Midterm Exam

Zhitian Liu

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 (http://www.bu.edu/cas/files/2017/02/GRS-Academic-Conduct-Code-Final.pdf).

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.

```
data=read.csv("https://raw.githubusercontent.com/lzt666666/MA678midterm-exam/main/data_collection.csv")
head(data)
```

	battle.number <int>	battle.deck <chr>	result <chr>	king.stower_destroyed <int>	king.stower_destroy <int>	ar
1	1	A	lose	0	0	
2	2	A	win	0	0	
3	3	A	win	0	0	
4	4	A	win	0	0	
5	5	A	win	0	0	
6	6	B	win	0	0	

6 rows | 1-7 of 10 columns

```
#My friend Cherry and I are both super fans of Clash Royale, a mobile strategy video game devel
oped and published by Supercell. Cherry has a card deck and she's really proud of it. I prepar
ed two different card decks; I wonder which card deck is better to use to defeat Cherry in a fr
iendly battle. Now lets look at the dataset.
```

explain of variables: Result: win, lose or tie game. king'stower_destroyed: 0-my king's tower wasn't destroyed by Cherry; 1- my king's tower was destroyed by Cherry king'stower_destroy: 0–I destroy Cherry's king's tower; 1—I didn't destroy Cherry's king's tower. arenatower_destroyed: number of my arena tower destroyed by Cherry. arenatower_destroy: number of Cherry's arena tower destroyed by me sudden death mode: Is the game go to the sudden death mode? Yes or no. time_used: time used for each battle. measured in second.

Questions want to be solved: Which battle deck has a better chance to win? Can I predict the result of a game ?

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.

```
summary(data)
```

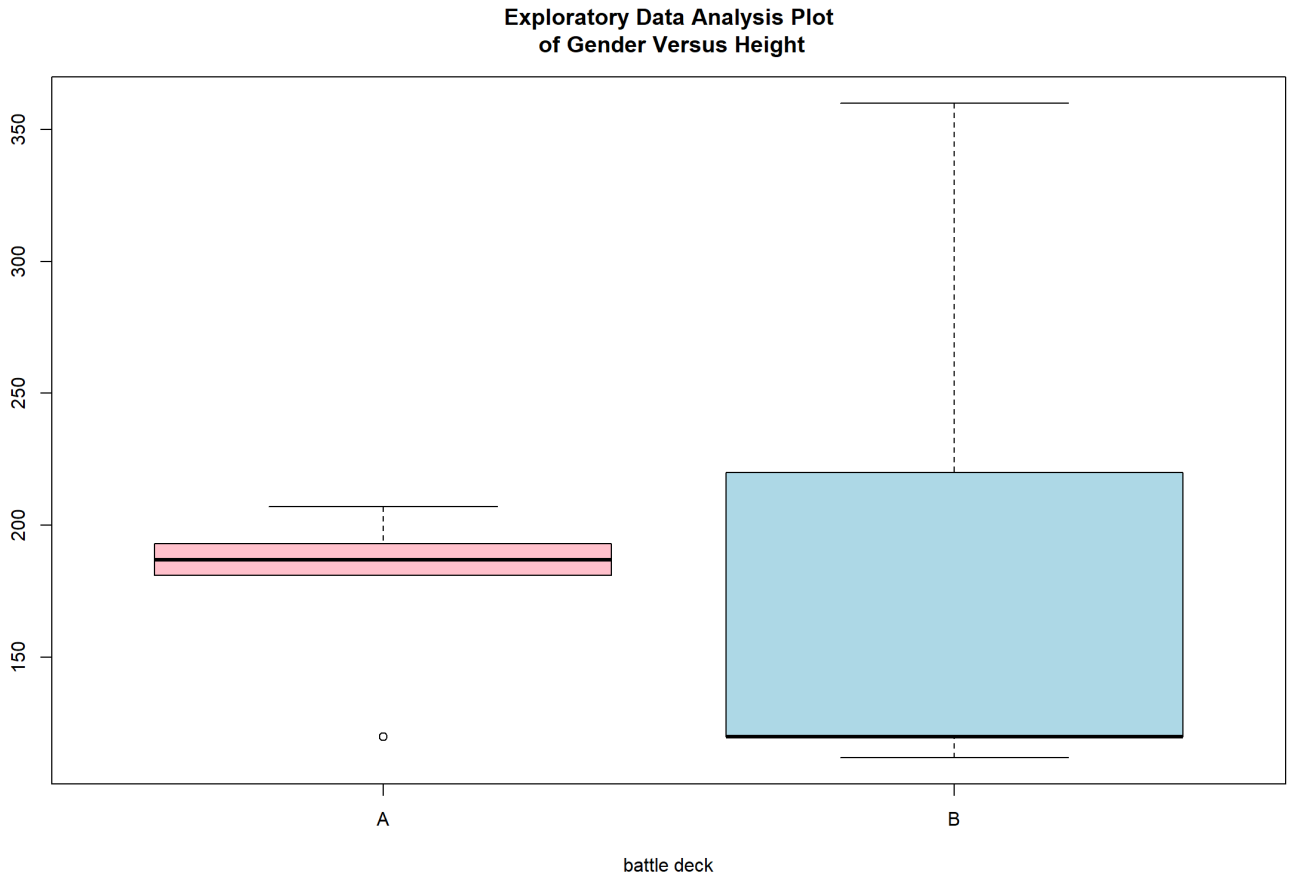
```
##  battle.number   battle.deck          result          king.stower_destroyed
##  Min.   : 1.00   Length:10          Length:10          Min.   :0
##  1st Qu.: 3.25   Class :character   Class :character   1st Qu.:0
##  Median : 5.50   Mode  :character   Mode  :character   Median :0
##  Mean   : 5.50                                         Mean   :0
##  3rd Qu.: 7.75                                         3rd Qu.:0
##  Max.   :10.00                                         Max.   :0
##  king.stower_destroy arenatower_destroyed arenatower_destroy sudden.death.mode
##  Min.   :0           Min.   :0.0          Min.   :0          Length:10
##  1st Qu.:0           1st Qu.:0.0          1st Qu.:1          Class :character
##  Median :0           Median :0.5          Median :1          Mode  :character
##  Mean   :0           Mean   :0.6          Mean   :1
##  3rd Qu.:0           3rd Qu.:1.0          3rd Qu.:1
##  Max.   :0           Max.   :2.0          Max.   :2
##    time_used
##  Min.   :112.0
##  1st Qu.:120.0
##  Median :184.0
##  Mean   :182.0
##  3rd Qu.:203.5
##  Max.   :360.0
```

```
##data cleaning
#Firstly, we need to clean the data, we noticed there are two columns with all the value to be
 0, It doesn't give any information, so we delete this two columns.
data=select(data,battle.deck,result, arenatower_destroyed,arenatower_destroy,sudden.death.mode,
time_used )
#Because the outcome is Ternary, It's hard to analyze, so I group the battle which I lost and t
ied together as "NOT WIN" to make the outcome bin
data$result[which(data$result=="lose"|data$result=="tie")]="not_win"

#We are going to use logistic regression later, so I tranformed the outcome "Result" to 0-1 for
mat.
data$result=ifelse(data$result[]=="win",1,0)

#EDA
#In order to better understand the data, We are going to do some EDA. FIrstly,we need to see th
e relationship between result and different battle deck.
A=c(0,1,1,1,1);B=c(1,0,0,0,1)
data_eda=data.frame(A,B)
data_eda
```

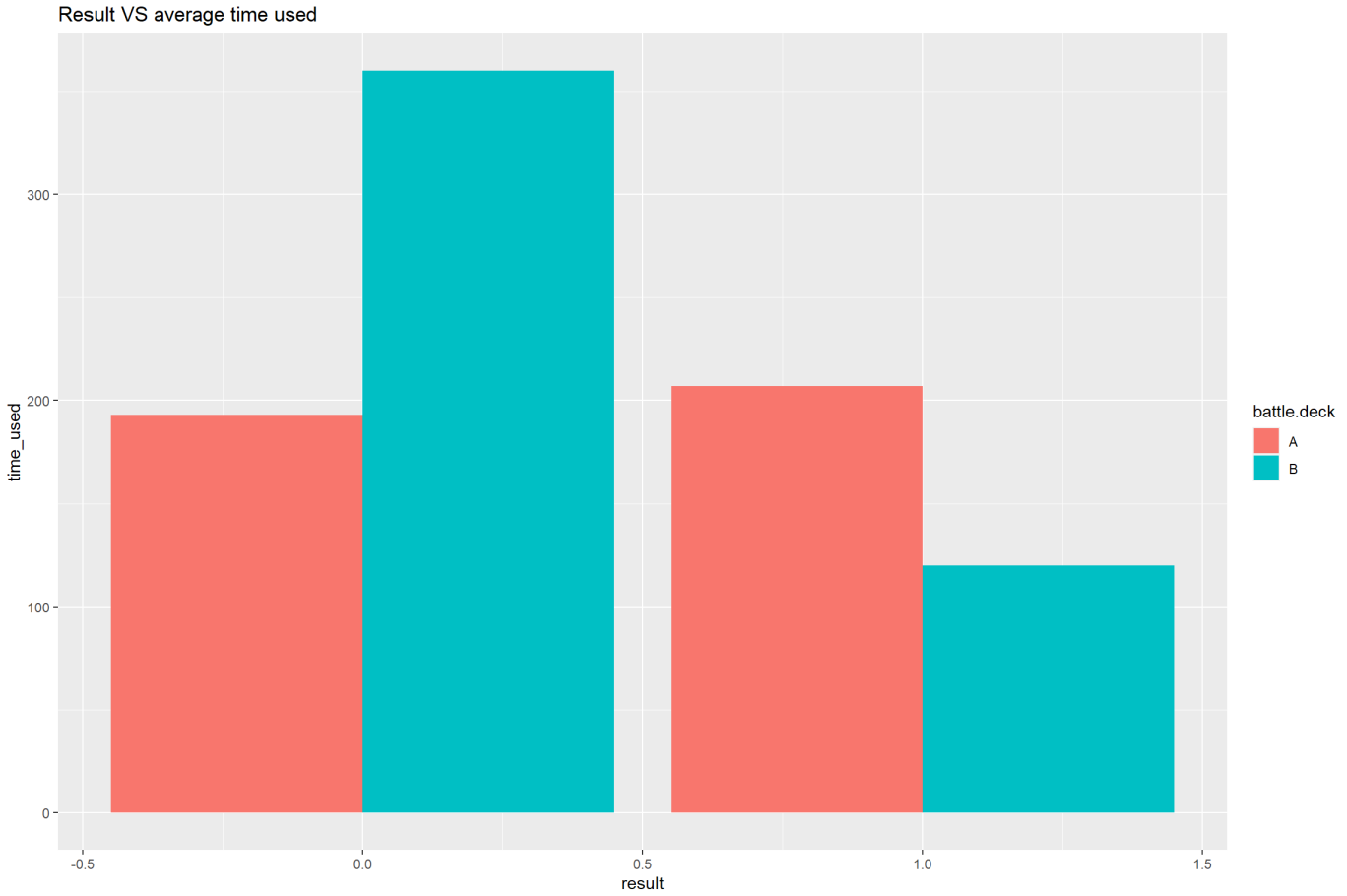
A <dbl>	B <dbl>
0	1
1	0
1	0
1	0
1	1

5 rows

```
#As we can see, use battle deck A is more likely to win the game. We need to prove that in the
 later analysis.
#The following boxplot shows the relationship between battle deck and time, Obviously, the game
with battle deck A generally ends in 180-200 seconds, while the game with battle deck B fluctua
tes greatly.
boxplot(time_used~battle.deck,data=data,xlab="battle deck",ylab="",col=c("pink","lightblue"),
        main="Exploratory Data Analysis Plot\n of Gender Versus Height")
```




```
#The bar chart below shows the time spent on two battle decks when the results of winning and l
osing are different. For battle deck A, regardless of winning or losing, the average game is of
ten relatively stable at around 200, but for battle deck B, the winning game is often much shor
ter than the losing game.
ggplot(data, aes(result, time_used)) + geom_bar(aes(fill = battle.deck), position = "dodge",sta
t='identity')+ggtitle("Result VS average time used")
```



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.

```
library(pwr)
pwr.t.test(n=5,power=0.8,sig.level=0.05, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 5
##               d = 2.024439
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
#After running a two-sample t test, the effect siz of the data is 2.02, which is quite weird, I think this is due to my data size is too small.
#If I use d=0.5 as my objective effect size, let's see how many observations do I need
pwr.t.test(n=NULL,d=0.5, power=0.8,sig.level=0.05, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.76561
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
#If we want effect size to be normal, I need 64 sample size for each group.
```

To see what sample size is appropriate for our data. we use

$$|\mu_a - \mu_b|/\sigma$$

to calculate the effect size.

```
d1=abs(mean(data$time_used[1:5])-mean(data$time_used[6:10]))/sd(data$time_used)
#
pwr.t.test(n=NULL,d=d1,power=0.8,sig.level=0.05, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 1134.595
##               d = 0.1176744
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
#It seems that we need 1135 observations for each group in ou r data.
```

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.

```
#The outcome of my data is binary, so I'm pretty sure I need to fit a logistic regression mode
l. I decide to use glm function with link "logit".

#After checking my data again, I found that it is hard to include the columns "arenatower_desto
ry" and "arenatower_destroyed" to the fitted model. When I tried to add this two columns to the
model, they Seriously affect the accuracy of the model (Displayed by AIC and p value).
#I thought it was because the two columns of data were categorical, so I transfer them to dummy
variable.

#create dummy variables


data$arenatower_destroy%>%as.factor()
```

```
##  [1] 0 1 1 1 1 2 1 1 0 2
## Levels: 0 1 2
```

```
data_dummy=fastDummies::dummy_cols(data,select_columns =c("arenatower_destroy","arenatower_dest
royed") )
#fit a logistic model with dummy variable
fit1=glm(result~time_used+arenatower_destroy_1+arenatower_destroyed_0+arenatower_destroyed_2, d
ata=data_dummy, family=binomial(link="logit"))
summary(fit1)
```

```
##
## Call:
## glm(formula = result ~ time_used + arenatower_destroy_1 + arenatower_destroyed_0 +
##     arenatower_destroyed_2, family = binomial(link = "logit"),
##     data = data_dummy)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
## -0.00008   0.00000   0.00000   0.00000   0.00005   1.17741  -0.00003   0.00000
##        9        10
## -1.17741   0.00005
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)               32.2738 18281.6366   0.002    0.999
## time_used                 -0.2689   152.3470  -0.002    0.999
## arenatower_destroy_1      25.1876 28318.8615   0.001    0.999
## arenatower_destroyed_0    18.5233 18761.5851   0.001    0.999
## arenatower_destroyed_2   -19.8588 38056.8395  -0.001    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13.4602  on 9  degrees of freedom
## Residual deviance:  2.7726  on 5  degrees of freedom
## AIC: 12.773
##
## Number of Fisher Scoring iterations: 20
```

```
#However, the p-value is still close to 1, which is bad. I'm quite confused about this, I guess another reason for this maybe is this these two columns of data directly determine the outcome. As now, I don't know how to deal with it, so I choose not to use them in this project.
#model selection
fit2=glm(result~battle.deck+time_used, data=data, family=binomial(link="logit"))
summary(fit2)
```

```
##
## Call:
## glm(formula = result ~ battle.deck + time_used, family = binomial(link = "logit"),
##     data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.67535  -0.07974   0.32537   0.75531   1.04496
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  12.17130   15.24618   0.798    0.425
## battle.deckB -4.76066    6.15717  -0.773    0.439
## time_used    -0.05725    0.07836  -0.731    0.465
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13.4602  on 9  degrees of freedom
## Residual deviance:  8.1835  on 7  degrees of freedom
## AIC: 14.184
##
## Number of Fisher Scoring iterations: 8
```

```
#After filter different parameters into the model (including interactions and log scale...I didn't show the process in the code), according to p-value and AIC, I decided to use a single variable-time_used in the model besides the indicator variable "battle.deck".
```

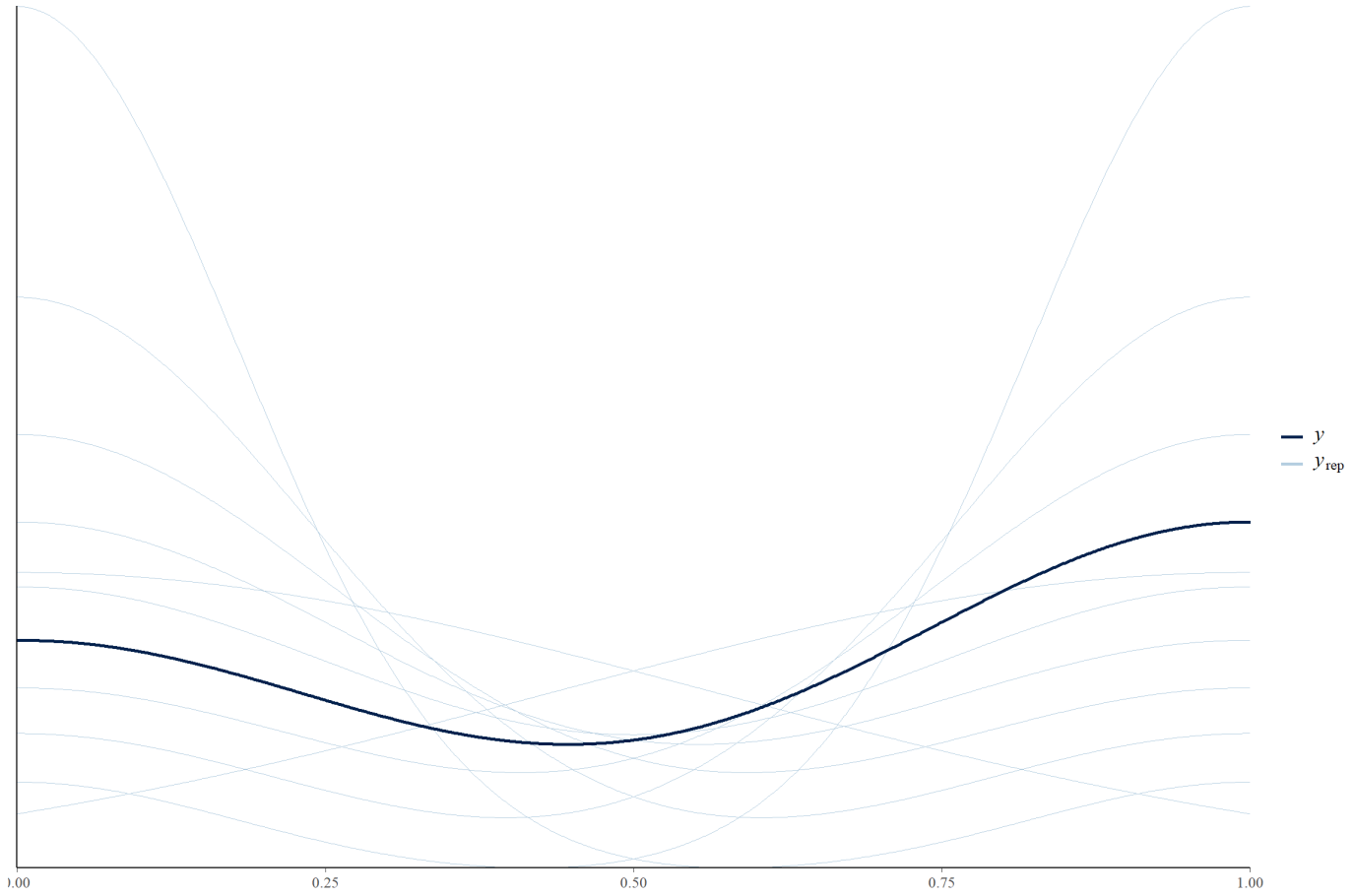
Validation (10pts)

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

```
#As we can see from the above output, the fitting model isn't great, the p-value is around 0.5 for each variable, but it is understandable since I only have 10 observations.
#We now use Leave One Out (LOO) Cross Validation to check our model. In order to use Loo function, we need to refit the same model using "stan_glm"
fit3=stan_glm(result~battle.deck+time_used, data=data, family=binomial(link="logit"),refresh=0)
print(loo(fit3))
```

```
##
## Computed from 4000 by 10 log-likelihood matrix
##
##          Estimate  SE
## elpd_loo     -7.2 2.3
## p_loo         2.7 1.0
## looic        14.3 4.5
## ------
## Monte Carlo SE of elpd_loo is NA.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     8     80.0%   461
##  (0.5, 0.7]   (ok)       1     10.0%   554
##    (0.7, 1]   (bad)      1     10.0%   2060
##    (1, Inf)   (very bad) 0      0.0%   <NA>
## See help('pareto-k-diagnostic') for details.
```

```
#From the output, we noticed that All Pareto k estimates are ok (k < 0.7), and elpd_loo is clos
e to 0.

#We can also do a Posterior Predictive Checks to see the fit of our model.
result_rep=posterior_predict(fit3)
ppc_dens_overlay(data$result,result_rep) + scale_y_continuous(breaks=NULL)
```



```
#From the below output, We noticed that although the light blue line is sparse(which I don't kn
ow why, I thought it will generate 4000 simulations), We can observed there are many similariti
es between the original data and the predicted value pattern.

#Hence, I think this model is already very good with so few observations.
```

Inference (10pts)

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

we calculate the 95% confidence interval (CI) using

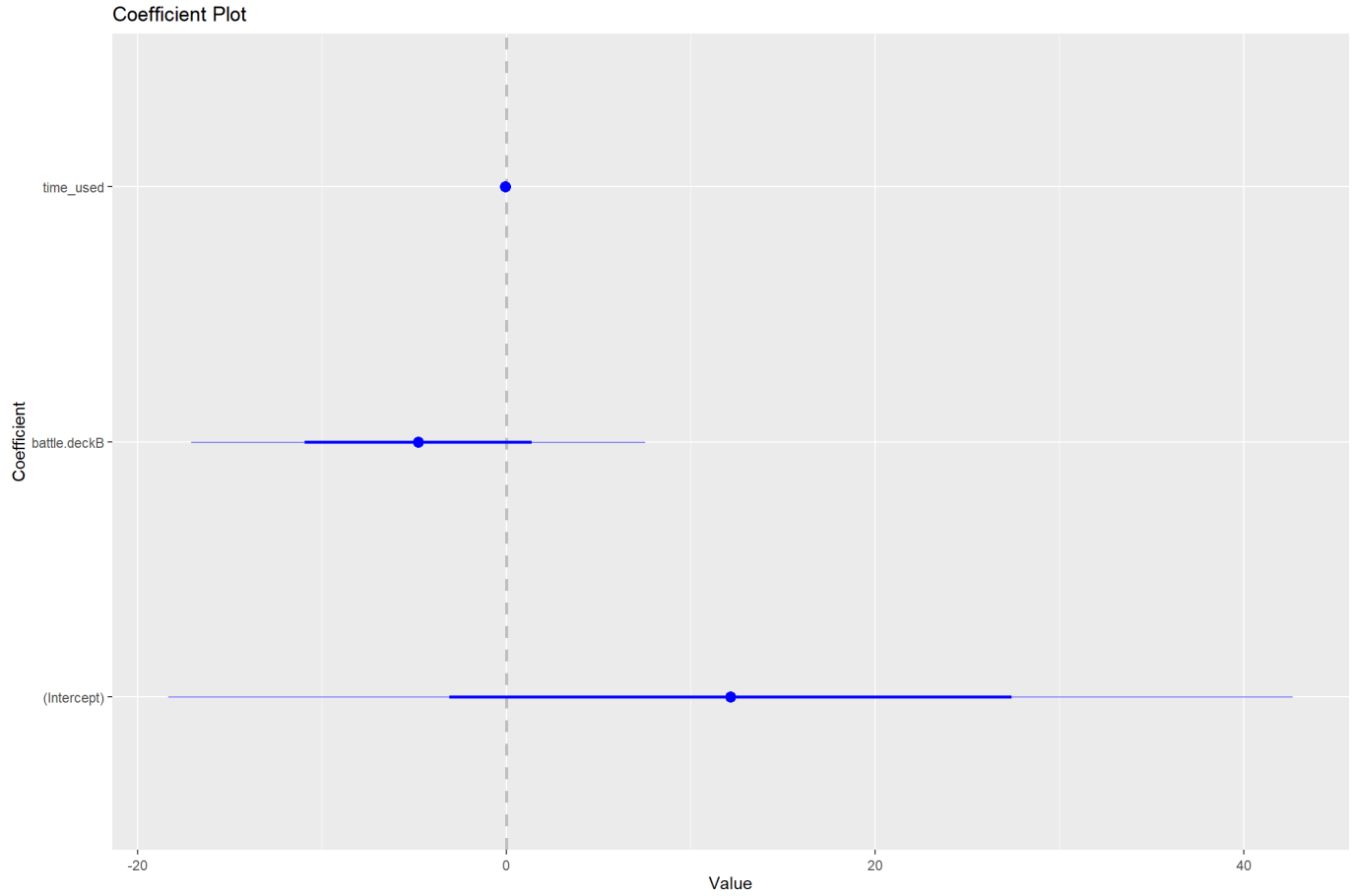
$$\hat{\beta}_i \pm 2s.e.$$

```
#confidence interval for battle.deckB
coef=summary(fit2)$coefficients
CI_deckB=c((coef[2]-2*coef[2,2]),(coef[2]+2*coef[2,2]))
#confidence interval for time_used
CI_time_used=c((coef[3]-2*coef[3,2]),(coef[3]+2*coef[3,2]))
CI_time_used;CI_deckB
```

```
## [1] -0.2139782  0.0994710
```

```
## [1] -17.075001   7.553671
```

```
#The 95%confidence interval for variable time_used is (-0.21,0.1); 95%confidence interval for v
ariable deck B is (-17.07,7.55)
#we can also use function confinct()
#We can plot the CI for the model we choosed before
coefplot(fit2, vertical=FALSE, var.las=1, frame.plot=TRUE)
```




```
# Also we can do a Wald test to test the overall effect of variable battle.deck.
wald.test(b = coef(fit2), Sigma = vcov(fit2), Terms = 2)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 0.6, df = 1, P(> X2) = 0.44
```

```
#The p-value is 0.44,chi-squared test statistic is 0.6 with 1 degrees of freedom, The effect of Battle deck is not significant, which is acceptable because I only have 10 observations in my data.
```

I'm not sure how to compare the comparison of interest

Discussion (10pts)

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

The Questions we asked before: Which battle deck has a better chance to win? Can I predict the result of a game ?

```
#First of all: the answer to the previous question is: By using deck A, I will have a better ch
ance to win the game. And yes, I can predict the result of the game according to which battle d
eck I used and the time I will used in a battle. Although the model is not in a good fit.

#I obtain the answers by interpreting the coefficient of the model:
#The coefficient of battle.deckB is -4.7, which means with same time of game, carry the battle
 deck A, versus carry deck B, changes the log odds of result by -4.7. By using the below formul
a, I know deck A has a better chance to win.
```

$$p = \frac{exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}$$

.

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.

```
#There are 2 main problems occurs in the study.

#The first problem that arises is that I failed to solve how to fit the the 2 categorical varia
bles into the model, and this 2 variables are directly related to the outcome.

# The second problem is that the sample size of this data is too small, resulting in each test
 not being significant, EDA and model are very unconvincing, and the real relationship between
 variables are very foggy.

# In addition to these two problems, there are some minor loopholes. For example, I don’t unde
rstand the power analysis problem very well. I know the sample size is not enough, But I don’t
 know which sample size generated from test should I use, 1134 or 64 for each group?

# Another point is that I found that this model is not meaningful in making predictions. I need
the time i used in a game to make predictions about the outcome, but at that time, the game is
 already overed.


# In the following study, generally speaking, I have to understand the principles behind these
 r functions more deeply, and also frequently review the previous knowledge. I have already for
gotten some of the knowledge of logistic regression.

# The things need to be done right away is to figure out why my model becomes very bad as soon
 as the two columns of categorical data are added.
```

Comments or questions

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