Midterm Exam

Your Name

11/2/2020

```
library("readxl")
library("ggplot2")
library("pwr")
library("MASS")
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:MASS':
##
##     select
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("rstanarm")
```

```
## Loading required package: Rcpp
## This is rstanarm version 2.21.1
## - See https://mc-stan.org/rstanarm/articles/priors for changes to default priors!
## - Default priors may change, so it's safest to specify priors, even if equivalent to the defaults.
## - For execution on a local, multicore CPU with excess RAM we recommend calling
##   options(mc.cores = parallel::detectCores())
```

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.

```
#load data into Ra
growth <- read_excel("Data_Exe.xlsx")
growth <- as.data.frame(growth)
head(growth)
```

```
##   Date GrowthA GrowthB
## 1 1015      15       9
## 2 1016      29      14
## 3 1017      15      15
## 4 1018      33      17
## 5 1019      27       9
## 6 1020      16      13
```

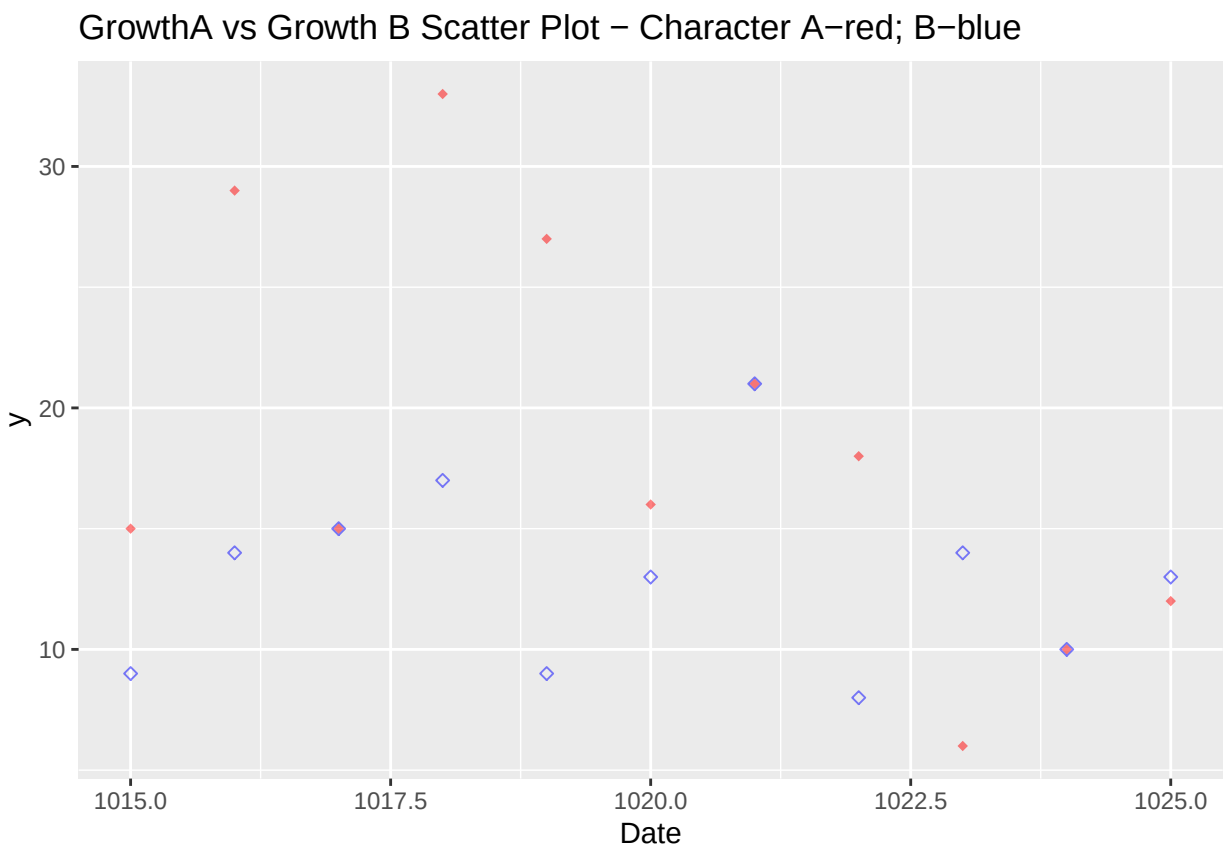
What My Data is: For this exercise, I'm collecting data of the daily growth in tags of 2 different characters from the same mobile game. The tags are counted from the same fan-work platform in order to simplify the comparison. Variable Reference:
Date - A sequence of numberic value in mmdd format.
GrowthA - The daily growth in tag number of character A (the new character)
GrowthB - The daily growth in tag number of character B (the old character)

Comparison of Interest:
If there's a difference in daily tag's growth trend between a new character and an old one.

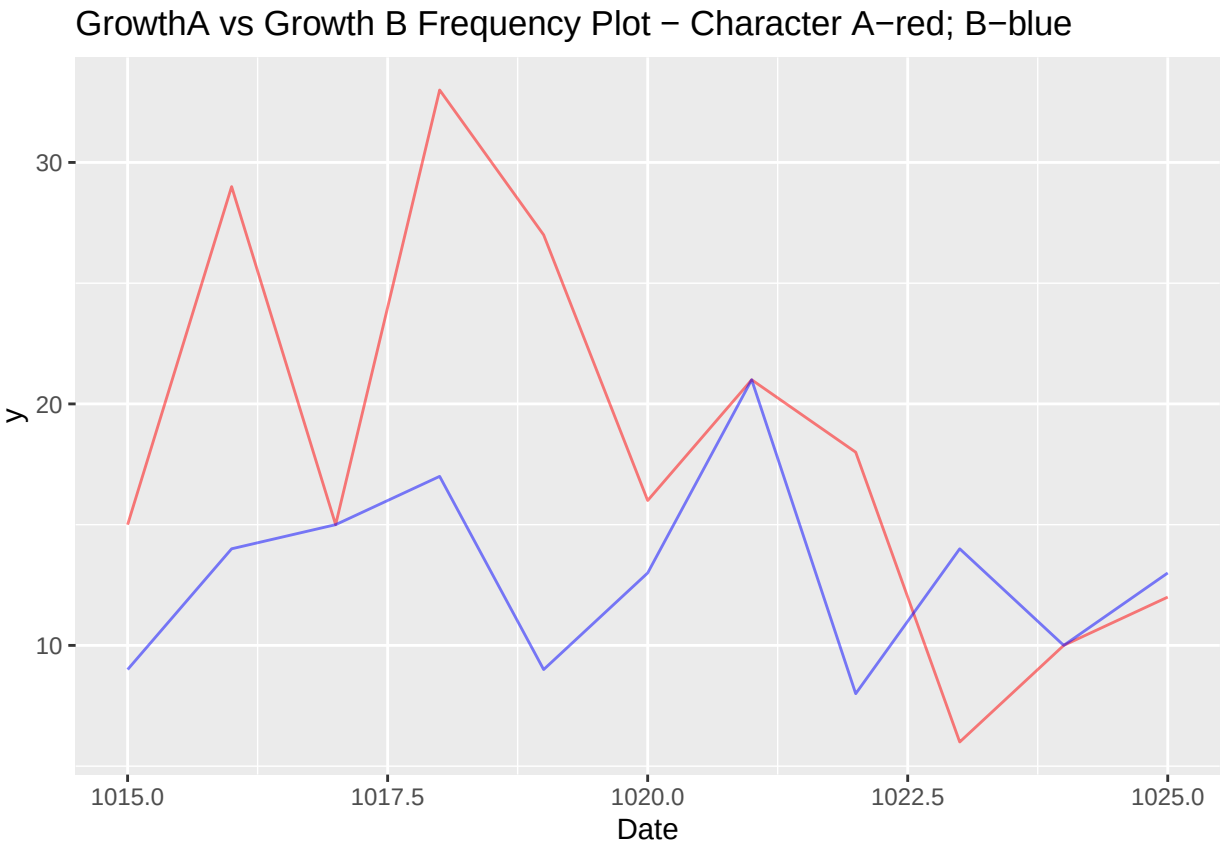
EDA (10pts)

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 = growth, aes(x = Date, y)) +
     geom_point(aes(x = Date, y = GrowthA), color=alpha('red', 0.5), shape=18) +
     geom_point(aes(x = Date, y = GrowthB),color=alpha('blue', 0.5), shape=23) +
     ggtitle("GrowthA vs Growth B Scatter Plot - Character A-red; B-blue")
```




```
ggplot(data = growth, aes(x = Date, y)) +
     geom_line(aes(x = Date, y = GrowthA), color=alpha('red', 0.5)) +
     geom_line(aes(x = Date, y = GrowthB), color=alpha('blue', 0.5)) +
     ggtitle("GrowthA vs Growth B Frequency Plot - Character A-red; B-blue")
```



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=11, power=0.8, sig.level=0.05, type = "two.sample", alternative="two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 11
##               d = 1.255951
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

The level of effect size I will be able to detect is around 1.26.

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.

```r
mean(growth$GrowthA)
```

```
## [1] 18.36364
```

```r
var(growth$GrowthA)
```

```
## [1] 70.05455
```

```r
mean(growth$GrowthB)
```

```
## [1] 13
```

```r
var(growth$GrowthB)
```

```
## [1] 15.2
```

```r
#Negative Binomial
M1 <- glm.nb(GrowthA ~ Date, data = growth, link = "identity")
M2 <- glm.nb(GrowthB ~ Date, data = growth, link = "identity")
```

After serveral comparism of fits, I decided to use identity link function. Despite the fact that log link prevents the negative data which make sense in this context, it has worse fit compared with identity link.
This could partly because our data sparse greatly, which makes it not so realistic to fit those points with a curvilinear line.

Validation (10pts)

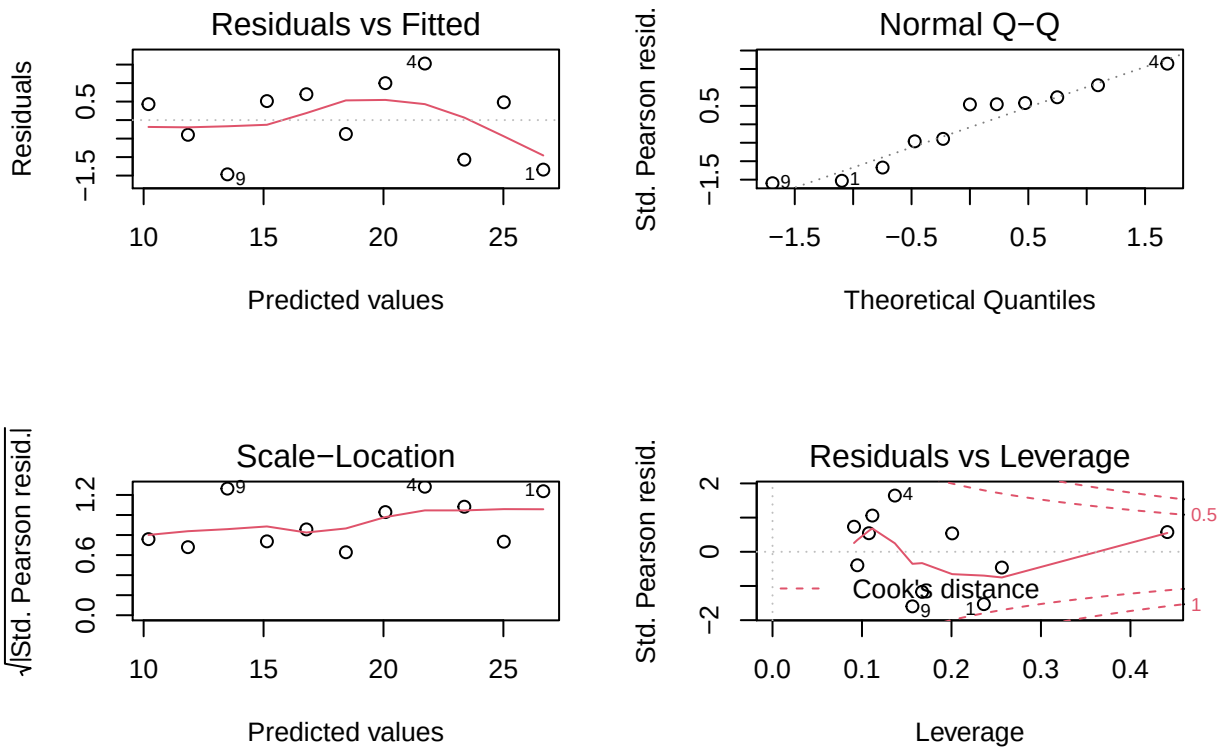
Please perform a necessary validation and argue why your choice of the model is appropriate.

```r
summary(M1)
```

```
##
## Call:
## glm.nb(formula = GrowthA ~ Date, data = growth, link = "identity",
##     init.theta = 14.41536194)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7438  -0.8048   0.4142   0.5747   1.3568
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1697.5857   607.1733   2.796  0.00518 **
## Date          -1.6462     0.5944  -2.769  0.00561 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(14.4154) family taken to be 1)
##
##     Null deviance: 17.453  on 10  degrees of freedom
## Residual deviance: 10.932  on  9  degrees of freedom
## AIC: 76.915
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  14.4
```

```
##           Std. Err.:  11.0
##
##  2 x log-likelihood:  -70.915
```

```
par(mfrow=c(2,2))
plot(M1)
```

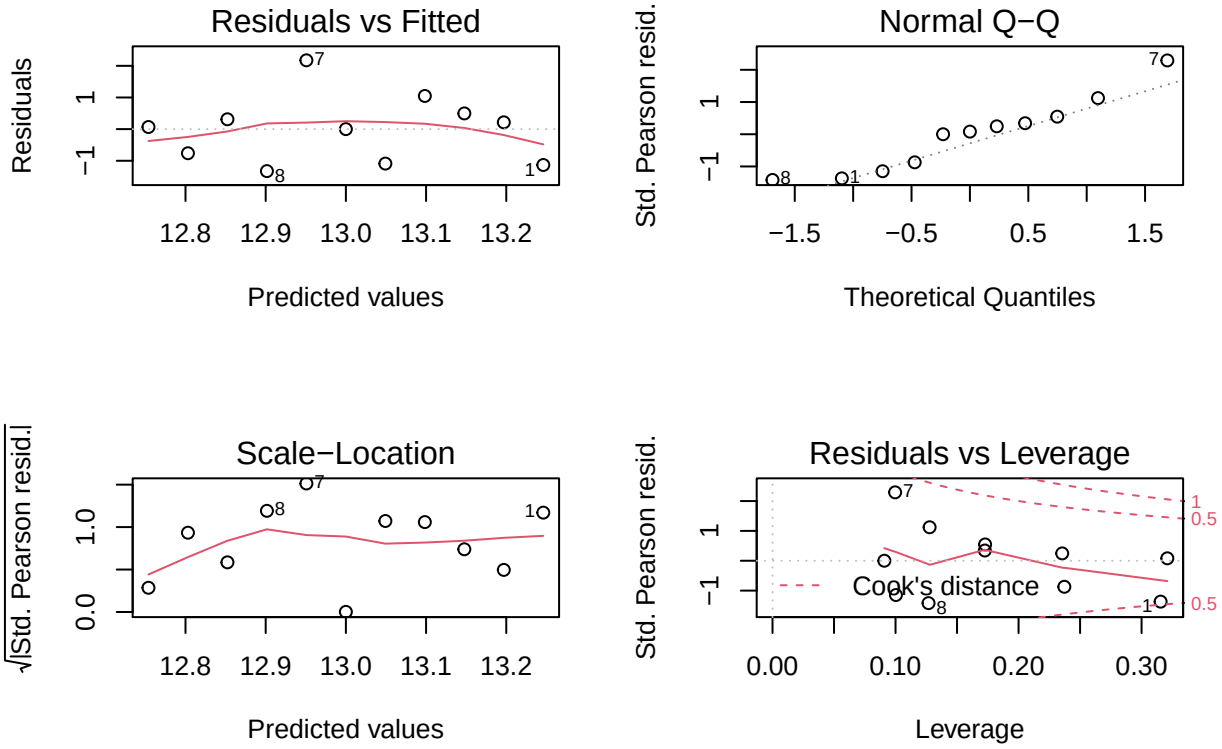



```
summary(M2)
```

```
##
## Call:
## glm.nb(formula = GrowthB ~ Date, data = growth, link = "identity",
##     init.theta = 221.9497595)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.43303  -0.97608   0.06681   0.39562   1.98248
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  63.22106  360.77041   0.175    0.861
## Date         -0.04924    0.35368  -0.139    0.889
##
## (Dispersion parameter for Negative Binomial(221.9498) family taken to be 1)
##
##     Null deviance: 10.807  on 10  degrees of freedom
## Residual deviance: 10.790  on  9  degrees of freedom
## AIC: 65.554
##
## Number of Fisher Scoring iterations: 1
##
```

```
##
##               Theta:  222
##           Std. Err.:  1676
##
##  2 x log-likelihood:  -59.554
```

```
par(mfrow=c(2,2))
plot(M2)
```



Because my dataset is discrete and I'm performing counting of events happened, the optimal choice should be Poisson distribution.
Meanwhile, since the "mean equals to variance" assumption is not fulfilled as printed above, I decided to look into negative binomial distribution with identity as the link function. Reasons are argued as in previous question.
Here I'm using glm.nb function to help create 2 models for each growth.

Inference (10pts)

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

```
confint(M1)
```

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

```
##                   2.5 %       97.5 %
## (Intercept) 428.624066 2941.1736544
## Date         -2.862605   -0.4025847
```

```
confint(M2)
```

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

```
##                    2.5 %    97.5 %
```

```
## (Intercept) -676.4142618 797.83146
## Date          -0.7690278   0.67625
```

Because I only have one varible for each set, statistical inference is hard to perform. Here I'm comparing the confidence interval between the result, it seems the two models differ from each other a lot.
The difference can be observed from summary of the models, I will explain this more in next discussion part.

Discussion (10pts)

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

Comparing the qq-plots:
Without taking into account the fact of small sample size, the two datasets display the pattern of coming from different distribution.

Analyzing the Residual vs Fitted & Scale Location: The model for GrowthA shows a better fit and a smoother magnitude change of standardized residuals.

Overall implications:
Compare with old character's tag growth, the new character's tag growth has a stronger negative correlation with date. As time goes, the new character's tag growth slow down. In contrast, the old character's tag growth remained flexible without a significant correlation with date.

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.

The biggest concern is that, the size of dataset is really small. Since the variability is really big, as shown above in the scatterplot, it will be better if I can include more observations. What's more, we have only one variable which is time.
We should be cautious that there might be other reasons for tag growth to perform this way, and the model may be more functional if we look into a longer time period.

Comments or questions

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