Midterm Exam

Jingyi Niu

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.

```
# load library
library(stringr)
library(ggplot2)
library(cowplot)
```

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.

Explain My question is that how will the roles and time I chose to play the game Honor of Kings influence my win rate, based on this question I collected all my record data from 10-23 to 10-04 to see whether there is any relationship between variables. To improve the reliability, I collect data from qualifying not arcade mode. I select qualifying and collect the KDA information as well as the time and character name.

```
data <- read.csv("C:/Users/dell/Desktop/data_collection.csv")
dim(data)
```

```
## [1] 81  7
```

After reading dataset, I extract the hour of the date from the time, which is more convenient for analysis.

```
data$date <- str_sub(data$time,1,5)
data$hours <- str_sub(data$time,7,8)
```

In order to show the data better, I output the data overview.The meanings of each field are as follows:
1. character: Hero name. There are 16 types in this dataset.
2. position: Hero positioning. There are four types of “Mid”, “Sup”, “Top” and “Bot”
3. kill: I kill the number of hero in the game.
4. death: I am killed the number of hero in the game.
5. assist: I helped my teammates kill the hero count.
6. win: Victory or not.One means victory, nine means defeat.
7. time: Game start time.
8. date: Game start date.
9. hours: Game start hours.

```
data$hours <- as.integer(data$hours)
factorVar <- c("character","position","date")
for(var in factorVar){
  data[,var] <- as.factor(data[,var])
}
summary(data)
```

```
##          character  position      kill           death            assist
##  Wang Zhaojun :22   Bot: 1   Min.   :0.000   Min.   : 0.000   Min.   : 1.00
##  Lian Po      :13   Mid:38   1st Qu.:1.000   1st Qu.: 2.000   1st Qu.: 5.00
##  Zhang Fei    : 9   Sup:22   Median :2.000   Median : 4.000   Median : 9.00
##  Ying Zheng   : 6   Top:20   Mean   :2.704   Mean   : 3.914   Mean   : 9.42
##  Mi laidi     : 5            3rd Qu.:4.000   3rd Qu.: 6.000   3rd Qu.:13.00
##  Mo Zi        : 5            Max.   :9.000   Max.   :10.000   Max.   :24.00
##  (Other)      :21
##       win              time                date        hours
##  Min.   :0.0000   Length:81          10-08  :15   Min.   : 0.00
##  1st Qu.:0.0000   Class :character   10-22  : 9   1st Qu.: 2.00
##  Median :1.0000   Mode  :character   10-17  : 8   Median :17.00
##  Mean   :0.5679                      10-23  : 8   Mean   :13.93
##  3rd Qu.:1.0000                      10-05  : 6   3rd Qu.:21.00
##  Max.   :1.0000                      10-11  : 6   Max.   :23.00
##                                      (Other):29
```

EDA (10pts)

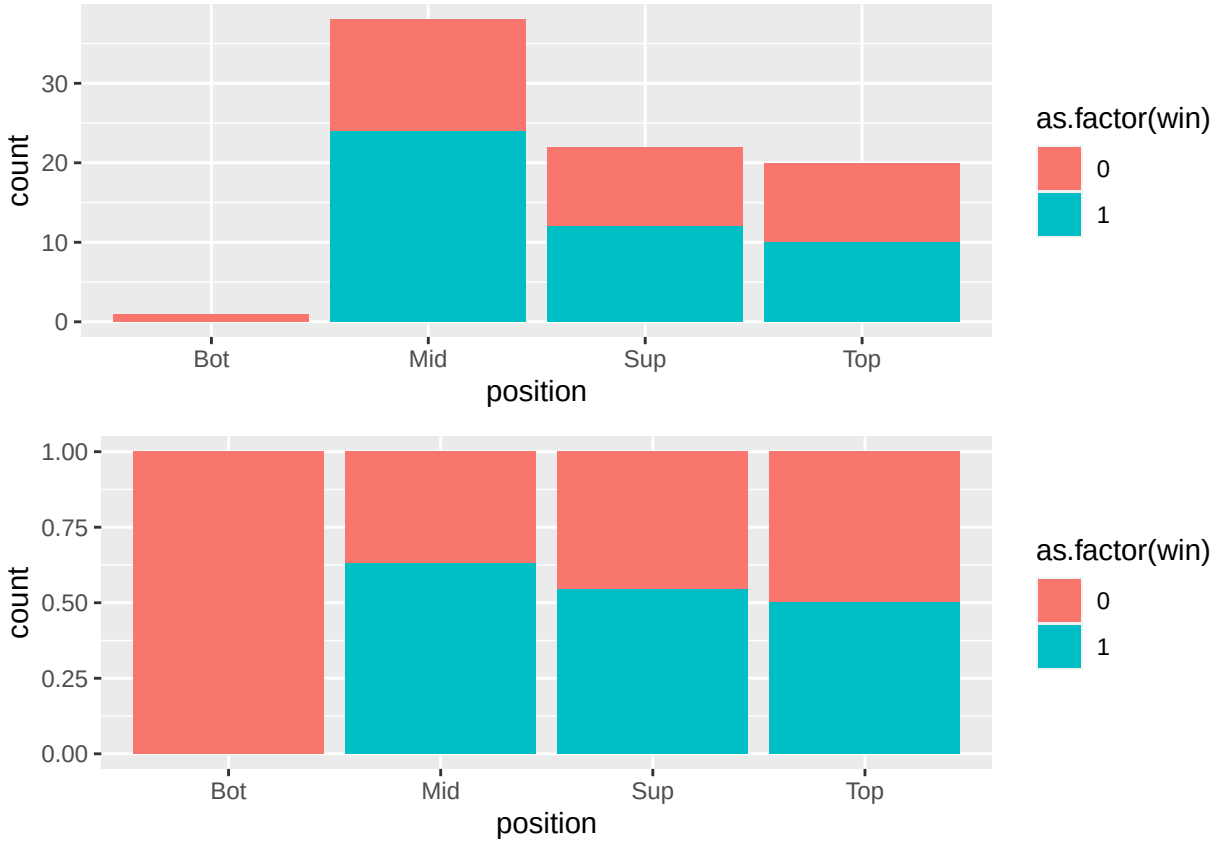
Please create one (maybe two) figure(s) that highlights the contrast of interest. Make sure you think ahead and match your figure with the analysis. For example, if your model requires you to take a log, make sure you take log in the figure as well.

First, I explored the number of wins and losses on different dates and their relative proportions.

```
p1 <- ggplot(data, aes(x=date)) +
  geom_bar(aes(fill=as.factor(win)))
p2 <- ggplot(data, aes(x=date)) +
  geom_bar(aes(fill=as.factor(win)),position='fill')
plot_grid(p1,p2,ncol=1)
```

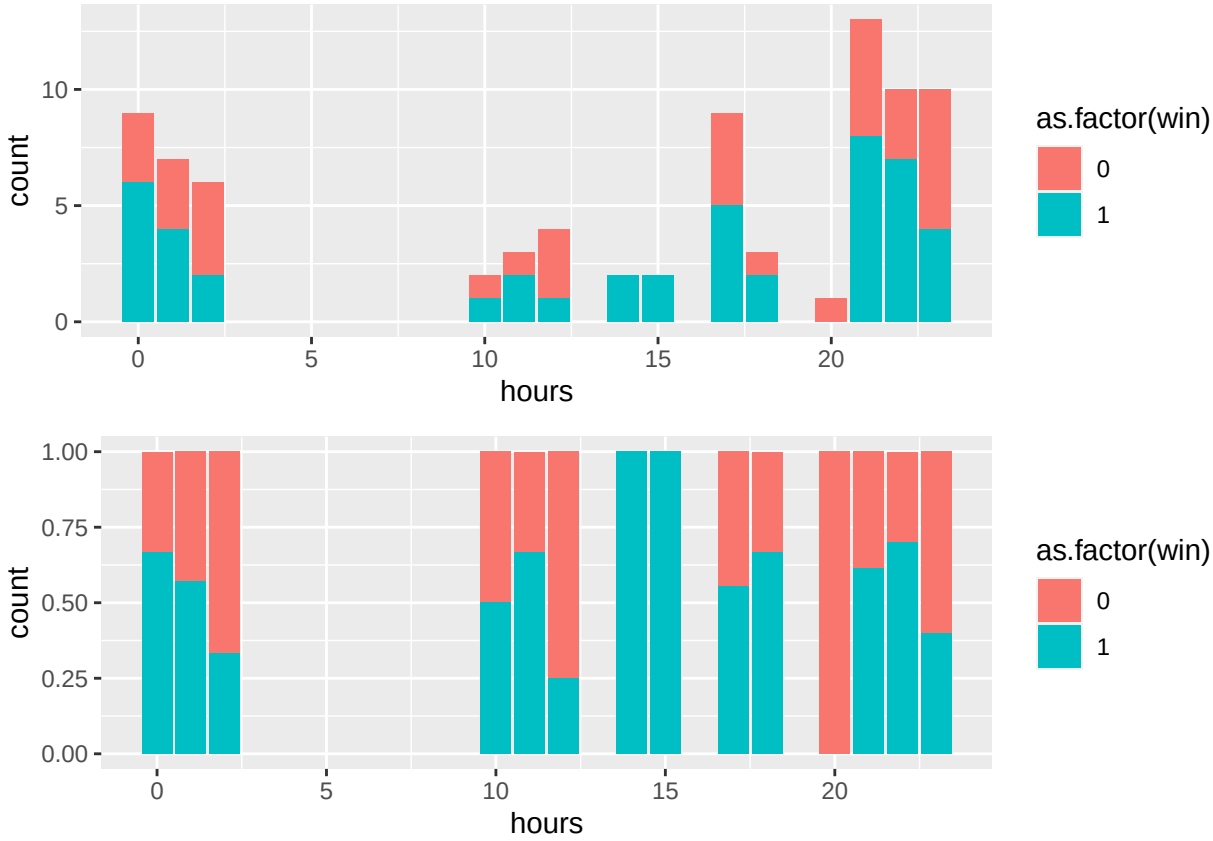
Then, I explored the number of wins and losses in different positions and their relative proportions.

```
p1 <- ggplot(data, aes(x=position)) +
  geom_bar(aes(fill=as.factor(win)))
p2 <- ggplot(data, aes(x=position)) +
  geom_bar(aes(fill=as.factor(win)),position='fill')
plot_grid(p1,p2,ncol=1)
```



Finally, I explored the number of winning and losing games and their relative proportions in different time periods.

```
p1 <- ggplot(data, aes(x=hours)) +
  geom_bar(aes(fill=as.factor(win)))
p2 <- ggplot(data, aes(x=hours)) +
  geom_bar(aes(fill=as.factor(win)),position='fill')
plot_grid(p1,p2,ncol=1)
```



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.

```
set.seed(202011)
data$gp <- runif(dim(data)[1])
dataTrain <-subset(data,data$gp <= 0.8)
dataTest <-subset(data,data$gp > 0.8)
```

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.

In order to predict the outcome, I used logistic regression model.

```
model <-glm(win~position+kill+death+assist+date+hours, data=dataTrain, family=binomial(link="logit"))
model_step<-step(model)
```

```
## Start:  AIC=61.9
```

```
## win ~ position + kill + death + assist + date + hours
##
##            Df Deviance    AIC
## - date     12   40.516 56.516
## <none>          21.895 61.895
## - hours     1   24.568 62.568
## - assist    1   27.281 65.281
## - position  3   32.921 66.921
## - kill      1   41.645 79.645
## - death     1   60.001 98.001
##
## Step:  AIC=56.52
## win ~ position + kill + death + assist + hours
##
##            Df Deviance    AIC
## - assist    1   41.827 55.827
## - position  3   46.381 56.381
## <none>          40.516 56.516
## - hours     1   44.977 58.977
## - kill      1   55.075 69.075
## - death     1   76.621 90.621
##
## Step:  AIC=55.83
## win ~ position + kill + death + hours
##
##            Df Deviance    AIC
## <none>          41.827 55.827
## - position  3   48.589 56.589
## - hours     1   45.540 57.540
## - kill      1   57.722 69.722
## - death     1   76.955 88.955
```

```
dataTrain$LOG_pred <- predict(model_step, newdata=dataTrain, type="response")
dataTest$LOG_pred <- predict(model_step, newdata=dataTest, type="response")
```

Validation (10pts)

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

Next, I verify the regression results of the model.

```
loglikelihood<-function(y, py) {
  pysmooth<-ifelse(py==0, 1e-12, ifelse(py==1, 1-1e-12, py))
  sum(y * log(pysmooth) + (1-y)*log(1 -pysmooth))
}
accuracyMeasures<-function(pred, truth, name="model",value=0.5) {
  dev.norm<--2*loglikelihood(as.numeric(truth), pred)/length(pred)
  ctable<-table(truth==1,pred=(pred>value))
  accuracy <-sum(diag(ctable))/sum(ctable)
  precision <-ctable[2,2]/sum(ctable[,2])
  recall <-ctable[2,2]/sum(ctable[2,])
  f1 <-2*precision*recall/(precision+recall)
  data.frame(model=name, accuracy=accuracy,precision=precision,recall=recall, f1=f1, dev.norm)
```

```
}
accuracyMeasures(dataTest$LOG_pred,dataTest$win)
```

```
##   model accuracy precision recall  f1 dev.norm
## 1 model      0.6 0.5833333  0.875 0.7 1.830504
```

Inference (10pts)

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

This part outputs the model results, and the regression results can be seen intuitively.

```
summary(model_step)
```

```
##
## Call:
## glm(formula = win ~ position + kill + death + hours, family = binomial(link = "logit"),
##     data = dataTrain)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.08634  -0.49040   0.07374   0.35362   2.80727
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -13.42239 2399.54514  -0.006  0.99554
## positionMid   13.40424 2399.54497   0.006  0.99554
## positionSup   16.01106 2399.54489   0.007  0.99468
## positionTop   14.30079 2399.54499   0.006  0.99524
## kill           1.04219    0.35148   2.965  0.00303 **
## death         -1.04730    0.27506  -3.808  0.00014 ***
## hours          0.09966    0.05633   1.769  0.07684 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 89.974  on 65  degrees of freedom
## Residual deviance: 41.827  on 59  degrees of freedom
## AIC: 55.827
##
## Number of Fisher Scoring iterations: 15
```

Discussion (10pts)

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

From the regression results:
1. Kill count is a positive factor in winning 2. The number of deaths has a negative effect on success

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. 1. The amount of data is too small to conduct a comprehensive study.
2. Multicollinearity exists between data, such as the position of class variables.
3. Hours variables should be treated as categories. Due to the small amount of data, dividing the training set and the test set cannot meet the requirements of stratified sampling.

Comments or questions

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