file including the summary, the details of your analysis, and all the R codes used to produce the analysis. Please do not suppress the codes in your pdf file.
    2. R markdown file used to produce the pdf file
  - Each group only needs to submit one set of files
  - Use the following file naming convensation; fail to do so will receive 10% reduction in the grade:
    - SectionNumber_hw01_FirstNameLastNameFirstInitial.fileExtension
    - For example, if you are in Section 1 and have two students named John Smith and Jane Doe, you should name your file the following
      - Section1_hw01_JohnS_JaneD.Rmd
      - Section1_hw01_JohnS_JaneD.pdf
  - Although it sounds obvious, please write the name of each members of your group on page 1 of your report.
  - This lab can be completed in a group of up to 3 people. Each group only needs to make one submission. Although you can work by yourself, we encourage you to work in a group.
  - When working in a group, do not use the "division-of-labor" approach to complete the lab. That is, do not divide the lab by having Student 1 completed questions 1 - 3, Student 2 completed questions 4 - 6, etc. Asking your teammate to do the questions for you takes away your own opportunity to learn.
- Other general guidelines:
  - Please read the instructions carefully.
  - Please read the questions carefully.
  - Use only techniques and R libraries that are covered in this course.
  - If you use R libraries and/or functions to conduct hypothesis tests not covered in this course, you will have to explain why the function you use is appropriate for the hypothesis you are asked to test
  - Thoroughly analyze the given dataset. Detect any anomalies, including missing values, potential of top and/or bottom code, etc, in each of the variables.
  - Your report needs to include a comprehensive Exploratory Data Analysis (EDA) analysis, which includes both graphical and tabular analysis, as taught in this course.

  - Your analysis needs to be accompanied by detailed narrative. Remember, make sure your that when your audience (in this case, the professors and your classmates) can easily understand your your main conclusion and follow your the logic of your analysis. Note that just printing a bunch of graphs and model results, which we call "output dump", will likely receive a very low score.
  - Your rationale of any decisions made in your modeling needs to be explained and supported with empirical evidence. Remember to use the insights generated from your EDA step to guide your modeling step, as we discussed in live sessions.
  - All the steps to arrive at your final model need to be shown and explained very clearly.
- Students are expected to act with regards to UC Berkeley Academic Integrity.
