NTU Exercise1

Yu Tian

2022-08-04

Package

```
#package
library(dplyr)

#set seed
set.seed(0623)
```

Preparation

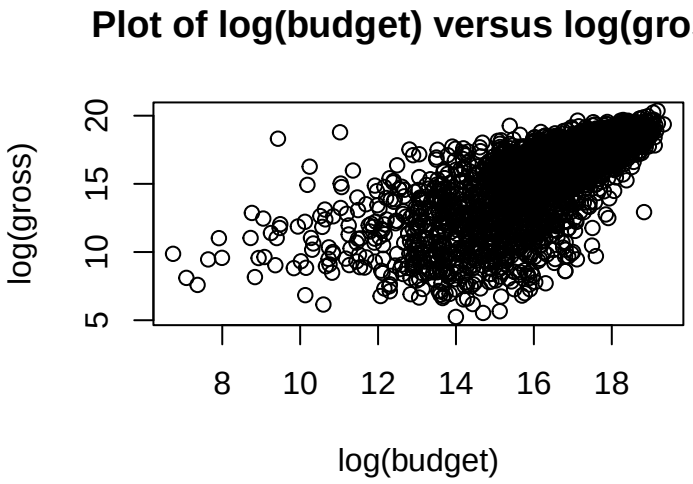
```
# Read (and import) the full exercise data set into R using read.csv()
data1 <- read.csv(file = 'Data_exercise1.csv')

# view the data example in R
data1 %>% head()
```

```
##   Year                                          Title Sequel
## 1 1995                                 Casper (1995)      0
## 2 1995                                  Se7en (1995)      0
## 3 1995                       The Usual Suspects (1995)      0
## 4 1995 Halloween: The Curse of Michael Myers (1995)      1
## 5 1995                                GoldenEye (1995)      0
## 6 1995                                 Clueless (1995)      0
##                                       ReleaseDate
## 1       Release Date: 26 May 1995 (USA) See more \xd4_
## 2 Release Date: 22 September 1995 (USA) See more \xd4_
## 3 Release Date: 15 September 1995 (USA) See more \xd4_
## 4 Release Date: 29 September 1995 (USA) See more \xd4_
## 5  Release Date: 17 November 1995 (USA) See more \xd4_
## 6      Release Date: 21 July 1995 (USA) See more \xd4_
##
## 1 Furious that her late father only willed her his gloomy-looking mansion rather than his millions, C
## 2                                                    A film about two homicide detectives' desperate
## 3
## 4
## 5
## 6
##   Runtime                         SoundMix Color            AspectRatio
## 1 100 min      Sound Mix: DTS | Dolby SR Color Aspect Ratio: 1.85 : 1
## 2 127 min Sound Mix: DTS | Dolby Digital Color Aspect Ratio: 2.35 : 1
## 3 106 min       Sound Mix: Dolby Digital Color Aspect Ratio: 2.35 : 1
## 4  87 min        Sound Mix: Ultra Stereo Color Aspect Ratio: 1.85 : 1
```

```
## 5 130 min Sound Mix: DTS | Dolby Digital Color Aspect Ratio: 2.35 : 1
## 6  97 min       Sound Mix: Dolby Digital Color Aspect Ratio: 1.85 : 1
##                                                                                       MPAA
## 1                                         Rated PG for mild language and thematic elements
## 2 Rated R for grisly afterviews of horrific and bizarre killings, and for strong language
## 3                         Rated R for violence and a substantial amount of strong language
## 4                                   Rated R for strong horror violence, and some sexuality
## 5        Rated PG-13 for a number of sequences of action/violence, and for some sexuality
## 6             Rated PG-13 for sex related dialogue and some teen use of alcohol and drugs
##   RateScore NumRate  NumReview  NumCritic ProductionDate
## 1       6.0  79,232    88 user  25 critic           null
## 2       8.6 930,951 1,050 user 203 critic           null
## 3       8.6 679,461 1,165 user 158 critic           null
## 4       4.9  19,470   328 user 105 critic           null
## 5       7.2 186,512   400 user 132 critic           null
## 6       6.8 111,526   245 user  87 critic           null
##                      FilmingDate     gross   budget critic numrate numreview
## 1      27 January 1994 - 8 June  112842727 56236798     25   79232        88
## 2   12 December 1994 - 10 March  112614910 37116286    203  930951      1050
## 3        13 June 1994 - 29 July    1259704  6748416    158  679461      1165
## 4  28 October 1994 - 5 December   16990668  5623680    105   19470       328
## 5      16 January 1995 - 6 June  112393959 65234685    132  186512       400
## 6 21 November 1994 - 7 February   55565345 13496831     87  111526       245
##   ratescore nummon day daynum budget_class competition ActorScore
## 1       6.0      5  26    176            5          40   9.060655
## 2       8.6      9  22    292            4          46   8.579231
## 3       8.6      9  15    285            1          36  -1.127334
## 4       4.9      9  29    299            1          60  -1.450700
## 5       7.2     11  17    347            5          47   7.248681
## 6       6.8      7  21    231            2          40   1.483862
##   ActorScore_budget Country BigSix country_num ThreeD  mpaa gross_ind NewBigSix
## 1          6.246445     USA      0           1      0    PG         9         1
## 2          5.803904     USA      0           1      0     R         9         0
## 3         -1.886623     USA      0           1      0     R         2         0
## 4          5.657837     USA      0           1      0     R         4         0
## 5          1.762772     USA      0           1      0 PG-13         9         0
## 6          8.622864     USA      1           1      0 PG-13         7         1
```

1. Make a plot using log(budget) on the x-axis and log(gross) on the y-axis.

```r
# simple plot
plot(log(data1$budget),   #log(budget) on the x-axis
     log(data1$gross),    #log(gross) on the y-axis
     main = "Plot of log(budget) versus log(gross)",
     xlab = "log(budget)", ylab = "log(gross)")
```

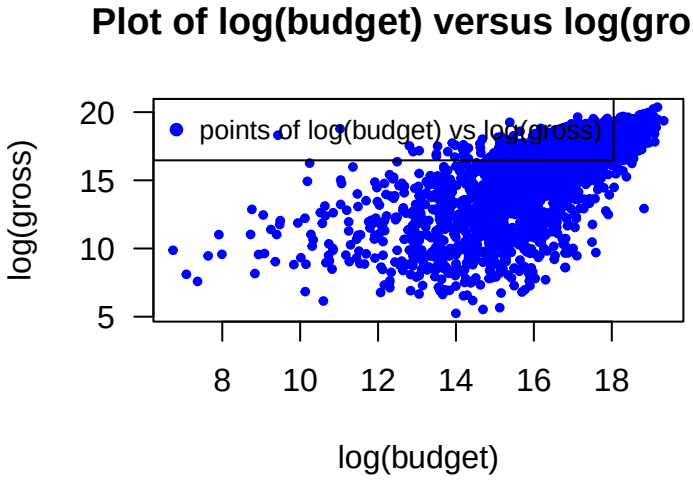


2.

add legend to the plot. Change the color/shape/size of the dots.

make the y-axis lable horizontal

```
plot(log(data1$budget), log(data1$gross),
     col = "blue",  # change the color
     pch = 20,  # change the shape
     cex = 0.8,  # change the size
     las = 1,  # make the y-axis lable horizontal
     main = "Plot of log(budget) versus log(gross)",
     xlab = "log(budget)", ylab = "log(gross)")

# add a legend
legend("topleft", "points of log(budget) vs log(gross)",
       pch = 20, col="blue",
       pt.cex = 1.2, cex = 0.8)
```



3.

Pre

```
# package
library(rvest)
library(stringr)

# read the website "Basketball-Reference.com"
BR=read_html('https://www.basketball-reference.com/')
```

Scrape data (more than 100 observations, e.g. player game points) from Basketball-Reference.com

```
# teams information
teams=BR %>%
  html_nodes("#teams .left a") %>%
  html_text()

teams_link=BR %>%
  html_nodes("#teams .left a")%>%
  html_attr('href')

teams_info=data.frame(teams,teams_link)
```

```
# all player names by looping over all teams
players_name=c()
for (i in 1:nrow(teams_info)) {
  team_link=paste0('https://www.basketball-reference.com',teams_info$teams_link[i])

  BR=read_html(team_link)

  players=BR %>%
    html_nodes(".iz+ .left a") %>%
    html_text()

  players_link=BR %>%
    html_nodes(".iz+ .left a") %>%
    html_attr('href')

  players_name_i=data.frame(players,players_link)

  players_name=rbind(players_name,players_name_i)
}
View(players_name)
```

```
# players information by looping all player names
players_info=c()
for (i in 1:nrow(players_name)) {
  player_link=paste0('https://www.basketball-reference.com',
                     players_name$players_link[i])

  BR=read_html(player_link)

  players_FGP=BR %>%
    html_nodes(".p2 div:nth-child(1) p+ p") %>%
```

```
    html_text()

  if(length(players_FGP)==0) {
    players_FGP = 0
  }

  players_FG3P=BR %>%
    html_nodes(".p2 div:nth-child(2) p+ p") %>%
    html_text()

  if(length(players_FG3P)==0) {
    players_FG3P = 0
  }

   players_FTP=BR %>%
    html_nodes(".p2 div:nth-child(3) p+ p") %>%
    html_text()

   if(length(players_FTP)==0) {
    players_FTP = 0
  }

  players_eFGP=BR %>%
    html_nodes(".p2 div:nth-child(4) p+ p") %>%
    html_text()

  if(length(players_eFGP)==0) {
    players_eFGP = 0
  }

   players_PER=BR %>%
    html_nodes(".p3 div:nth-child(1) p+ p") %>%
    html_text()

   if(length(players_PER)==0) {
    players_PER = 0
  }

  players_info_i=data.frame(players_name$players[i],
                            as.numeric(players_FGP),
                            as.numeric(players_FG3P),
                            as.numeric(players_FTP),
                            as.numeric(players_eFGP),
                            as.numeric(players_PER))

  players_info=rbind(players_info,players_info_i)
}

# view the players information in R
players_info %>% head(15)
```

```
##    players_name.players.i. as.numeric.players_FGP. as.numeric.players_FG3P.
## 1             Nikola Jović                     0.0                      0.0
## 2               Jamal Cain                     0.0                      0.0
```

```
## 3           Marcus Garrett                  23.8                  25.0
## 4              Darius Days                   0.0                   0.0
## 5           Jamaree Bouyea                   0.0                   0.0
## 6           Victor Oladipo                  43.8                  34.8
## 7         Orlando Robinson                   0.0                   0.0
## 8           Dewayne Dedmon                  52.7                  33.8
## 9             Caleb Martin                  45.5                  36.3
## 10             Tyler Herro                  44.0                  38.5
## 11       Haywood Highsmith                  35.7                  30.3
## 12         Duncan Robinson                  43.5                  40.6
## 13              Kyle Lowry                  42.5                  36.8
## 14               Max Strus                  44.6                  39.1
## 15            Gabe Vincent                  39.6                  34.3
##    as.numeric.players_FTP. as.numeric.players_eFGP. as.numeric.players_PER.
## 1                      0.0                     0.0                     0.0
## 2                      0.0                     0.0                     0.0
## 3                     40.0                    26.2                     4.7
## 4                      0.0                     0.0                     0.0
## 5                      0.0                     0.0                     0.0
## 6                     79.0                    49.3                    16.4
## 7                      0.0                     0.0                     0.0
## 8                     73.5                    56.6                    15.1
## 9                     72.5                    53.1                    12.7
## 10                    85.1                    52.1                    14.3
## 11                    28.6                    44.6                     6.5
## 12                    86.0                    61.0                    11.5
## 13                    81.3                    51.1                    18.0
## 14                    75.6                    59.9                    12.6
## 15                    83.1                    50.5                     9.4
```

Run a regression model to answer a question. (Linear regression)

Question: Use players_PER as dependent variable; and players_FGP, players_FG3P, players_FTP, players_eFGP as independent variables. Is there a strong correlation between dependent variable and other independent variables? Does this linear regression model fit the data well?

- – PER: Player Efficiency Rating
- – FGP: Field Goal Percentage
- – FG3P: 3-Point Field Goal Percentage
- – FTP: Free Throw Percentage
- – eFGP: Effective Field Goal Percentage

```r
# replace zero(0) with NA and remove the NA value
players_info_1 <- players_info %>%
  mutate_all(~na_if(., 0.0)) %>%
  na.omit(players_info_1)

# Change columns (variables) name
colnames(players_info_1) <- c("players_name", "players_FGP", "players_FG3P", "players_FTP", "players_eF

players_info_1 %>% head(15)
```

```
##          players_name players_FGP players_FG3P players_FTP players_eFGP
## 3      Marcus Garrett        23.8         25.0        40.0         26.2
## 6      Victor Oladipo        43.8         34.8        79.0         49.3
## 8      Dewayne Dedmon        52.7         33.8        73.5         56.6
## 9        Caleb Martin        45.5         36.3        72.5         53.1
## 10        Tyler Herro        44.0         38.5        85.1         52.1
## 11  Haywood Highsmith        35.7         30.3        28.6         44.6
## 12    Duncan Robinson        43.5         40.6        86.0         61.0
## 13         Kyle Lowry        42.5         36.8        81.3         51.1
## 14          Max Strus        44.6         39.1        75.6         59.9
## 15       Gabe Vincent        39.6         34.3        83.1         50.5
## 16     Omer Yurtseven        52.6          9.1        62.3         52.8
## 17        Bam Adebayo        55.8         14.0        74.1         55.9
## 18       Jimmy Butler        46.0         32.1        84.1         49.2
## 19      Rui Hachimura        47.7         36.0        77.9         51.7
## 21 Kristaps Porziņģis        44.4         35.3        82.0         50.2
##    players_PER
## 3          4.7
## 6         16.4
## 8         15.1
## 9         12.7
## 10        14.3
## 11         6.5
## 12        11.5
## 13        18.0
## 14        12.6
## 15         9.4
## 16        17.4
## 17        20.0
## 18        21.1
## 19        12.9
## 21        19.8
```

```
# run a linear regression model
lr_players_info <- lm(players_PER ~ players_FGP + players_FG3P + players_FTP + players_eFGP, data = play
# show the regression results
summary(lr_players_info)
```

```
##
## Call:
## lm(formula = players_PER ~ players_FGP + players_FG3P + players_FTP +
##     players_eFGP, data = players_info_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5755  -1.6967  -0.1989   1.4055  10.5956
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -11.12465    1.95752  -5.683 2.59e-08 ***
## players_FGP    0.77943    0.05575  13.981  < 2e-16 ***
## players_FG3P   0.12026    0.02938   4.093 5.17e-05 ***
## players_FTP    0.12425    0.01701   7.304 1.58e-12 ***
## players_eFGP  -0.45456    0.06765  -6.719 6.50e-11 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.932 on 390 degrees of freedom
## Multiple R-squared:  0.4908, Adjusted R-squared:  0.4856
## F-statistic: 93.99 on 4 and 390 DF,  p-value: < 2.2e-16
```

From the p value above, we can find that these coefficient of all independent variable is significant. However, from the R-squared value above, we can find this model id not fitting the data very well.

4. Output the results using 'stargazer'

```
# package
library(stargazer)

# output the regression results
stargazer(lr_players_info, type = "text", title = "Linear regression model of player information")
```

```
##
## Linear regression model of player information
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                              players_PER
## ------------------------------------------------
## players_FGP                   0.779***
##                                (0.056)
##
## players_FG3P                  0.120***
##                                (0.029)
##
## players_FTP                   0.124***
##                                (0.017)
##
## players_eFGP                  -0.455***
##                                (0.068)
##
## Constant                     -11.125***
##                                (1.958)
##
## ------------------------------------------------
## Observations                     395
## R2                              0.491
## Adjusted R2                     0.486
## Residual Std. Error       2.932 (df = 390)
## F Statistic           93.989*** (df = 4; 390)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```