



โครงงานย่อย

วิชา Business Data Analytic

เรื่อง ทำนายค่าเหนื่อยนักกีฬาบาสเกตบอลโดยใช้ Multiple Regression

จัดทำโดย

เขตโสภณ ขุนพารเพิง 60070127

ภูวนันต์ โลกเจริญลาภ 60070154

นำเสนอโดย

อ.วารุณี บัววิรัตน์

หลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิทยาการข้อมูลและการวิเคราะห์เชิงธุรกิจ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าคุณทหารลาดกระบัง

ปีการศึกษา 2561

คำนำ

รายงานวิชา Business Data Analytics เล่มนี้จัดทำเพื่อศึกษาค้นคว้าวิธีทำเพื่อทำนายค่าหนึ่งจากหลายๆ ตัวแปรซึ่งผู้จัดทำได้รับมอบหมายจากอาจารย์ผู้สอนให้ศึกษาเพิ่มเติมจาก เอกสาร อินเทอร์เน็ต และแหล่งข้อมูลต่างๆ เพื่อนำสิ่งที่ค้นคว้ามาทำเป็นรายงานเพื่อเป็นประโยชน์ในการเรียนการสอนของตนเองและอาจารย์ต่อไป

ผู้จัดทำได้ทำการศึกษาข้อมูลเกี่ยวกับนักกีฬาบาสเก็ตบอล และนำข้อมูลนำมาเป็นข้อมูลในการทำ Model นี้ โดยหาความสัมพันธ์ของตัวแปรต่างๆ ที่ผู้จัดทำสนใจ ซึ่งผู้จัดทำหวังเป็นอย่างยิ่งว่ารายงานเล่มนี้จะ เป็นประโยชน์ต่อผู้ที่สนใจและผู้ที่น่าไปใช้ให้เกิดประโยชน์ตามสิ่งที่หวัง

ผู้จัดทำ

สารบัญ

| | |
|--------------------|------|
| ความเป็นมา | 1 |
| วัตถุประสงค์ | 2 |
| สมมุติฐาน | 2 |
| ข้อมูล | 2 |
| การเก็บข้อมูล | 3-5 |
| การวิเคราะห์ข้อมูล | 6-10 |
| สรุปผลการวิเคราะห์ | 10 |
| ประโยชน์ที่ได้ | 10 |

ความเป็นมา

ในปัจจุบันการแข่งขันกีฬามีอิทธิพลมากต่อมูลค่าต่างๆ เช่น Brand สินค้าต่างๆ จึงทำให้ต้องเลือกนักกีฬาที่เหมาะสมซึ่งเราก็อยากทราบว่าค่าเหนื่อยของนักกีฬานั้นมีตัวแปรอะไรบ้างถึงทำให้ค่าเหนื่อยนั้นมากขึ้นการที่ค่าเหนื่อยของนักกีฬาเยอะนั้นทำให้ทราบอยู่แล้วว่่านักกีฬาคนนั้นจะต้องเก่งไม่จริงก็ไม่มีทีมไหนที่จะให้ค่าเหนื่อยสูงยิ่งค่าเหนื่อยสูงก็ทำให้รู้ว่่านักกีฬาคนนั้นเก่งทำให้การที่ให้นักกีฬาคนนั้นเป็น Brand Ambassador ให้กับ Brand จะทำให้มูลค่าของ Brand เราสูงขึ้น

วัตถุประสงค์

- ต้องการที่จะทราบว่าค่าเหนื่อย (Salary) ของนักบาสเกตบอลนั้นมีตัวแปรใดบ้างที่มีความสัมพันธ์ในการเพิ่มหรือลดค่าเหนื่อยของนักบาสแต่ละคน และสามารถที่จะทำนายค่าเหนื่อยของนักบาสตามตัวแปรที่มีได้

กำหนดสมมติฐาน

สมมติฐาน (เงื่อนไข)

1. e_i และ e_j เป็นอิสระกัน
2. ตัวแปรอิสระ X ต้องไม่มีความสัมพันธ์กันเอง เพื่อป้องกันการเกิด Multicollinearity

สมมติฐาน

1. ทดสอบว่าตัวแปรค่าเหนื่อย (Salary) มีความสัมพันธ์กับตัวแปรอื่นๆหรือไม่ที่ $\alpha = 0.05$
2. ทดสอบว่ามีตัวแปรใดบ้างที่อยู่ในสมการ

ข้อมูล

ข้อมูลที่มีคือ ข้อมูลของนักบาสเกตบอลในลีกอเมริกา ที่มีชื่อลีกว่า NBA (National Basketball Association) โดยเก็บข้อมูลของนักบาสในฤดูกาล (2017-2018) เพื่อมาเปรียบเทียบกับฤดูกาล (2016-2017) เพราะต้องการที่จะทำนายถึงค่าเหนื่อยที่จะได้รับในฤดูกาลถัดไปของนักบาสเกตบอล NBA



การเก็บข้อมูล

เก็บข้อมูลจาก เว็บไซต์ <https://www.basketball-reference.com/contracts/players.html> โดยใช้วิธี Web Scraping ผ่านโค้ดภาษา R จาก GitHub : <https://github.com/koki25ando/NBA-Players-2017-18-dataset> ซึ่งเลือกเก็บมาจากนักบาสเก็ตบอลของฤดูกาล (2016 - 2017) และ นักบาสเก็ตบอลของฤดูกาล (2017 - 2018) แต่ตัวข้อมูลที่จะเลือกใช้ในการทำนายนั้นจะเป็นของฤดูกาล (2017 - 2018)

จัดการกับข้อมูลต่างๆ (ซึ่งมีทั้งค่า Missing Value และอื่นๆ) และทำการ Export ออกมาเป็นไฟล์ .csv โดยใช้ภาษา R ดังนี้

1. โหลดข้อมูลและเตรียมข้อมูลเข้า Rstudio และดูองค์ประกอบของข้อมูลดังรูปที่ (1),(2)

PREPARATION

Require packages

```
library(data.table)
library(corrplot)
library(GGally)
library(tidyverse)
library(PerformanceAnalytics)
library(plotly)
```

Data Preparation

```
salary.table <- read.csv("nba/NBA_season1718_salary.csv")
ss <- read.csv("nba/Seasons_Stats.csv")
```

รูปที่ (1) : แสดงการ import data นักกีฬาบาสเก็ตบอลปี 2017, 2018

```
$ BPM      : num  NA NA NA NA NA NA NA NA NA NA NA ...
$ VORP     : num  NA NA NA NA NA NA NA NA NA NA NA ...
$ FG       : int  144 102 174 22 21 1 340 5 226 125 ...
$ FGA      : int  516 274 499 86 82 4 936 16 813 435 ...
$ FG.      : num  0.279 0.372 0.349 0.256 0.256 0.25 0.363 0.313 0.278 0.287 ...
$ X3P      : int  NA NA NA NA NA NA NA NA NA NA NA ...
$ X3PA     : int  NA NA NA NA NA NA NA NA NA NA NA ...
$ X3P.     : num  NA NA NA NA NA NA NA NA NA NA NA ...
$ X2P      : int  144 102 174 22 21 1 340 5 226 125 ...
$ X2PA     : int  516 274 499 86 82 4 936 16 813 435 ...
$ X2P.     : num  0.279 0.372 0.349 0.256 0.256 0.25 0.363 0.313 0.278 0.287 ...
$ eFG.     : num  0.279 0.372 0.349 0.256 0.256 0.25 0.363 0.313 0.278 0.287 ...
$ FT       : int  170 75 90 19 17 2 215 0 209 132 ...
$ FTA      : int  241 106 129 34 31 3 282 5 321 209 ...
$ FT.      : num  0.705 0.708 0.698 0.559 0.548 0.667 0.762 0 0.651 0.632 ...
```

รูปที่ (2) : องค์ประกอบของข้อมูล

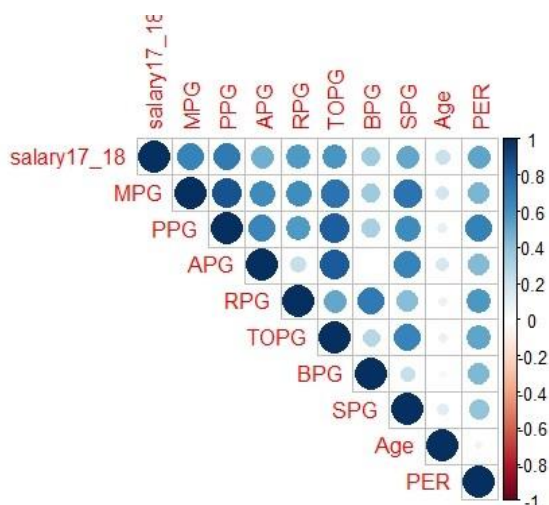
2. เมื่อเห็นว่องค์ประกอบของข้อมูลมีตัวแปรที่เป็น NA (Null) เยอะพอสมควรเลยทำ Data Cleaning ดังรูปที่ (3)

```
# Merging Data
stats_salary <- merge(stats17, salary.table, by.x = "Player", by.y = "Player")
names(stats_salary)[40] <- "salary17_18"
stats_salary <- stats_salary[-39]

# CORRELATION CHECK
corrplot(cor(stats_salary %>%
  select(salary17_18, MPG:SPG,
    Age, PER, contains("%")),
  use = "complete.obs"),
  method = "circle", type = "upper")
```

รูปที่ (3) : รูปแสดงการ Cleaning ข้อมูล

3. นำ Data มา Merge รวมกัน และดูว่าข้อมูลมีความสัมพันธ์กันหรือไม่ (ทั้งในเชิงลบและเชิงบวก หรือก็คือค่าสหสัมพันธ์ของข้อมูล ใกล้ -1.0 หรือ 1.0 หรือไม่) ดังรูปที่ (4)



รูปที่ (4) : Correlation แสดงความสัมพันธ์ของข้อมูล

Result :

ข้อมูลที่ได้มีความสัมพันธ์กันจะส่วนมาก (Correlation เข้าใกล้ 1.0)

4. Export ข้อมูลออกมาเป็น .csv และนำเข้าใช้ใน SPSS ต่อไป
สรุปได้ว่าข้อมูลที่ได้มีความสัมพันธ์กันมาก

```
# export as csv
write.csv(stats_salary_regression, file = "nba_salary_regression.csv")
```

รูปที่ (5) : แสดงการ Export ไฟล์เป็น CSV เพื่อทำใน SPSS

5. ได้ข้อมูลที่เหลือดังรูปที่ (6)

| B | C | D | E | F | G | H | I |
|-------------|----------|----------|----------|----------|----------|----------|----------|
| salary17_18 | MPG | PPG | APG | RPG | TOPG | BPG | SPG |
| 1312611 | 7.409091 | 2.181818 | 0.181818 | 1.636364 | 0.454545 | 0.590909 | 0.045455 |
| 2116955 | 13.75385 | 4.953846 | 1.923077 | 1.061538 | 1.015385 | 0.138462 | 0.384615 |
| 5504420 | 28.725 | 12.7375 | 1.875 | 5.0625 | 1.1125 | 0.5 | 0.8 |
| 7319035 | 29.06557 | 8.721311 | 1.622951 | 7.393443 | 1.540984 | 0.721311 | 0.983607 |
| 27734405 | 32.25 | 14 | 4.955882 | 6.823529 | 1.705882 | 1.279412 | 0.764706 |
| 9769821 | 14.10606 | 8.106061 | 0.863636 | 4.212121 | 0.5 | 0.242424 | 0.287879 |
| 6.00E+06 | 15.06383 | 7.361702 | 0.489362 | 6.212766 | 0.787234 | 0.680851 | 0.574468 |
| 10845506 | 15.54762 | 6.738095 | 0.714286 | 2.857143 | 0.833333 | 0.119048 | 0.428571 |
| 5725000 | 15.51471 | 5.970588 | 0.588235 | 1.264706 | 0.485294 | 0.117647 | 0.544118 |
| 4187599 | 20.25974 | 7.961039 | 0.571429 | 6.623377 | 1.324675 | 1.272727 | 0.480519 |

รูปที่ (6) : รูปแสดงข้อมูลหลังจากจัดการข้อมูล

ซึ่งได้แก่

- Salary17_18 : ค่าเหนื่อย (รายได้) ของนักบาสเกตบอล
- MPG : Minutes Played Per Game
- PPG : Points Per Game
- APG : Assists Per Game
- RPG : Total Rebounds Per Game
- TOPG : Turnovers Per Game
- BPG : Blocks Per Game
- SPG : Steals Per Game

การวิเคราะห์ข้อมูล

ใช้วิธี Multiple Linear Regression ในการวิเคราะห์ข้อมูลเพื่อทำนายหารายได้ของนักบาสเกตบอลในฤดูกาล (2017 - 2018)

1. ใช้ข้อมูลที่ Export มาจาก Rstudio ชื่อ nba_salary_regresssion.csv
2. ทำ Multiple Linear Regression
 - a. Analyze > Regression > Linear
 - b. Method = "Enter"
 - c. ย้ายตัวแปร Salary ไปไว้ช่อง dependent และเลือกทุกตัวแปรที่เหลือไปไว้ในช่อง Independents
 - d. ปุ่ม Statistics เลือก Estimates, Model fit, Collinearity diagnostics, Durbin-Watson และกด Continue
 - e. ปุ่ม Plots ย้าย "*ZRESID" ไป Y และ "*ZPRED" ไป X และกด Continue
 - f. ปุ่ม Save ตรง Predicted Values เลือก Unstandardized ตรง Residuals เลือก Unstandardized และกด Include the covariance matrix และกด Continue
 - g. กด OK
3. ตรวจสอบว่า e_i และ e_j เป็นอิสระกันดังรูปที่ (7)

Model Summary^b

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Durbin-Watson |
|-------|-------------------|----------|-------------------|----------------------------|---------------|
| 1 | .747 ^a | .558 | .551 | \$5,078,459.882 | 1.913 |

a. Predictors: (Constant), SPG, BPG, PPG, APG, RPG, MPG, TOPG

b. Dependent Variable: salary17_18

รูปที่ (7) : รูปแสดงการตรวจสอบความเป็นอิสระกัน

สรุป ค่า Durbin-Watson 1.913 มีค่าใกล้ 2 จึงยอมรับ H_0 หรือค่าความคลาดเคลื่อนเป็นอิสระต่อกัน

4. ทดสอบว่าเกิด Multicollinearity ในตัวแปร X หรือไม่ ดังรูปที่ (8)

Coefficients^a

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
|-------|------------|-----------------------------|------------|---------------------------|--------|------|-------------------------|-------|
| | | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | - 721376.312 | | | -3.872 | .000 | | |
| | | 2792908.716 | | | | | | |
| | MPG | 30564.983 | 68125.572 | .035 | .449 | .654 | .170 | 5.881 |
| | PPG | 686814.910 | 97014.583 | .546 | 7.080 | .000 | .171 | 5.833 |
| | APG | 1059086.795 | 295017.464 | .252 | 3.590 | .000 | .206 | 4.848 |
| | RPG | 916087.239 | 177186.897 | .295 | 5.170 | .000 | .313 | 3.190 |
| | TOPG | - 818817.759 | | -.282 | -3.309 | .001 | .140 | 7.143 |
| | | 2709446.696 | | | | | | |
| | BPG | 470136.392 | 871988.048 | .025 | .539 | .590 | .468 | 2.136 |
| | SPG | 631254.518 | 981145.189 | .034 | .643 | .520 | .373 | 2.680 |

a. Dependent Variable: salary17_18

รูปที่ (8) : รูปแสดงการทดสอบ Multicollinearity

จากตารางต้องดูจากค่า Tolerance ต้องเข้าใกล้ 1 จึงจะไม่เกิด Multicollinearity หรือดูที่ VIF ต้องเข้าใกล้ 1 มากยิ่งดี และไม่เกิน 5

5. ดูว่าตัวแปรใดบ้างที่อยู่ในสมการ จากตาราง Coefficients ดังรูปที่ (9)

| Coefficients ^a | | | | | | | | |
|---------------------------|------------|-----------------------------|------------|---------------------------|--------|------|-------------------------|-------|
| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | Collinearity Statistics | |
| | | B | Std. Error | Beta | t | Sig. | Tolerance | VIF |
| 1 | (Constant) | - 721376.312 | | | -3.872 | .000 | | |
| | | 2792908.716 | | | | | | |
| | MPG | 30564.983 | 68125.572 | .035 | .449 | .654 | .170 | 5.881 |
| | PPG | 686814.910 | 97014.583 | .546 | 7.080 | .000 | .171 | 5.833 |
| | APG | 1059086.795 | 295017.464 | .252 | 3.590 | .000 | .206 | 4.848 |
| | RPG | 916087.239 | 177186.897 | .295 | 5.170 | .000 | .313 | 3.190 |
| | TOPG | - 818817.759 | | -.282 | -3.309 | .001 | .140 | 7.143 |
| | | 2709446.696 | | | | | | |
| | BPG | 470136.392 | 871988.048 | .025 | .539 | .590 | .468 | 2.136 |
| | SPG | 631254.518 | 981145.189 | .034 | .643 | .520 | .373 | 2.680 |

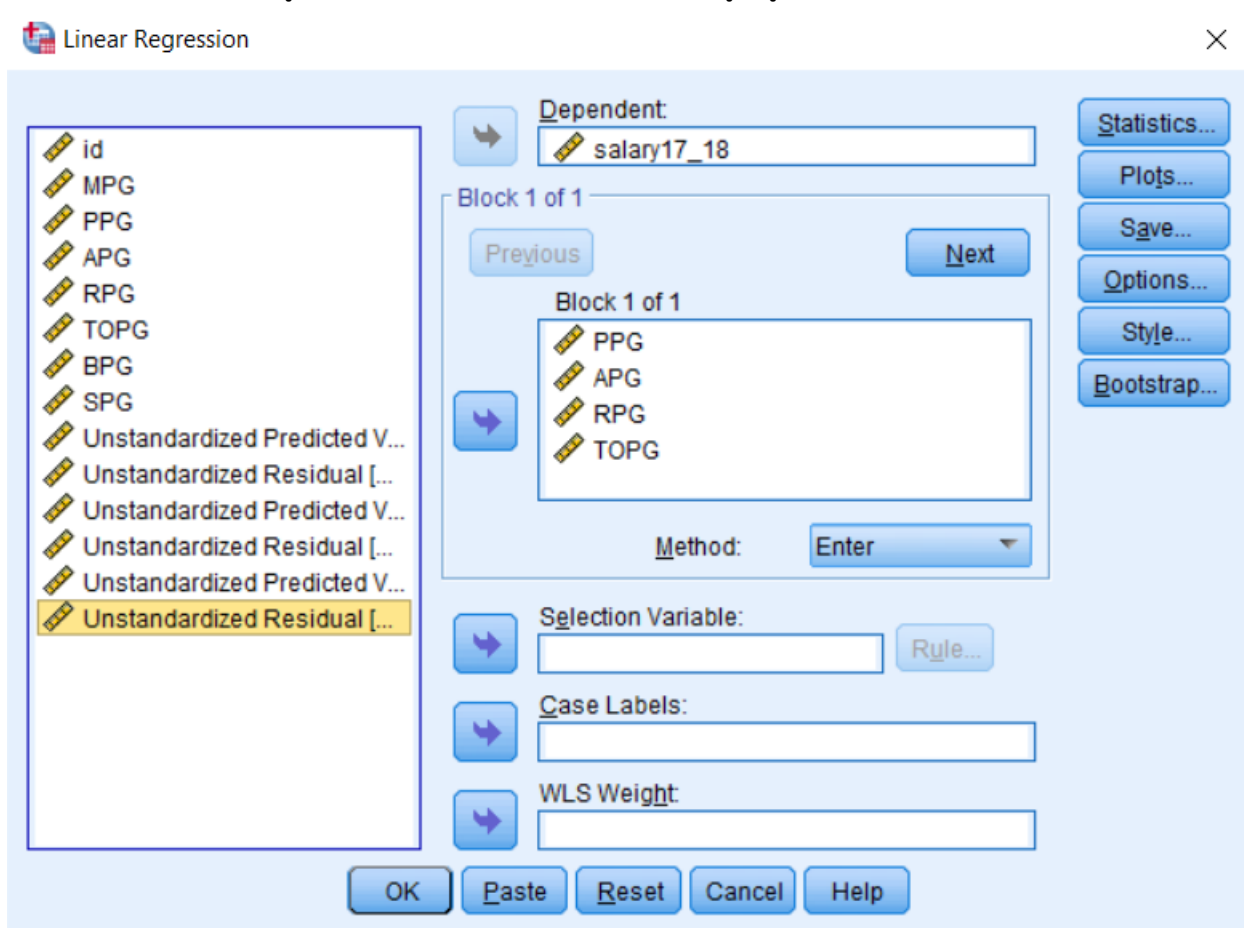
a. Dependent Variable: salary17_18

รูปที่ (9) : รูปแสดงตาราง Coefficients

ดูว่าค่า Sig < 0.05 หรือไม่ หากน้อยกว่าถือว่าอยู่ในสมการ แต่หากมากกว่าให้นำออกจากสมการ

สรุป ค่าที่อยู่ในสมการได้แก่ (Constant), PPG, APG, RPG และ TOPG

6. เลือกตัวแปรเข้าสู่สมการใหม่อีกรอบตามตัวแปรที่เหลืออยู่ดังรูปที่ (10)



Linear Regression

Dependent: salary17_18

Block 1 of 1

Previous Next

Block 1 of 1

PPG
APG
RPG
TOPG

Method: Enter

Selection Variable: Rule...

Case Labels:

WLS Weight:

OK Paste Reset Cancel Help

Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...

id
MPG
PPG
APG
RPG
TOPG
BPG
SPG
Unstandardized Predicted V...
Unstandardized Residual [...]
Unstandardized Predicted V...
Unstandardized Residual [...]
Unstandardized Predicted V...
Unstandardized Residual [...]

รูปที่ (10) : รูปแสดงการเลือกตัวแปรที่เหมาะสมเข้าสู่สมการ

7. หาวว่าตัวแปรใดบ้างที่อยู่ในสมการตามตาราง Coefficient รูปที่ (11)

Coefficients^a

| Model | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
|-------|-----------------------------|--------------|---------------------------|--------|------|-------------------------|-------|
| | B | Std. Error | Beta | | | Tolerance | VIF |
| 1 | (Constant) | - 490612.994 | | -4.875 | .000 | | |
| | | 2391914.254 | | | | | |
| | PPG | 724914.202 | .576 | 9.853 | .000 | .297 | 3.367 |
| | APG | 1145191.329 | .273 | 4.273 | .000 | .249 | 4.016 |
| | RPG | 1021336.617 | .329 | 7.719 | .000 | .560 | 1.785 |
| | TOPG | - 800658.453 | -.286 | -3.434 | .001 | .146 | 6.855 |
| | | 2749070.874 | | | | | |

a. Dependent Variable: salary17_18

รูปที่ (11) : รูปแสดงตาราง Coefficients

จะได้สมการดังนี้ :

$$Y = -2391914.254 + 724914.202PPG + 1145191.329APG + 1021336.617RPG - 2749070.874TOPG$$

สรุปผลการวิเคราะห์

ถ้าใน 1 เกมส์เพิ่ม 1 แด้ม จะเพิ่มค่าเหนื่อย \$724914.202

ถ้าใน 1 เกมส์เพิ่ม 1 Assist จะเพิ่มค่าเหนื่อย \$1145191.329

ถ้าใน 1 เกมส์เพิ่ม 1 Rebound จะเพิ่มค่าเหนื่อย \$1021336.617

ถ้าใน 1 เกมส์เสีย Turnover เพิ่ม 1 ครั้ง จะลดค่าเหนื่อย \$2749070.874

(Turnover หมายถึงบุกอยู่แล้วเสียการครองบอล)

ประโยชน์ที่ได้

ในการจ้างผู้เล่นคนหนึ่งการดูค่าความสามารถจะทำให้เราต่อรองถึงค่าจ้างที่เราจะจ้างผู้เล่นคนนั้นต่อไปได้ ให้เราได้ทราบว่าคุณค่าที่เราจะจ้างขั้นต่ำควรจะอยู่ที่ราคาเท่าไร