Midterm Exam

Zhe Yu

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.

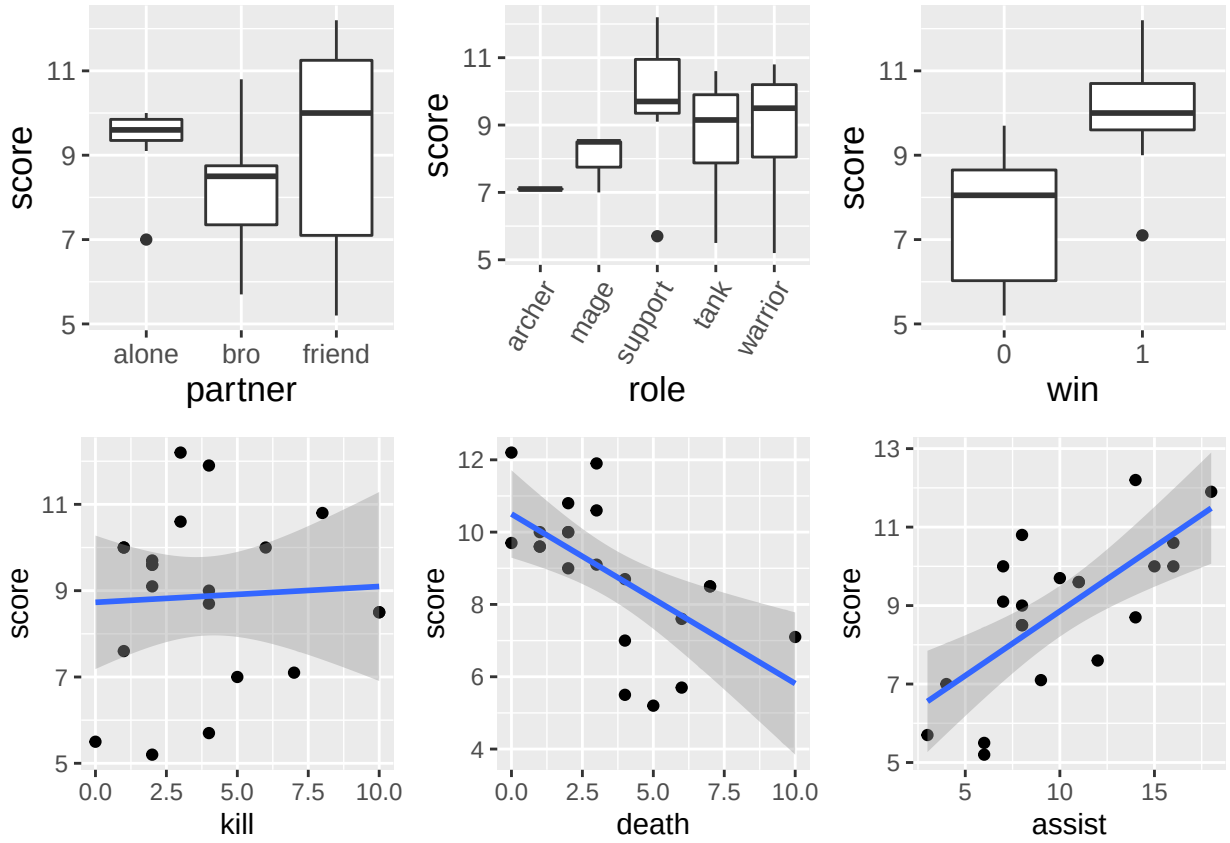
Data Description (10pts)

The data was collected from a Moba (Multiplayer Online Battle Arena) game, just like the game League Of Legend, and the data contain the results of each game I played with different people. I tried to collect all the data in a closest period, and I collected the information of winning, the performance score, the role I played and so on. So the question is: how factors (partner, number of kill...) affect my performance score? For in each game, there was an unique score assessing my performance in that game.
Column names:
`Partner`: bro means playing with my brother; alone means I play alone; friends means playing with my friends.
`win`: win this game or not.
`kill`: the number of killing.
`death`: the number of death.
`assist`: the number of assist ally killing enemy.
`score`: after each game, system will give a score to measure performance.
`role`: the role played in each game.

```
#import the data
game <- read.csv("game.csv",header = T)
game$partner <- as.factor(game$partner)
game$role <- as.factor(game$role)
game$win <- as.factor(game$win)
game <- game[-c(15,18,22),]
game$num <- as.factor(as.character(rep(1:7,3)))
```

EDA (10pts)

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



In the plots, it seems that when I play with my friends, I play support or I win the game that I will have higher score. The number of killing has a small impact on the score for the slope is small. But the number of death and assistance have negative and positive impact separately on score.

Power Analysis (10pts)

```
pwr.anova.test(k=3 , n =7 , sig.level = 0.05, power = 0.8)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 3
##               n = 7
##               f = 0.7379229
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

```
pwr.anova.test(k=3 , f =0.25 , sig.level = 0.05, power = 0.8)
```

```
##
```

```
##      Balanced one-way analysis of variance power calculation
##
##               k = 3
##               n = 52.3966
##               f = 0.25
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

For I only have 7 observations in each group, so the effect size is 0.74 which is a little bit bigger. When using variance comparison, 0.25 always be the appropriate value for ES. And if I want to reach that value, I have to set around 52 observations in each group. If the effect size is too big, it will make the Type M error.

Modeling (10pts)

```
#complete pooling
fit_g <- lm(score~partner+role+win+kill+death+assist,data = game)
summary(fit_g)
```

```
##
## Call:
## lm(formula = score ~ partner + role + win + kill + death + assist,
##     data = game)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78728 -0.27814 -0.02834  0.24488  1.29392
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.003273   1.824448   3.290  0.00814 **
## partnerbro    -0.007817   0.612949  -0.013  0.99008
## partnerfriend -0.271008   0.511556  -0.530  0.60783
## rolemage      -0.521544   1.294080  -0.403  0.69541
## rolesupport    0.514687   1.455536   0.354  0.73098
## roletank       0.151585   1.420368   0.107  0.91712
## rolewarrior   -0.235576   1.363695  -0.173  0.86630
## win1           0.399528   0.589321   0.678  0.51319
## kill           0.428398   0.114081   3.755  0.00375 **
## death         -0.463681   0.158782  -2.920  0.01529 *
## assist         0.260338   0.068368   3.808  0.00344 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6953 on 10 degrees of freedom
## Multiple R-squared:  0.9364, Adjusted R-squared:  0.8729
## F-statistic: 14.73 on 10 and 10 DF,  p-value: 0.0001051
```

```
fit_h=step(fit_g,direction="backward")
```

```
## Start:  AIC=-8.85
## score ~ partner + role + win + kill + death + assist
##
##           Df Sum of Sq     RSS      AIC
```

```
## - partner  2    0.1928  5.0269 -12.0242
## - role     4    1.2783  6.1124 -11.9182
## - win      1    0.2222  5.0563  -9.9018
## <none>                  4.8341  -8.8454
## - death    1    4.1224  8.9565   2.1050
## - kill     1    6.8168 11.6509   7.6280
## - assist   1    7.0095 11.8436   7.9725
##
## Step:  AIC=-12.02
## score ~ role + win + kill + death + assist
##
##          Df Sum of Sq     RSS      AIC
## - role    4    1.6488  6.6757 -14.0670
## - win     1    0.3096  5.3364 -12.7692
## <none>                 5.0269 -12.0242
## - death   1    6.6940 11.7208   3.7537
## - assist  1    7.6499 12.6768   5.4002
## - kill    1    8.5662 13.5931   6.8658
##
## Step:  AIC=-14.07
## score ~ win + kill + death + assist
##
##          Df Sum of Sq     RSS      AIC
## - win     1    0.6284  7.3041 -14.1780
## <none>                 6.6757 -14.0670
## - kill    1   15.1412 21.8169   8.8014
## - assist  1   16.3537 23.0294   9.9373
## - death   1   18.6378 25.3134  11.9231
##
## Step:  AIC=-14.18
## score ~ kill + death + assist
##
##          Df Sum of Sq    RSS     AIC
## <none>                 7.304 -14.178
## - kill    1    18.718 26.022  10.502
## - death   1    25.398 32.702  15.301
## - assist  1    25.631 32.935  15.450
```

```
summary(fit_h)
```

```
##
## Call:
## lm(formula = score ~ kill + death + assist, data = game)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1107 -0.3491 -0.1037  0.4052  1.5965
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.22863    0.54189  11.494 1.94e-09 ***
## kill         0.38979    0.05906   6.600 4.50e-06 ***
## death       -0.51616    0.06713  -7.688 6.24e-07 ***
## assist       0.29197    0.03780   7.724 5.87e-07 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6555 on 17 degrees of freedom
## Multiple R-squared:  0.904,  Adjusted R-squared:  0.887
## F-statistic: 53.34 on 3 and 17 DF,  p-value: 7.352e-09
```

First I tried to put all the variables into the model, but seems some variables are not significant. So I used `step` function to drop some variables. And the remaining variables are `kill`, `death`, `assist`. all the categorical variables had been dropped. But I will still try to use multilevel analysis to see if there are some difference between group.

```
#varying intercept
fit2 <- lmer(score~kill+death+assist+(1|partner),data = game)
round(coef(fit2)$partner,digits = 2)
```

```
##        (Intercept) kill death assist
## alone         6.23 0.39 -0.52   0.29
## bro           6.23 0.39 -0.52   0.29
## friend        6.22 0.39 -0.52   0.29
```

```
fit3 <- lmer(score~kill+death+assist+(1|role),data = game)
round(coef(fit3)$role,digits = 2)
```

```
##         (Intercept) kill death assist
## archer         6.24  0.4 -0.51   0.29
## mage           6.08  0.4 -0.51   0.29
## support        6.32  0.4 -0.51   0.29
## tank           6.10  0.4 -0.51   0.29
## warrior        6.14  0.4 -0.51   0.29
```

```
fit5 <- lmer(score~kill+death+assist+(1|win),data = game)
round(coef(fit5)$win,digits = 2)
```

```
##   (Intercept) kill death assist
## 0        6.22 0.38  -0.5   0.28
## 1        6.37 0.38  -0.5   0.28
```

With the results above, it is clear that partner indeed has no impact on the score for with different partner the intercept are the same. And for `role` group and `win` group, there are some differences but they are small so I will not take any of them into consideration.

```
#varying intercept and scope.
fit4 <- lmer(score~kill+death+assist+(1+role|partner),data = game)
```

```
## boundary (singular) fit: see ?isSingular
```

```
coef(fit4)
```

```
## $partner
##          rolemage rolesupport    roletank rolewarrior (Intercept)      kill
## alone  -0.1000039 -0.07154272  0.08955112   0.1731528    6.461353 0.3271198
## bro    -0.1178446 -0.08430606  0.10552716   0.2040434    6.461353 0.3271198
## friend  0.5974694  0.42742967 -0.53502007  -1.0344950    6.461353 0.3271198
##             death    assist
## alone  -0.4568363 0.2830024
## bro    -0.4568363 0.2830024
## friend -0.4568363 0.2830024
##
## attr(,"class")
```

```
## [1] "coef.mer"
```

```
fit6 <- lmer(score~kill+death+assist+(1+win|partner),data = game)
```

```
## boundary (singular) fit: see ?isSingular
```

```
coef(fit4)
```

```
## $partner
##          rolemage rolesupport    roletank rolewarrior (Intercept)      kill
## alone  -0.1000039 -0.07154272  0.08955112   0.1731528    6.461353 0.3271198
## bro    -0.1178446 -0.08430606  0.10552716   0.2040434    6.461353 0.3271198
## friend  0.5974694  0.42742967 -0.53502007  -1.0344950    6.461353 0.3271198
##             death    assist
## alone  -0.4568363 0.2830024
## bro    -0.4568363 0.2830024
## friend -0.4568363 0.2830024
##
## attr(,"class")
## [1] "coef.mer"
```

here I got the warning means that the model is too complex. So both of them are not appropriate.

Validation (10pts)

After trying different type of models, linear regression model stands out.

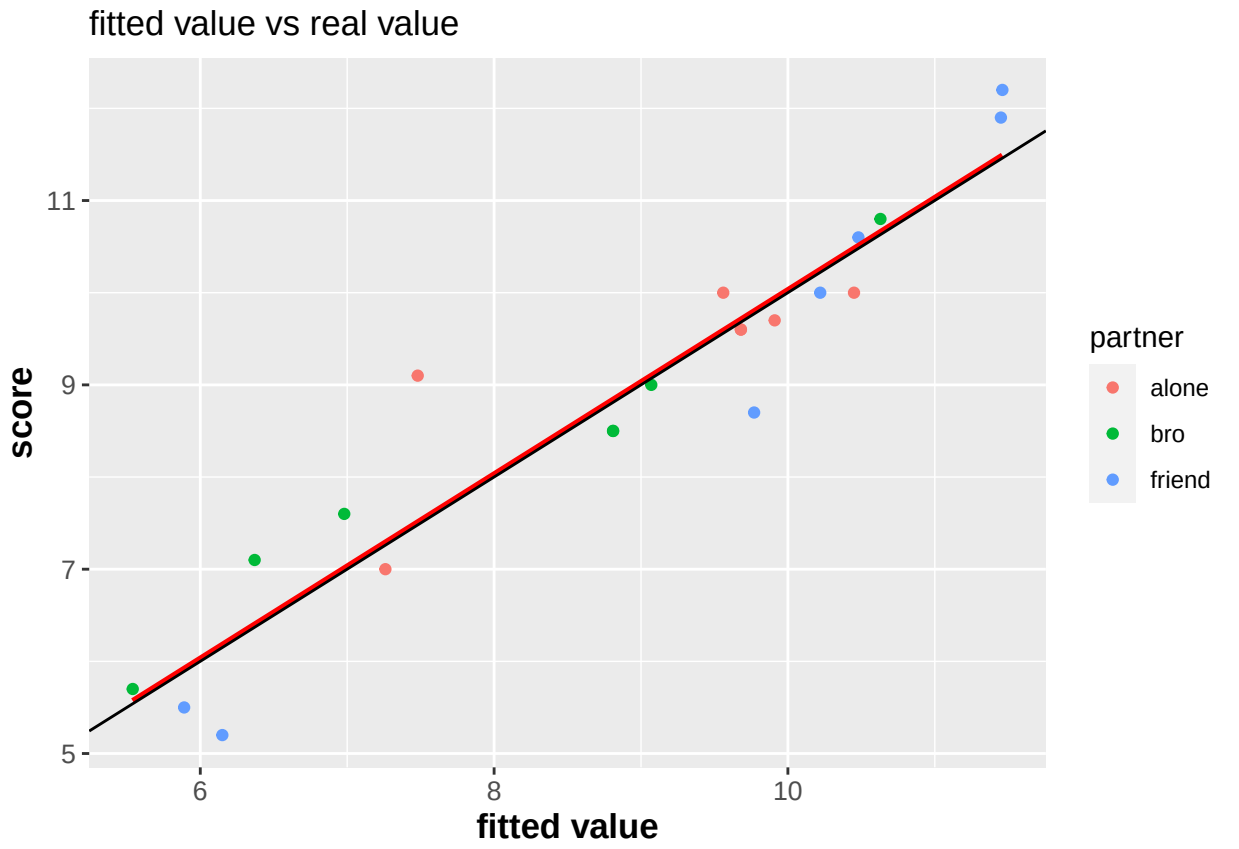
```
summary(fit_h)
```

```
##
## Call:
## lm(formula = score ~ kill + death + assist, data = game)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1107 -0.3491 -0.1037  0.4052  1.5965
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.22863    0.54189  11.494 1.94e-09 ***
## kill         0.38979    0.05906   6.600 4.50e-06 ***
## death       -0.51616    0.06713  -7.688 6.24e-07 ***
## assist       0.29197    0.03780   7.724 5.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6555 on 17 degrees of freedom
## Multiple R-squared:  0.904,  Adjusted R-squared:  0.887
## F-statistic: 53.34 on 3 and 17 DF,  p-value: 7.352e-09
```

The R-squared is close to 1, and the p-value is lower than 0.05. But just mentioned above, although the p-value imply that the model is a good fit, the effect size suggests that sample size should be larger to have a real good fit.

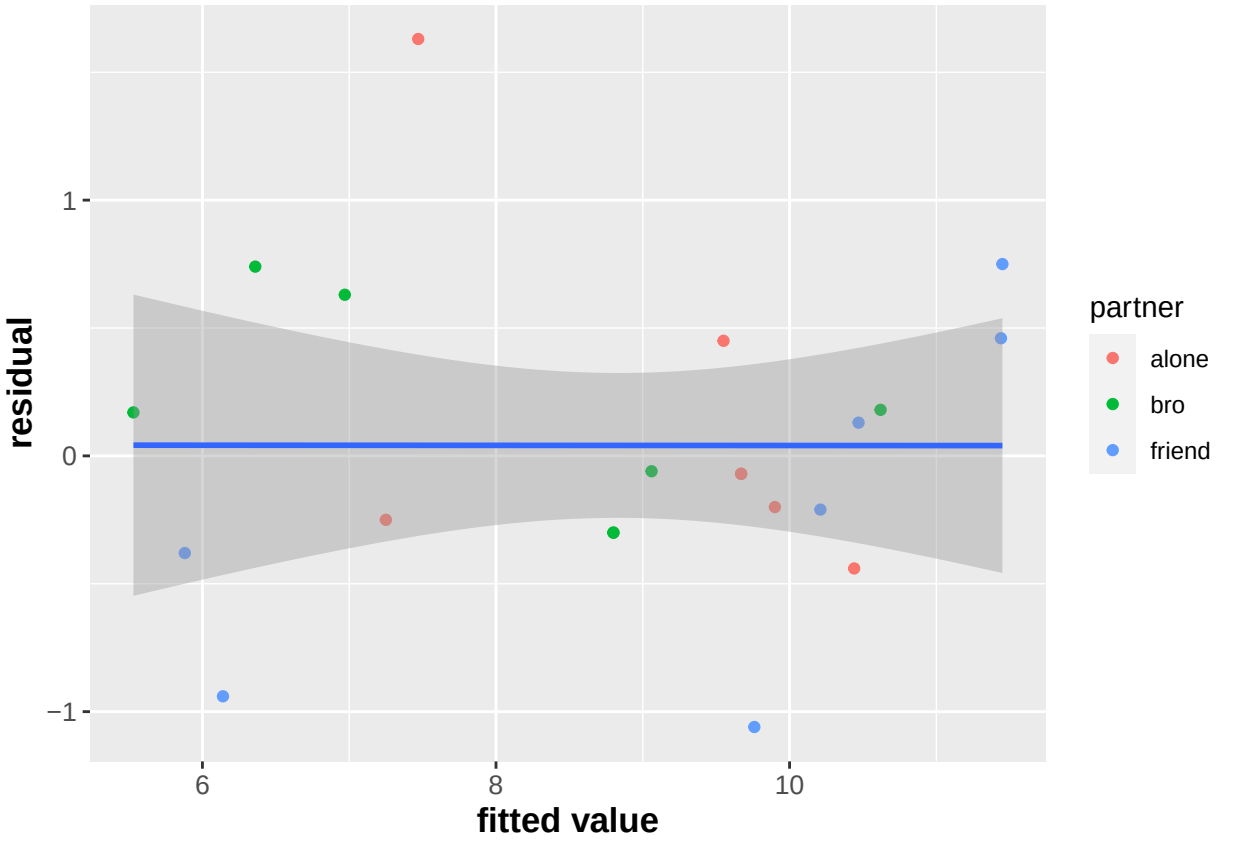
```
yhat <- 0.39*game$kill-0.52*game$death+0.29*game$assist+6.23
ggplot(game)+
  geom_point(mapping=aes(x=yhat,y=score,color=partner))+
  geom_smooth(mapping=aes(x=yhat,y=score),method = "lm",color="red",se=F)+
```

```
geom_abline(intercept = 0,slope = 1)+
theme(axis.text = element_text(size = 10),
      axis.title = element_text(size = 13, face = "bold")) +
labs(title = "fitted value vs real value",x = "fitted value")
```



because the model has multiple continuous predictors, so I draw the fitted value vs real value, and compared with line y=x. The red line is the regression line and the black one is line y=x. It is clear that two lines almost overlapped, and dots are all close to the line.

```
## `geom_smooth()` using formula 'y ~ x'
```



To see the fitted value vs residual, majority of residuals are close to 0. But still some dots are little bit fat from 0.

Inference (10pts)

```
round(coef(fit_h),digits = 2)
```

```
## (Intercept)        kill       death      assist
##        6.23        0.39       -0.52        0.29
```

If I do nothing-no killing, no death, no assistance-the score will be at around 6.23. And if I kill or assist to kill one more enemy in each game, the score will increase 0.39 and 0.29 separately. But if I dead one more time, the score will decrease 0.52. And there is no big difference when I play different roles or whether I won the game or who I play with. so I have stable performance in different situation.

Discussion (10pts)

After all the analysis, simple linear regression model is the best model for the data, due to all the categorical variables are not strongly correlated to the performance score. Although the p-value is under 0.05, I still cannot say that it is an optimal model for the effect size is not appropriate. And the number of death plays an important role in assessing the performance for the parameter of it is the highest. If I want to keep high score in the future game, saving life is more significant than killing.
And just as the GH textbook mentioned:"In various limiting cases, the classical and multilevel approaches coincide. When there is very little group-level variation, the multilevel model reduces to classical regression with no group indicators","When the number of groups is small (less than five, say), there is typically not enough information to accurately estimate group-level variation. As a result, multilevel models in this setting typically gain little beyond classical varying-coefficient models." In this data, using multilevel model just like the same as classical regression for the number of group is small and the group-level variation is small.

Also, in the EDA part seems all the categorical variables can play roles in the model, but actually none of them. And the line seems that number of killing has less impacts than assistance, but the results shows that killing affects more than number of assistance. Maybe it because the sample size is too small, so there are occasions to make it different.

Limitations and future opportunity. (10pts)

1.sample size is the most important aspect that improve the model. For now in each group I only have 7 observations. It is far from touching the truth.
2.In a game, there are lots of variables will influence the score, not only the variable listed in the data, so I can include more different variables(ranks, when I played, golds, participation rate, etc).
3.If I collect more data, maybe there will be correlation between partners, roles or results of the game with score.
4.EDA show different results with the model, try to use larger data to see if EDA shows the same conclusion with model.

Comments or questions

For there are three continuous outcome so it is a little bit difficult to draw the plots. If indeed that `partner` or `role` will influence the score, I still not sure how to visualize it.