Midterm Exam

Yuelin Jiang

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.

ANSWER: The data is count data of motorcycles passed at an intersection given the same amount of time (3 minutes). My interest of question is, whether the number of motorcycles with passengers are different between noontime and evening time. The analysis should take into account of the total motorcycles passed during this time.

```
motor <- read.csv("motorcycles.csv")
head(motor)
```

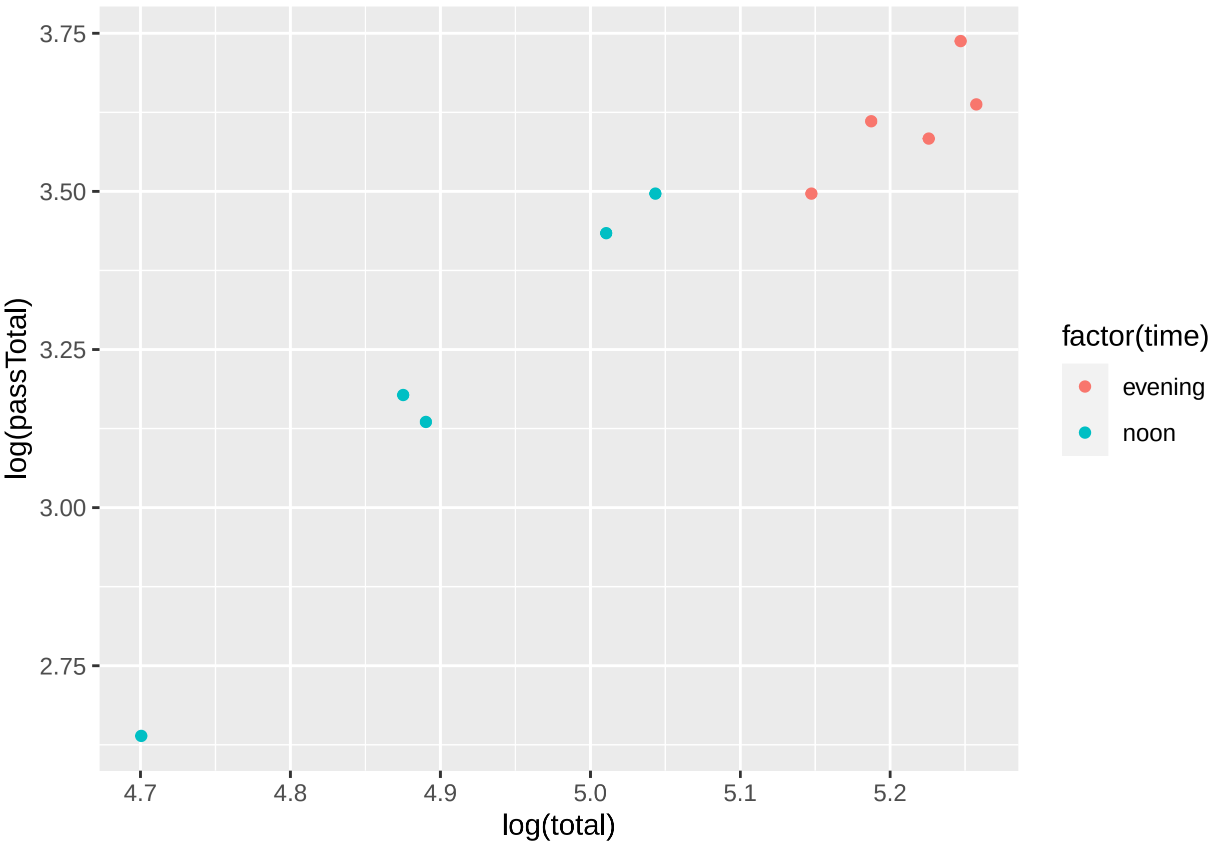
```
##   X total stuff passenger1 paseenger2orMore    time
## 1 1   110    12         12                2    noon
## 2 2   131    13         22                2    noon
## 3 3   155    12         32                1    noon
## 4 4   133    14         20                3    noon
## 5 5   150    16         26                5    noon
## 6 6   179     8         32                5 evening
```

```
motor$passTotal <- motor$paseenger2orMore + motor$passenger1
```

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.

```r
ggplot(data = motor)+
  geom_point(aes(x= log(total), y = log(passTotal) , color = factor(time) ))
```



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.

```r
pwr.t.test(n=10, d=NULL, sig.level=0.05,power=0.8, type = "one.sample")
```

```
##
##      One-sample t test power calculation
##
##               n = 10
##               d = 0.9960043
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

ANSWER: The calculated effect size(d=0.996) for my data is too large, meaning that I probably don't have a large enough sample size for my problem. I should not use this large an effect size, because even if my modeling comes out significant, my two comparison groups wouldn't differ more than 0.2 standard deviations.

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.

ANSWER: I choose to fit the data to a generalized linear model with Poisson distribution. The Poisson GLM model is appropriate because we have count data in a fixed time period, without inflated zero counts. Log link is standard when using a Poisson distribution. I also set offset to log(total), because the total motorcycles passed this intersection is the baseline for the number of motorcycles with passengers.

```
fit <- stan_glm(passTotal ~ factor(time) , family = poisson(link = "log"), offset = log(total), data = m

summary(fit, digits = 4)
```

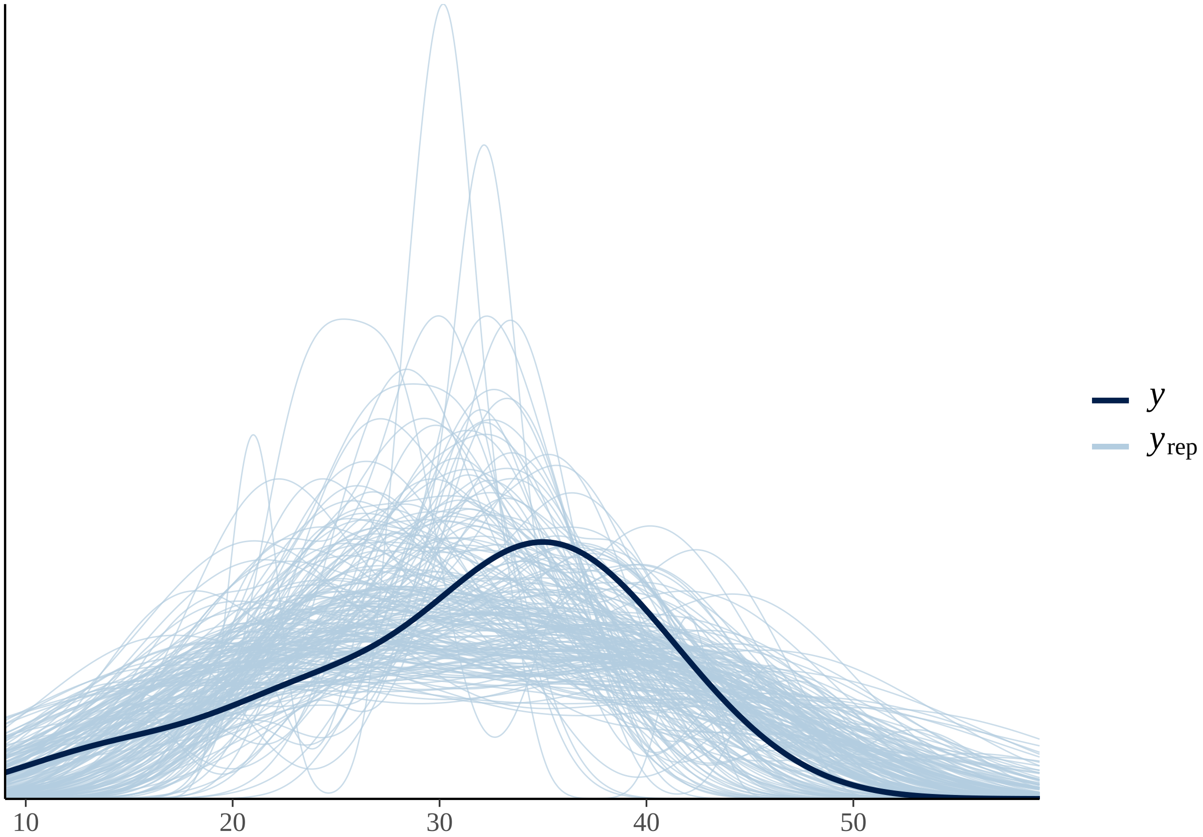
```
##
## Model Info:
##  function:     stan_glm
##  family:       poisson [log]
##  formula:      passTotal ~ factor(time)
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 10
##  predictors:   2
##
## Estimates:
##                    mean    sd      10%     50%     90%
## (Intercept)      -1.5978  0.0737 -1.6918 -1.5969 -1.5043
## factor(time)noon -0.0990  0.1156 -0.2446 -0.0979  0.0510
##
## Fit Diagnostics:
##            mean    sd      10%     50%     90%
## mean_PPD 31.1323  2.5380 27.9000 31.1000 34.4000
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
##                  mcse   Rhat   n_eff
## (Intercept)      0.0015 1.0011 2447
## factor(time)noon 0.0024 1.0011 2247
## mean_PPD         0.0469 1.0014 2932
## log-posterior    0.0250 1.0013 1644
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

Validation (10pts)

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

ANSWER: Let's first assess our model with posterior predictive checks.

```
post.fit = posterior_predict(fit)
ppc_dens_overlay(y = motor$passTotal, yrep=post.fit[1:200,])
```

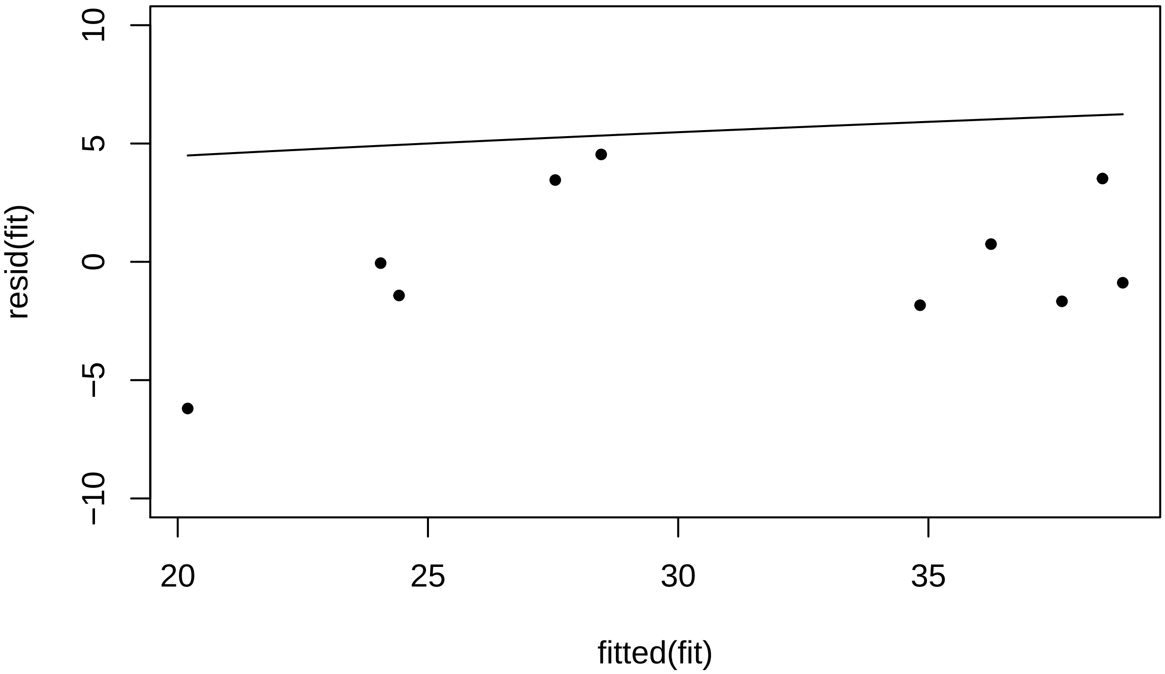


From this we can see that our model barely predicts the same as our data, with large variance too.

Let's make sure Poisson is the right GLM to use by further checking overdispersion: With a Poisson model, we expect our model's standard deviation to have a square root relationship with our fitted mean, so we can check overdispersion by plotting the residual and the curve.

We can also use residual plot and dispersiontest() to check our model.

```r
plot(fitted(fit),resid(fit),pch=20, ylim = c(-10,10))
curve(sqrt(x),add=T)
```



```
dispersiontest(fit)
```

```
##
##  Overdispersion test
##
## data:  fit
## z = -2.963, p-value = 0.9985
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##  0.3760726
```

ANSWER:

We can observe that most residuals are below the curve and dispersion test is 0.33, meaning it's unlikely we have an overdispersion problem. Therefore, we can be confident with our Poisson model.

Inference (10pts)

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

```
posterior_interval(fit)
```

```
##                          5%         95%
## (Intercept)      -1.7207291 -1.47920167
## factor(time)noon -0.2875348  0.09178898
```

ANSWER: Since the 95% confidence interval crosses zero, we cannot be certain that whether noon has more passengers than evening time.

Discussion (10pts)

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

ANSWER:

In conclusion, my analysis shows that this sample size is not large enough to draw a significant conclusion to answer my client's question. With the limited data we have, we also cannot conclude that noon time has more or fewer motorcycles with passengers than evening time, given the total motorcycles passed.

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 concerns are: 1. the sample size is too small, this can be fixed by increasing the number of observations; 2. 3min interval may be too arbitrary that does not take account into how many times the traffic light changes, which can affect the total number of motorcycles passed. This issue can be addressed by observing the intervals of traffic light change first, and then collect data according to the change of traffic light. 3. the model's prediction has large variance; this might be fixed by having larger sample size.

Comments or questions

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