STAT 579 Midterm 1

Yifan Zhu

October 15, 2016

1. read the data

```
library(readxl)
scimagojr <- read_excel(path = "C:/Users/fanne/Desktop/STAT579/STAT579mid1/scimagojr.xlsx")
```

2. rename "-"

```
# replace '-' with NA
scimagojr$`SJR Best Quartile`[scimagojr$`SJR Best Quartile` == "-"] <- NA
```

3.

```
# replace 'book serie' with 'book series'
scimagojr$Type[scimagojr$Type == "book serie"] <- "book series"

# replace 'conference and proceeding' with 'conference proceedings'
scimagojr$Type[scimagojr$Type == "conference and proceeding"] <- "conference proceedings"
```

4.

```
# check the first 20 types of periodicals whose 'SJR Best Quartile' is not
# avalaible
head(scimagojr$Type[is.na(scimagojr$`SJR Best Quartile`)], 20)
```

```
##  [1] "conference proceedings" "conference proceedings"
##  [3] "conference proceedings" "conference proceedings"
##  [5] "conference proceedings" "conference proceedings"
##  [7] "conference proceedings" "conference proceedings"
##  [9] "conference proceedings" "conference proceedings"
## [11] "conference proceedings" "conference proceedings"
## [13] "conference proceedings" "conference proceedings"
## [15] "conference proceedings" "conference proceedings"
## [17] "conference proceedings" "conference proceedings"
## [19] "conference proceedings" "conference proceedings"
```

```
# check the first 20 types of periodicals whose 'SJR Best Quartile' is
# avalaible
head(scimagojr$Type[!is.na(scimagojr$`SJR Best Quartile`)], 20)
```

```
##  [1] "journal" "journal" "journal" "journal" "journal" "journal" "journal"
##  [8] "journal" "journal" "journal" "journal" "journal" "journal" "journal"
## [15] "journal" "journal" "journal" "journal" "journal" "journal"
```

Comment: from the head part we can see it seems the type of those not assigned SJR Best Quatile are all conference proceeding, and those assigned SJR Best Quatile are all journal.

5.

```
# remove the 3 columns: 'SJR Best Quartile', 'Issn', 'Country'
scimagojr <- scimagojr[, !names(scimagojr) %in% c("SJR Best Quartile", "Issn",
    "Country")]
```

6.

```
# Calculate the median for all the metrics in the data set for each type
aggregate(. ~ Type, data = scimagojr[, !names(scimagojr) %in% c("Rank", "Title")],
    FUN = median)
```

```
##                      Type   SJR H index Total Docs. (2015)
## 1            book series 0.130       4                  9
## 2 conference proceedings 0.110       2                  0
## 3                journal 0.321      14                 39
##   Total Docs. (3years) Total Refs. Total Cites (3years)
## 1                   35         373                    7
## 2                   62           0                    6
## 3                  123        1150                   70
##   Citable Docs. (3years) Cites / Doc. (2years) Ref. / Doc.
## 1                     34                  0.11       27.24
## 2                     60                  0.01        0.00
## 3                    112                  0.61       28.50
```

Comment: Among these three types, journals have the highest median value in all fields of metrics, the conference proceedings have the lowest and the book series is in the middle. It seems that the type journal overall has more infuence than others, and the influence of conference proceedings is small.

7.

```
# pull out the records whose type is 'journal', remove the column 'Type'
scimagojr.journal <- subset(scimagojr, Type == "journal", select = -c(Type))
```

8.

```
# do linear regression of SJR using all the metrics
journallm <- lm(formula = SJR ~ ., data = scimagojr.journal[, !names(scimagojr.journal) %in%
    c("Title", "Rank")])

summary(journallm)
```

```
##
## Call:
## lm(formula = SJR ~ ., data = scimagojr.journal[, !names(scimagojr.journal) %in%
##     c("Title", "Rank")])
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11.8328  -0.1400   0.0272   0.1055  19.2079 
## 
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)    
## (Intercept)              -6.590e-03  6.699e-03  -0.984   0.3253    
## `H index`                 4.251e-03  1.533e-04  27.742  < 2e-16 ***
## `Total Docs. (2015)`      2.538e-06  5.332e-05   0.048   0.9620    
## `Total Docs. (3years)`   -3.329e-04  4.121e-05  -8.079 6.86e-16 ***
## `Total Refs.`            -6.873e-06  1.187e-06  -5.791 7.10e-09 ***
## `Total Cites (3years)`    2.244e-05  2.569e-06   8.733  < 2e-16 ***
## `Citable Docs. (3years)`  2.853e-04  4.109e-05   6.942 3.96e-12 ***
## `Cites / Doc. (2years)`   5.200e-01  2.897e-03 179.502  < 2e-16 ***
## `Ref. / Doc.`            -4.745e-04  1.682e-04  -2.821   0.0048 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.581 on 22597 degrees of freedom
## Multiple R-squared:  0.7971, Adjusted R-squared:  0.797 
## F-statistic: 1.11e+04 on 8 and 22597 DF,  p-value: < 2.2e-16
```

Comment: The variable “Total Doc. (2015)” is not significant. It is easy to understand because the total number of articles published in 2015 should not have much impact on the influence of a periodical. The quality rather than the quantity of articles matters.

9. restrict to the statistical journals

   (a)

```
# find all the periodicals that contain 'statistic' in the title
scimagojr.journal.stat <- scimagojr.journal[grep(pattern = "statistic", x = scimagojr.journal$Title
    ignore.case = T), ]
```

   (b)

```
# rows to be added to the data frame
addrow <- scimagojr.journal[scimagojr.journal$Title %in% c("Stat", "Biometrika",
    "Biometrics", "Biometrical Journal"), ]

# add rows to the data frame
scimagojr.journal.stat <- rbind(scimagojr.journal.stat, addrow)

# find the index of and remove records with skew result described in the
# problem
droprow.id <- scimagojr.journal.stat$Title %in% c("Vital & health statistics. Series 3, Analytical
    "National vital statistics reports : from the Centers for Disease Control and Prevention, Natio

scimagojr.journal.stat <- scimagojr.journal.stat[!droprow.id, ]
```

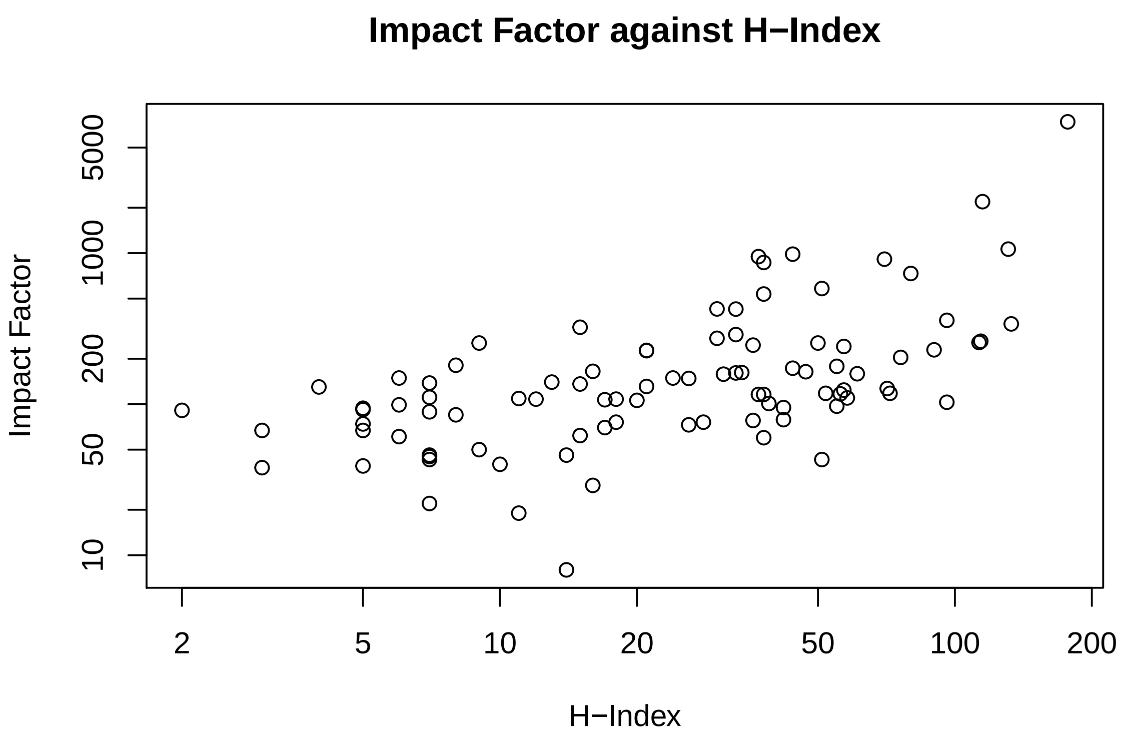
   (c)

```
# find the index of records whose values are all non-zero in a row
nozero.id <- as.vector(apply(scimagojr.journal.stat != 0, MARGIN = 1, FUN = prod) ==
    1)

# remove those records
scimagojr.journal.stat_nozero <- scimagojr.journal.stat[nozero.id, ]
```

10. (a)

```
# plot Impact Factor against H-Index
plot(x = scimagojr.journal.stat_nozero$`H index`, y = scimagojr.journal.stat_nozero$`Citable Docs.
    log = "xy", xlab = "H-Index", ylab = "Impact Factor", main = "Impact Factor against H-Index")
```



Comment: we can see the Impact Factor increases as the H-Index increases. It makes sense because the as the number of articles that have received at least H many citations over the whole period increases, the average citations per article in a period of time would almost surely increase.

(b)

```
# define a function to return the titles with top 5 values of x
top5title <- function(x) {
    return(scimagojr.journal.stat_nozero$Title[order(x, decreasing = T)[1:5]])
}

# apply to each column
apply(subset(scimagojr.journal.stat_nozero, select = -c(Rank, Title)), MARGIN = 2,
    FUN = top5title)
```

```
##      SJR
## [1,] "Journal of the Royal Statistical Society. Series B: Statistical Methodology"
## [2,] "Annals of Statistics"
## [3,] "Review of Economics and Statistics"
## [4,] "Statistical Methods in Medical Research"
## [5,] "Journal of the American Statistical Association"
##      H index
## [1,] "Physical Review E - Statistical, Nonlinear, and Soft Matter Physics"
## [2,] "Journal of the American Statistical Association"
## [3,] "Statistics in Medicine"
## [4,] "Physica A: Statistical Mechanics and its Applications"
## [5,] "Review of Economics and Statistics"
##      Total Docs. (2015)
## [1,] "Physical Review E - Statistical, Nonlinear, and Soft Matter Physics"
## [2,] "Physica A: Statistical Mechanics and its Applications"
## [3,] "Statistics and Probability Letters"
## [4,] "Communications in Statistics - Theory and Methods"
## [5,] "Journal of Statistical Computation and Simulation"
##      Total Docs. (3years)
## [1,] "Physical Review E - Statistical, Nonlinear, and Soft Matter Physics"
## [2,] "Physica A: Statistical Mechanics and its Applications"
## [3,] "Statistics in Medicine"
## [4,] "Statistics and Probability Letters"
## [5,] "Computational Statistics and Data Analysis"
##      Total Refs.
## [1,] "Physical Review E - Statistical, Nonlinear, and Soft Matter Physics"
## [2,] "Physica A: Statistical Mechanics and its Applications"
## [3,] "Journal of Statistical Mechanics: Theory and Experiment"
## [4,] "Statistics in Medicine"
## [5,] "Communications in Statistics - Theory and Methods"
##      Total Cites (3years)
## [1,] "Physical Review E - Statistical, Nonlinear, and Soft Matter Physics"
## [2,] "Physica A: Statistical Mechanics and its Applications"
## [3,] "Statistics in Medicine"
## [4,] "Computational Statistics and Data Analysis"
## [5,] "Journal of Statistical Physics"
##      Citable Docs. (3years)
## [1,] "Physical Review E - Statistical, Nonlinear, and Soft Matter Physics"
## [2,] "Physica A: Statistical Mechanics and its Applications"
## [3,] "Statistics in Medicine"
## [4,] "Statistics and Probability Letters"
## [5,] "Communications in Statistics - Theory and Methods"
##      Cites / Doc. (2years)
## [1,] "National health statistics reports"
## [2,] "Journal of the Royal Statistical Society. Series B: Statistical Methodology"
## [3,] "Annual Review of Statistics and Its Application"
## [4,] "Statistics Surveys"
## [5,] "Review of Economics and Statistics"
##      Ref. / Doc.
## [1,] "Annual Review of Statistics and Its Application"
## [2,] "Statistics Surveys"
## [3,] "Wiley Interdisciplinary Reviews: Computational Statistics"
## [4,] "Statistical Science"
## [5,] "Annals of Applied Statistics"
```