Midterm Exam

Rong Li

11/2/2020

Instruction

This is your midterm exam that you are expected to work on it alone. You may NOT discuss any of the content of your exam with anyone except your instructor. This includes text, chat, email and other online forums. We expect you to respect and follow the GRS Academic and Professional Conduct Code.

Although you may NOT ask anyone directly, you are allowed to use external resources such as R codes on the Internet. If you do use someone's code, please make sure you clearly cite the origin of the code.

When you finish, please compile and submit the PDF file and the link to the GitHub repository that contains the entire analysis.

Introduction

In this exam, you will act as both the client and the consultant for the data that you collected in the data collection exercise (20pts). Please note that you are not allowed to change the data. The goal of this exam is to demonstrate your ability to perform the statistical analysis that you learned in this class so far. It is important to note that significance of the analysis is not the main goal of this exam but the focus is on the appropriateness of your approaches.

Data Description (10pts)

Please explain what your data is about and what the comparison of interest is. In the process, please make sure to demonstrate that you can load your data properly into R.

My data is about the monthly mean takeout times of three groups: undergraduate students, graduate students and working people. The data contains 18 observations and 3 columns.
I wonder which group orders takeouts most oftenly.

Here are the first several rows of my data:

```
##   status monthly.average most.ordered
## 1      1               6            1
## 2      1               3            2
## 3      1               5            2
## 4      1              18            3
## 5      1              10            2
## 6      1               2            8
```

'status' refers to the group of the observation.
1: undergraduate students
2: graduate students
3: working people

'monthly.average' shows the average of takeout times every month.

'most.ordered' means the most ordered kind of food.
1: desserts and drinks
2: fast food
3: chinese cuisine
5: snacks
8: fruit
10: fresh product

The third column isn't related to our question, so we will not use it in the following analysis.

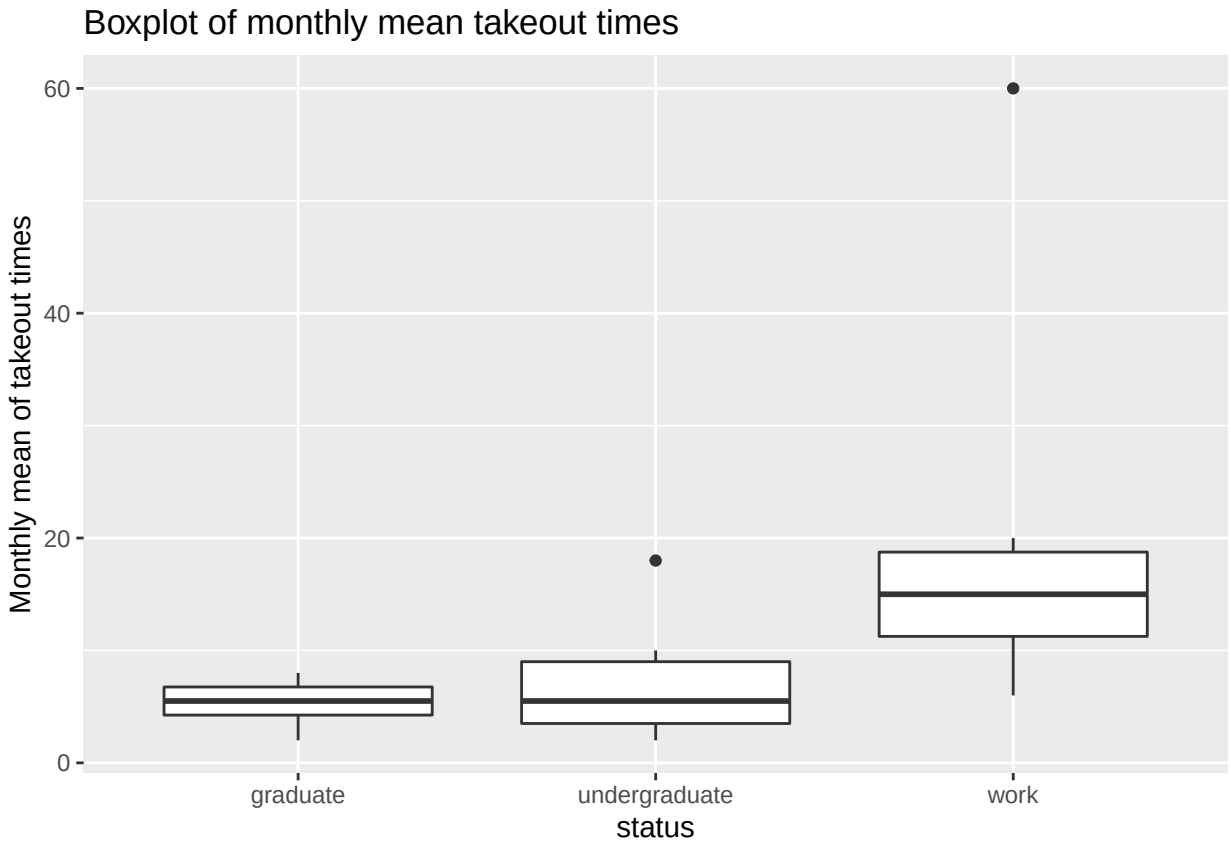
EDA (10pts)

Please create one (maybe two) figure(s) that highlights the contrast of interest. Make sure you think ahead and match your figure with the analysis. For example, if your model requires you to take a log, make sure you take log in the figure as well.

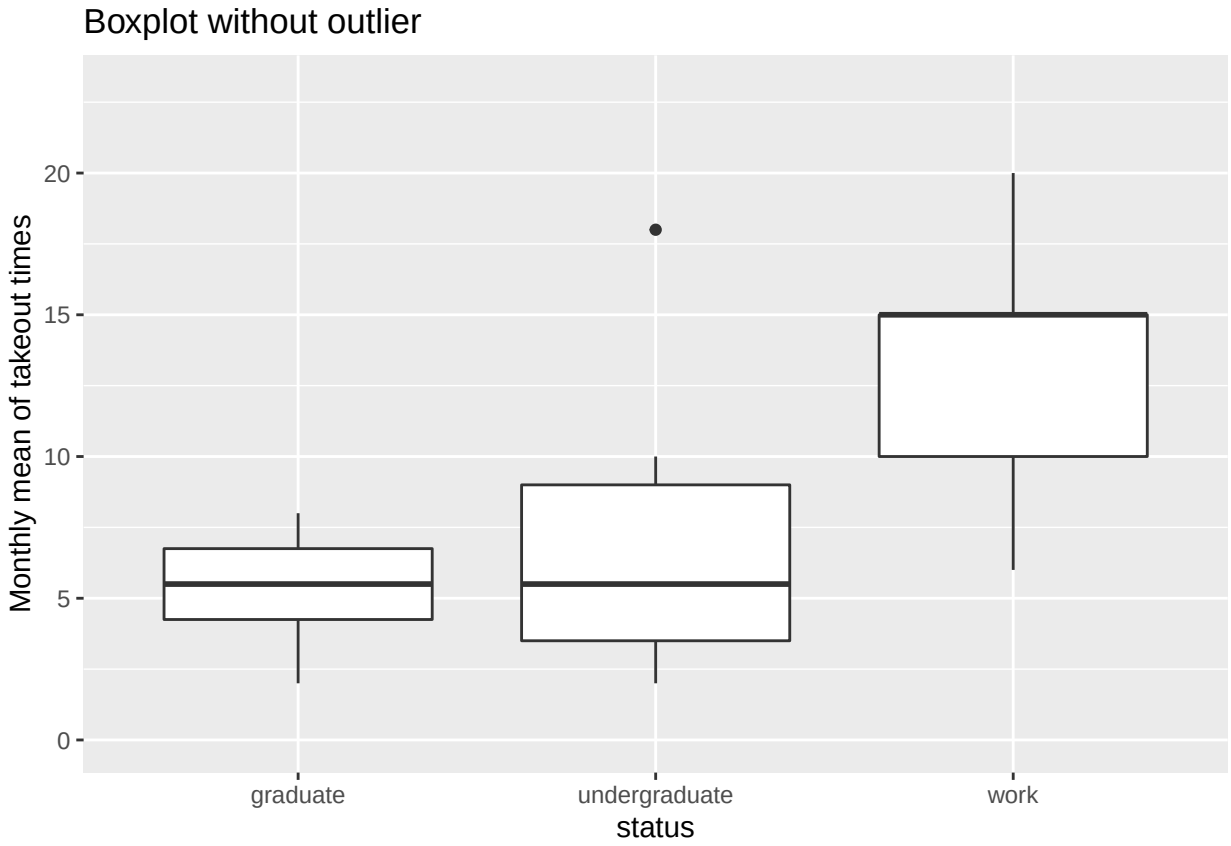
First I calculate the mean and standard deviation of each group.

```
##                    mean        sd
## undergraduate  7.333333  5.921711
## graduate       5.333333  2.160247
## work          21.000000 19.697716
```

Then, I create the boxplot of monthly mean takeout times of three groups.



There is a outlier in the last graph, so we should drop it.



From the graph, we can infer that there is a difference between each group.

Power Analysis (10pts)

Please perform power analysis on the project. Use 80% power, the sample size you used and infer the level of effect size you will be able to detect. Discuss whether your sample size was enough for the problem at hand. Please note that method of power analysis should match the analysis. Also, please clearly state why you should NOT use the effect size from the fitted model.

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 3
##               n = 6
##               f = 0.811284
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

The level of effect size is 0.811. It is large effect size.

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 3
##               n = 52.3966
##               f = 0.25
##       sig.level = 0.05
```

```
##           power = 0.8
##
## NOTE: n is number in each group
```

The result means we need more than 52 observations for each group. My sample size is not enough for the problem at hand. We should not use the effect size because of M-type error.

Modeling (10pts)

Please pick a regression model that best fits your data and fit your model. Please make sure you describe why you decide to choose the model. Also, if you are using GLM, make sure you explain your choice of link function as well.

I use the linear regression to predict the monthly mean of takeout times and I choose the status as predictor.

```
fit1 <- lm(monthly.average ~ factor(status), data = df)
summary(fit1)
```

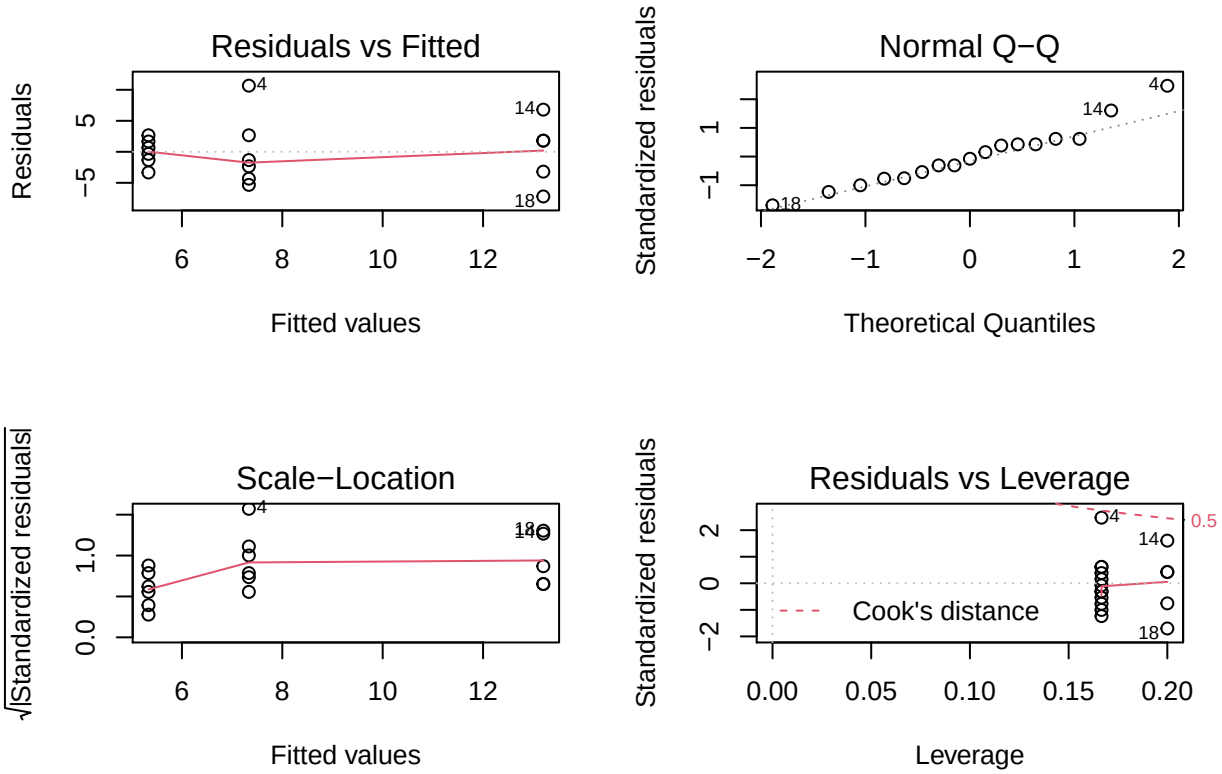
```
##
## Call:
## lm(formula = monthly.average ~ factor(status), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2000 -3.2000 -0.3333  1.8000 10.6667
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.333      1.932   2.761   0.0153 *
## factor(status)undergraduate     2.000      2.732   0.732   0.4762
## factor(status)work              7.867      2.865   2.746   0.0158 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.732 on 14 degrees of freedom
## Multiple R-squared:  0.3627, Adjusted R-squared:  0.2717
## F-statistic: 3.984 on 2 and 14 DF,  p-value: 0.04269
```

The model is $monthly.average = 5.333 + 2.000 \times indicator(undergraduate) + 7.867 \times indicator(work)$.
I choose this model because of the following reasons:
a. There is only one predictor;
b. The only predictor is indicator, so we should factor it;
c. The 'monthly.average' is a continues variable.

Validation (10pts)

Please perform a necessary validation and argue why your choice of the model is appropriate.

```
# External validation:
par(mfrow = c(2, 2))
plot(fit1)
```

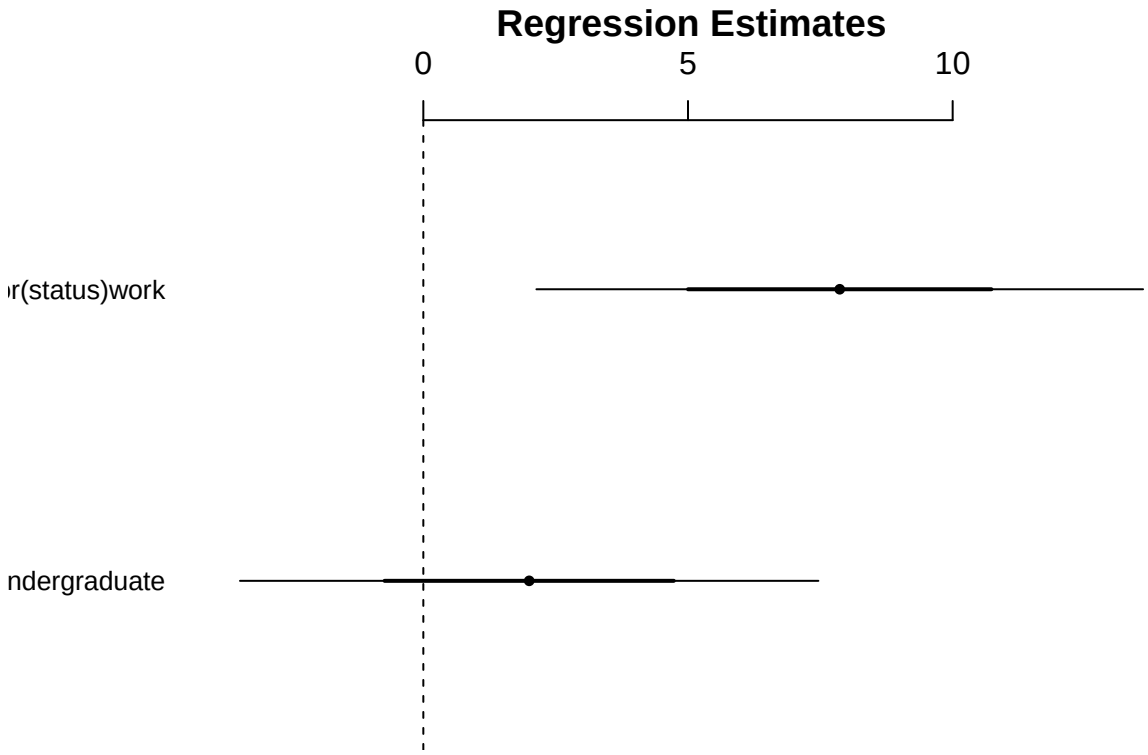


The p-value of the model is $0.04 < 0.05$, so we think there is a significant difference.
R-squared is 0.36 which means the model doesn't fit well.
The first and third plots look flat and the second plot looks like a straight line, so the current model is appropriate.

Inference (10pts)

Based on the result so far please perform statistical inference to compare the comparison of interest.

```
##                                  2.5 %    97.5 %
## (Intercept)                   1.190093  9.476574
## factor(status)undergraduate -3.859427  7.859427
## factor(status)work            1.721248 14.012085
```



I calculate the confidence interval of the slopes.
I'm 97.5% 'confident' that the true slope of 'undergraduate' lies between -3.86 and 7.86.
I'm 97.5% 'confident' that the true slope of 'work' lies between 1.72 and 14.01.
Obviously, the working group orders takeout most oftenly. And the undergraduate group is the second. Graduate group is the last.

Discussion (10pts)

Please clearly state your conclusion and the implication of the result.

My conclusion:
The working group orders takeouts most oftenly.

The model I use is:$monthly.average = 5.333 + 2.000 \times indicator(undergraduate) + 7.867 \times indicator(work)$.
Intercept. The intercept of 5.33 reflects the predicted monthly average of takeout times for graduate students.
Undergraduate. On average, the undergraduate students order takeouts 2 more times per month than graduate students.
Work. On average, the working people order takeouts 7.867 more times per month than graduate students.

Limitations and future opportunity. (10pts)

Please list concerns about your analysis. Also, please state how you might go about fixing the problem in your future study.
My concern:
The r-square is really small which means the linear regression model doesn't fit well.
Maybe we could collect more observations and do ANOVA as well.

Comments or questions

If you have any comments or questions, please write them here.