Midterm Exam

Anna Cook

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.

The data that I collected is about modes of transportation for people on the Charles River Esplanade in Boston. I have recorded whether people on the Esplanade were traveling by foot or by bike, and which direction they were going (east vs. west). The question I am trying to answer is whether the direction a person is traveling can be used to predict their mode of transportation. A subset of the raw data can be seen below. The Transport variable corresponds to 1 if the person was on foot, and 0 if the person was on a bike. The data was collected on the Charles River Esplanade on a Thursday during a 20 minute interval from 12:55 to 1:15 in the afternoon. There were 130 total people who passed by, making 130 total observations.

```
esplanade <- read_csv("Data Collection Assignment.csv")
```

```
## Warning: Missing column names filled in: 'X3' [3]
```

```
## Parsed with column specification:
## cols(
##   Transport = col_double(),
##   Direction = col_character(),
##   X3 = col_logical()
## )
```

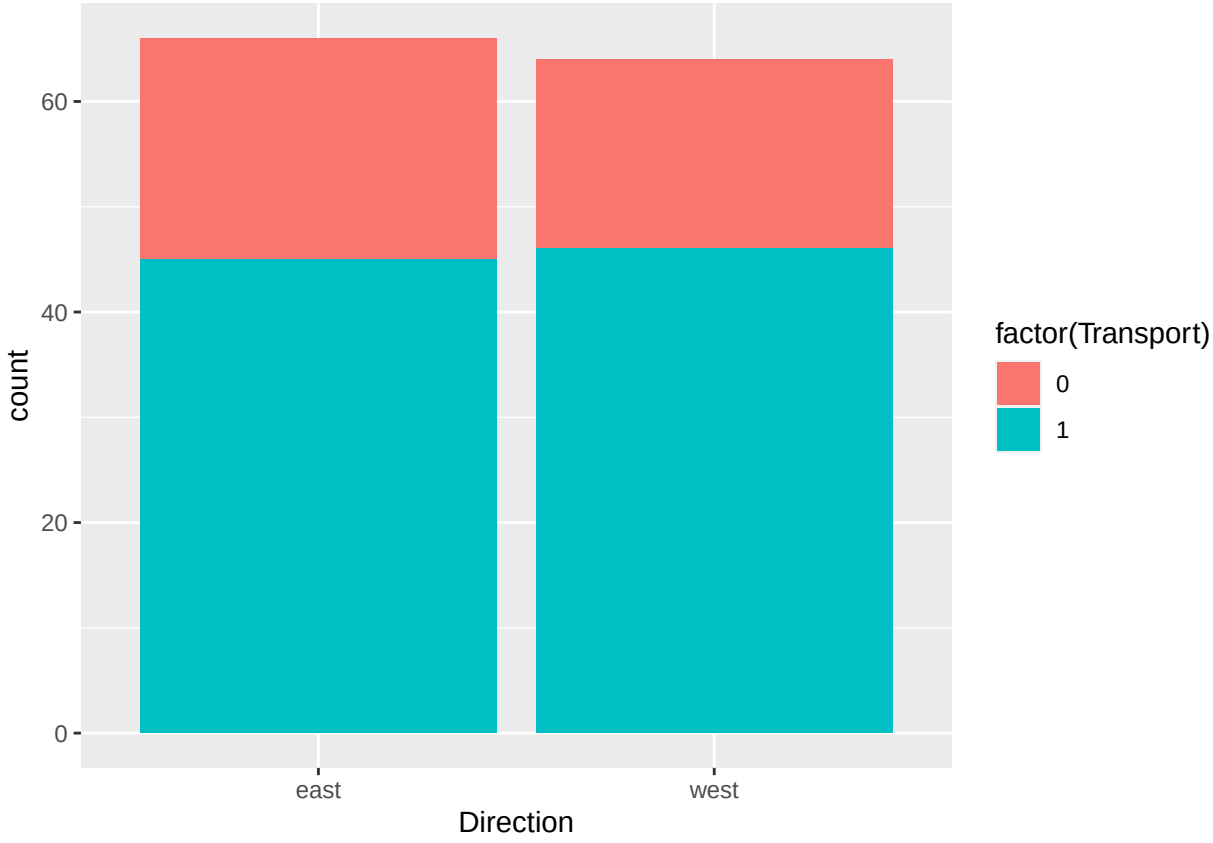
```
head(esplanade)
```

```
## # A tibble: 6 x 3
##   Transport Direction X3
##       <dbl> <chr>     <lgl>
## 1         1 west      NA
## 2         0 west      NA
## 3         0 west      NA
## 4         1 east      NA
## 5         1 west      NA
## 6         1 east      NA
```

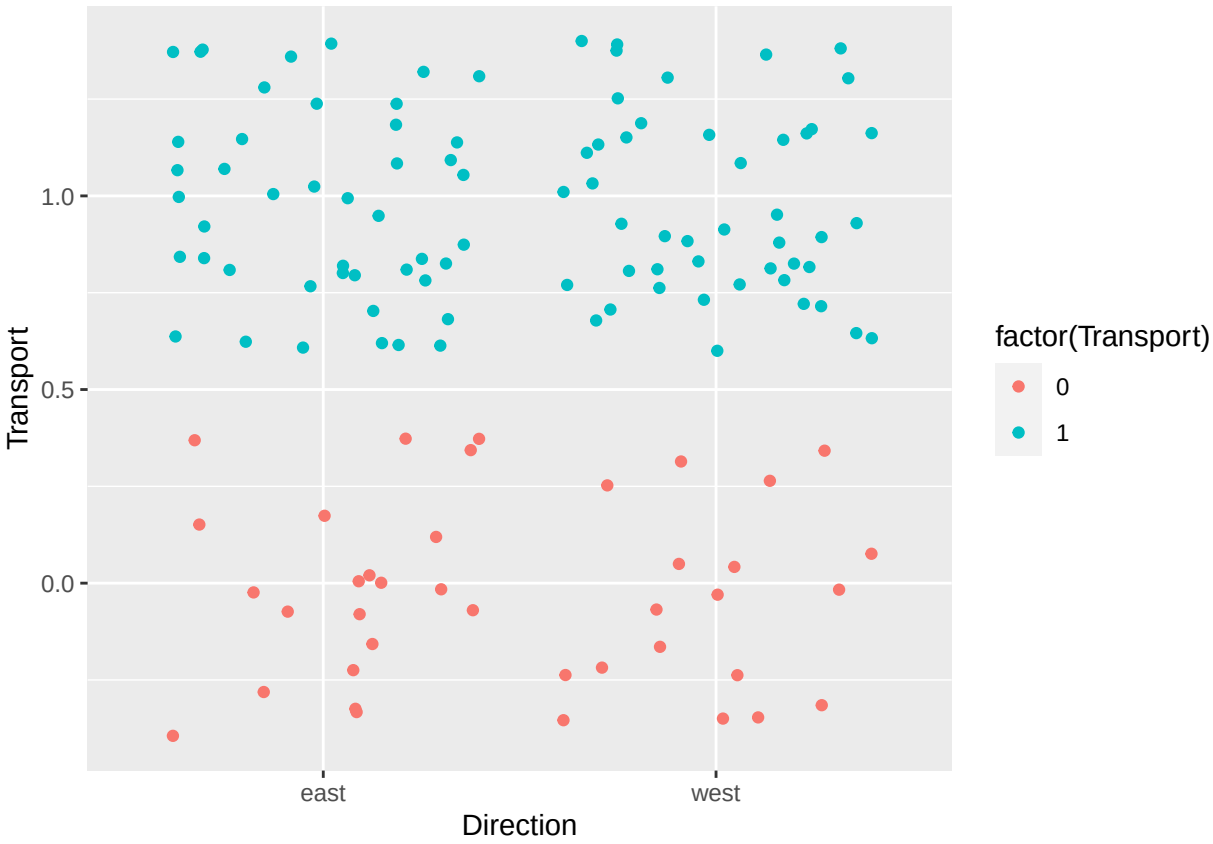
EDA (10pts)

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.

```
ggplot(data=esplanade, aes(x=Direction)) + geom_bar(data=esplanade, aes(fill=factor(Transport)))
```




```
ggplot(data=esplanade, aes(x=Direction, y=Transport, color=factor(Transport))) + geom_jitter()
```



From the plots above, there appears to be a difference in the rate of modes of transportation, with people traveling on foot (Transport=1) at higher rates than people traveling by bike (Transport=0). However, this does not seem to be dependent on the direction of travel. The number of people traveling east and west appear to be very similar. Additionally, it appears that there are similar proportions of walkers and bikers on both the east and west sides.

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.

```
pwr.t.test(n=130,d=NULL,sig.level=0.05,power=0.80,type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 130
##               d = 0.3487925
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Above I used a two sample t test for the power analysis, in order to compare outcomes for the east vs. west travellers. From this analysis, we can see that the effect size that could be detected is 0.35. This is a moderate effect size, so it appears that 130 observations is enough to detect a relationship between direction and

mode of transportation. The reason we should not use the effect size from the fitted model, is because the effect size from the fitted model tells us what effect we actually see in our data, not the effect size that can hypothetically be detected with a given power and significance level. Based on our power analysis, we could detect an effect size of 0.35, although we see a much smaller effect from the model.

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 chose to use a logistic regression model because my outcome variable is binary (either 0 or 1, for bike or foot). I am looking at proportions of people traveling on foot or on bike, rather than total counts, so logistic regression is a more appropriate choice than poisson regression.

```
model<-stan_glm(Transport ~ factor(Direction), data=esplanade, family=binomial(link="logit"), refresh =
model
```

```
## stan_glm
##  family:       binomial [logit]
##  formula:      Transport ~ factor(Direction)
##  observations: 130
##  predictors:   2
## ------
##                       Median MAD_SD
## (Intercept)           0.8    0.3
## factor(Direction)west 0.2    0.4
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
summary(model)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       binomial [logit]
##  formula:      Transport ~ factor(Direction)
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 130
##  predictors:   2
##
## Estimates:
##                         mean   sd   10%   50%   90%
## (Intercept)            0.8    0.3  0.4   0.8   1.1
## factor(Direction)west  0.2    0.4 -0.3   0.2   0.7
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 0.7    0.1  0.6   0.7   0.8
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
```

```
## MCMC diagnostics
##                         mcse Rhat n_eff
## (Intercept)             0.0  1.0  2810
## factor(Direction)west 0.0  1.0  2891
## mean_PPD                0.0  1.0  3008
## log-posterior           0.0  1.0  1263
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

We can interpret this model by taking the invlogit of the output, since the link function was logit. The probability that a person is traveling by foot given that they are going east is P(Transport = 1 | East) = invlogit(0.8) = 0.69. The probability that a person is traveling by foot given that they are going west is P(Transport=1 | West) = invlogit(0.8 + 0.2) = 0.73.

```
invlogit(0.8)
```

```
## [1] 0.6899745
```

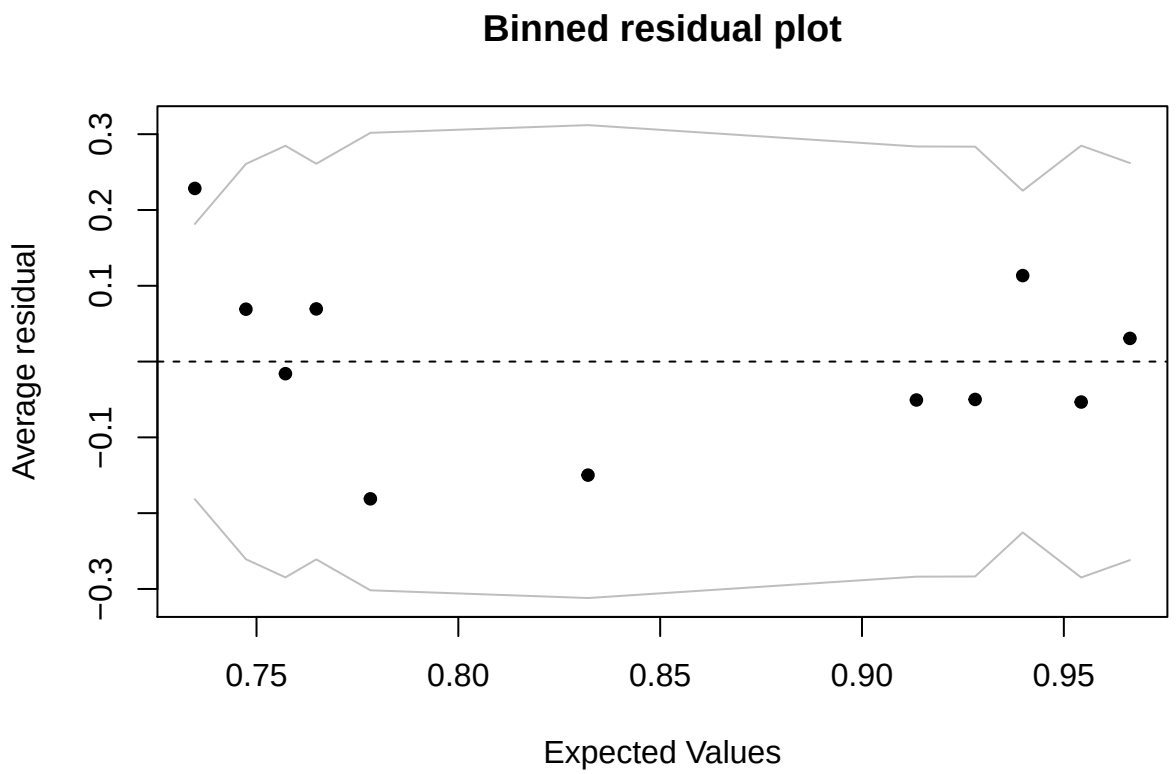
```
invlogit(0.8 + 0.2)
```

```
## [1] 0.7310586
```

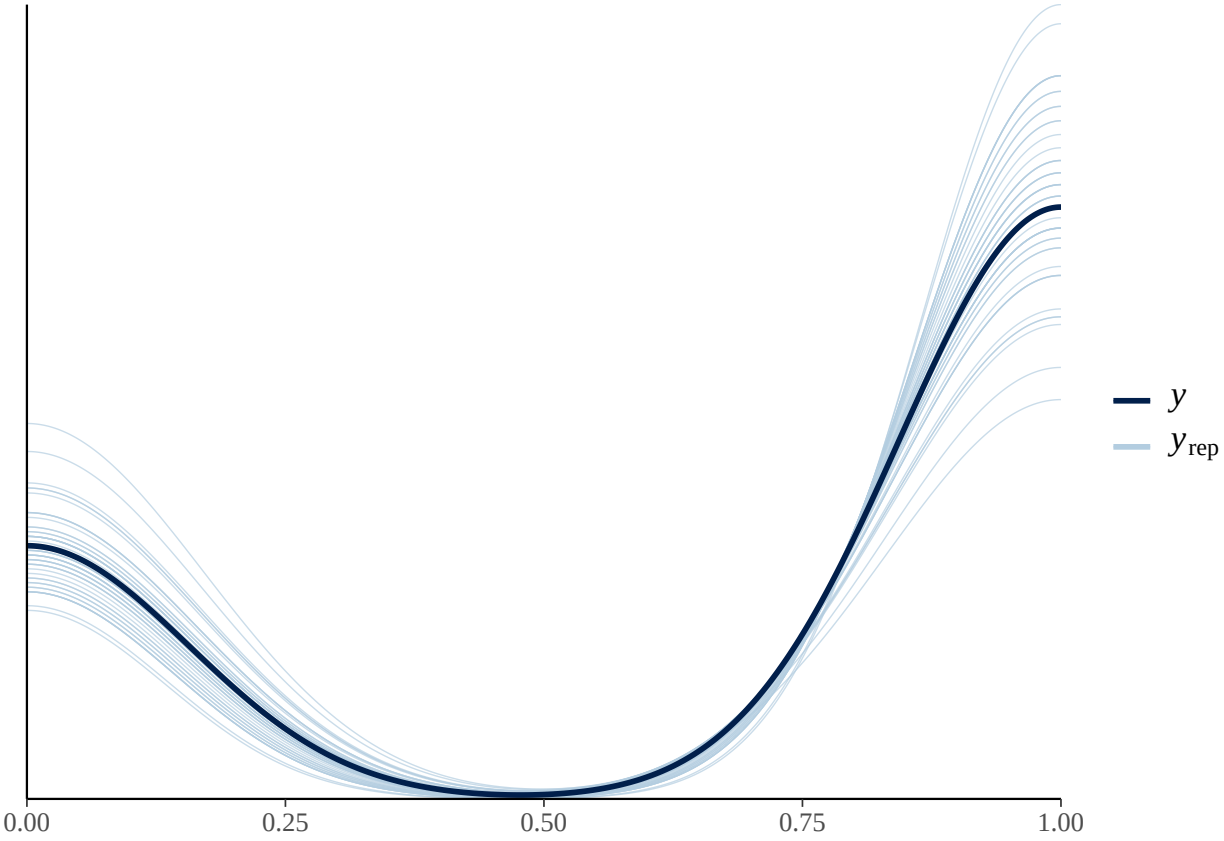
Validation (10pts)

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

```
binnedplot(jitter(predict(model)), jitter(resid(model)))
```




```
pp_check(model)
```



As you can see from the output above, the jittered binned residual plot shows residuals that are close to and evenly dispersed around zero. They are also mostly without the outer bounds, so this residual plot suggests that the model is fitting well.

Second, the posterior predictive check also shows that the model is fitting the data well, because the dark blue "y" line is closely aligned with the light blue "yrep" lines. This means that the model is fitting our data well, based on what would be predicted in other simulations.

Inference (10pts)

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

```
model
```

```
## stan_glm
##  family:       binomial [logit]
##  formula:      Transport ~ factor(Direction)
##  observations: 130
##  predictors:   2
## ------
##                       Median MAD_SD
## (Intercept)           0.8    0.3
## factor(Direction)west 0.2    0.4
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

As you can see from the model output, the coefficient for Direction is 0.2, and the MAD_SD is 0.4. Because the MAD_SD is larger than the coefficient (and therefore, the coefficient is not beyond 2 standard deviations

away from zero), this shows that this is not a statistically significant coefficient. Thus, we can conclude that direction is not a significant predictor of mode of transportation.

Additionally, when interpreting the model above, I found that the difference in probability of traveling by foot for the east and west bound travelers was only about 0.04, which is a very small difference.

Discussion (10pts)

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

Based on the above analyses, I have concluded that the direction that a person on the esplanade is traveling is not a significant predictor of the mode of transportation that they are using. That is, the proportion of travelers on foot and on bike was not significantly different for people traveling East versus those traveling West. Because this data was collected purely out of curiousity, it doesn't really have important implications. However, as someone who bikes along the Esplanade very regularly, it's interesting to know that the proportions of bikers vs. walkers doesn't significantly differ depending on a person's direction.

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.

One limitation was that the data was only collected at only one, fairly limited time period. In order to ensure a representative sample, it would be a good idea to collect data from different times of day, and different days of the week instead of just one day/time. Another option would be to record data for a longer interval of time. Additionally, data could be collected from multiple locations, so that location could be included as a predictor as well. Another opportunity for a future study would be to collect more information about the passers-by, such as gender. All of these options would allow for more predictors, and potentially a stronger model for predicting mode of transportation.

Comments or questions

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