# Modern Data Science Methods for Educational Research

## R for Data Analysis in Educational Research

## Data Analysis II

ผศ.ดร.สิวะโชติ ศรีสุทธิยากร
อ.ดร.ประภาศิริ รัชประภาพรกุล

ภาควิชาวิจัยและจิตวิทยาการศึกษา
คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

February 12, 2023

# การวิเคราะห์ความสัมพันธ์

# วัตถุประสงค์ลักษณะที่ 2

เพื่อวิเคราะห์ความสัมพันธ์ระหว่างเงินเดือนของอาจารย์มหาวิทยาลัย กับตำแหน่งทางวิชาการ สาขาวิชา เพศ และประสบการณ์ทำงานที่แตกต่างกัน

1. สำรวจความสัมพันธ์
   ระหว่างเงินเดือนอาจารย์มหาวิทยาลัยกับตัวแปรอิสระต่าง ๆ
2. สำรวจความสัมพันธ์ระหว่างตัวแปรอิสระ

# Importing Data

```
library(dplyr)
dat <- read.csv("TeacherSalaryData.csv",
                header = TRUE,
                stringsAsFactors = TRUE)
dat<-dat[,-1]
dat$discipline <- factor(dat$discipline,
                         levels = c("A","B"),
                         labels = c("Pure Science",
                                    "Applied Science"))
dat <- dat %>%
       mutate(rank = factor(rank,
                            levels = c("AsstProf",
                                       "AssocProf",
                                       "Prof")))
```

# Importing Data

```
1 head(dat)
```

```
      rank      discipline yrs.since.phd yrs.service  sex
1     Prof Applied Science             19           18 Male
2     Prof Applied Science             20           16 Male
3 AsstProf Applied Science              4            3 Male
4     Prof Applied Science             45           39 Male
5     Prof Applied Science             40           41 Male
6 AssocProf Applied Science              6            6 Male
```

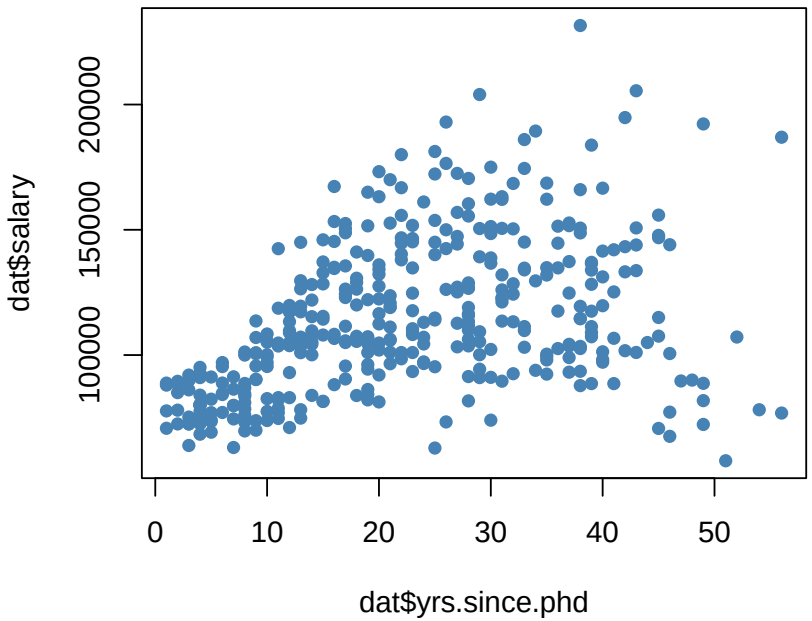
# ฟังก์ชัน `plot()`

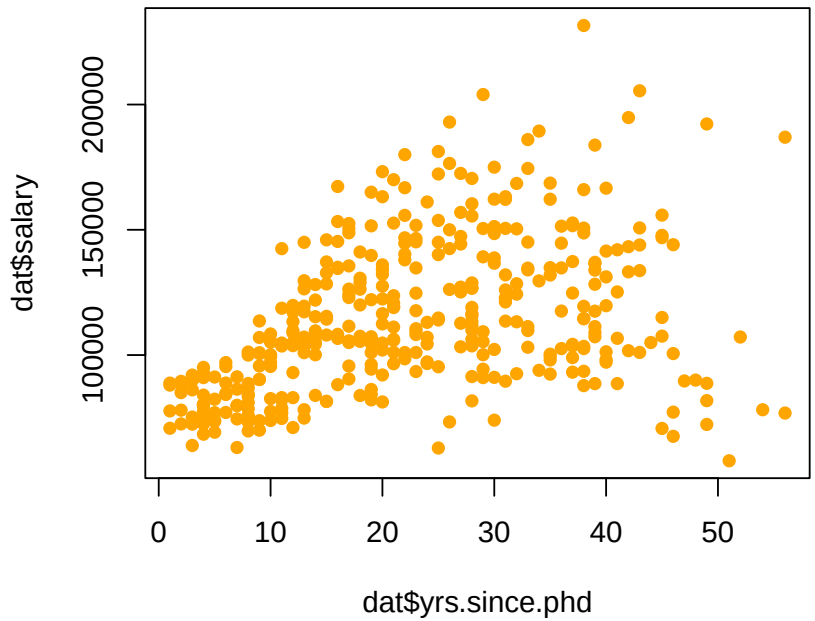
ฟังก์ชัน `plot()` เป็น generic graphic function ซึ่งสามารถใช้ plot แผนภาพที่แตกต่างกันได้ โดยขึ้นกับลักษณะข้อมูลที่นำเข้าในฟังก์ชัน

- ถ้า x และ y เป็นตัวแปรเชิงปริมาณทั้งคู่ ฟังก์ชันจะให้แผนภาพการกระจาย (scatter plot)
- ถ้า x เป็นตัวแปรจัดประเภท และ y เป็นตัวแปรเชิงปริมาณ จะให้ boxplot เชิงเปรียบเทียบ
- ถ้า x เป็นตัวแปรจัดประเภท และ y เป็นตัวแปรจัดประเภท จะให้ mosaic plot

# สำรวจความสัมพันธ์ระหว่างเงินเดือนกับตัวแปรอิสระเชิงปริมาณ

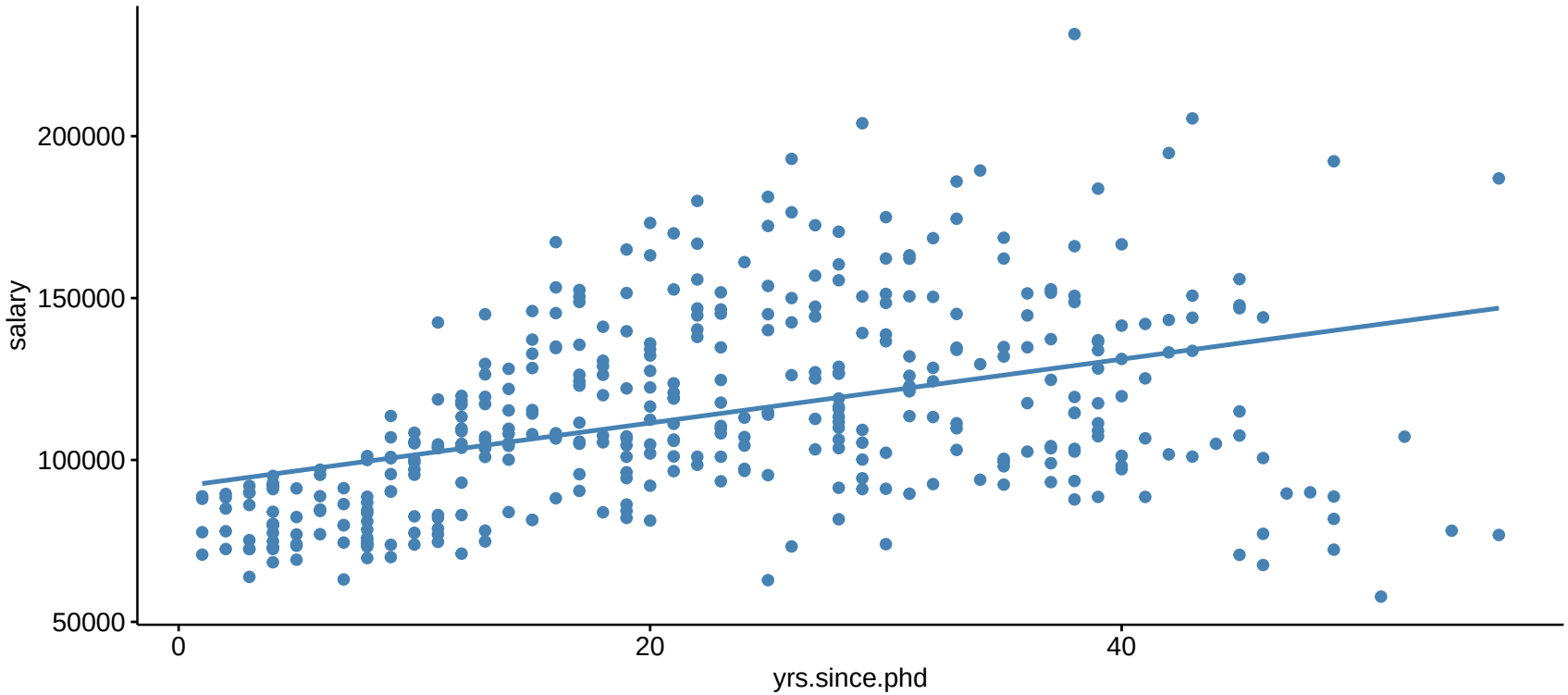
```
1 par(mfrow=c(1,2))
2 plot(dat$yrs.since.phd, dat$salary, pch=16, col="steelblue"
3 plot(dat$yrs.since.phd, dat$salary, pch=16, col="orange")
```



# สำรวจความสัมพันธ์ระหว่างเงินเดือนกับตัวแปรอิสระเชิงปริมาณ

```
1 library(ggpubr)
2 ggscatter(dat, x = "yrs.since.phd",
3                 y = "salary",
4           add = c("reg.line"), col= "steelblue")
```



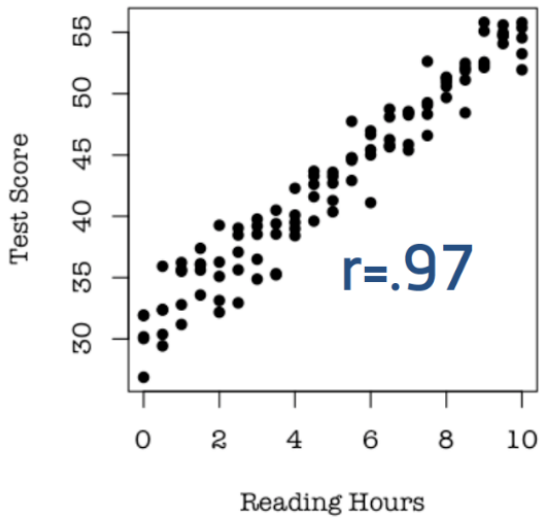
# สหสัมพันธ์

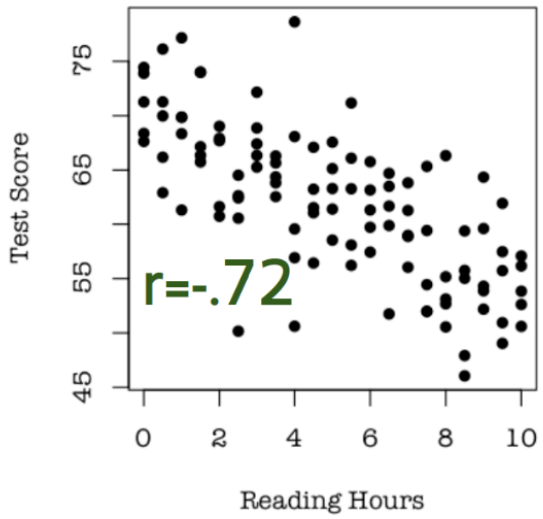
# สหสัมพันธ์ระหว่างเงินเดือนกับตัวแปรอิสระเชิงปริมาณ

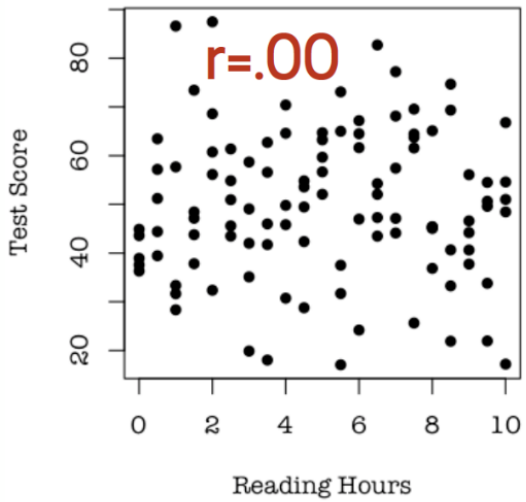
ฟังก์ชัน `cor()` ใช้หาค่า Correlation ระหว่างตัวแปรเชิงปริมาณที่กำหนด

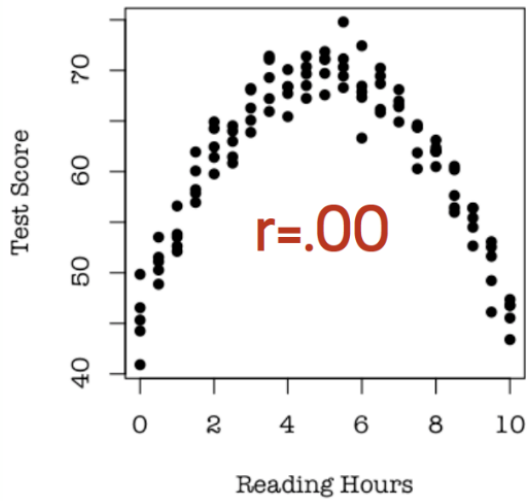
- ทิศทาง (+, -)
- ขนาด [-1, 1]
- นัยสำคัญทางสถิติ
- $R^2$ - coefficient of determination

# ขนาดและทิศทางของสหสัมพันธ์



# ขนาดของสหสัมพันธ์

| ค่าสัมบูรณ์ของสัมประสิทธิ์สหสัมพันธ์ | การแปลความหมาย |
|---|---|
| .90 - 1.00 | สหสัมพันธ์สูงมาก (very high correlation) |
| .70 - .90 | สหสัมพันธ์สูง (high correlation) |
| .50 - .70 | สหสัมพันธ์ปานกลาง (moderate correlation) |
| .30 - .50 | สหสัมพันธ์ต่ำ (low correlation) |
| .00 - .30 | สหสัมพันธ์ต่ำมาก (very low correlation) |

**Figure 1:** ที่มา : สิวะโชติ ศรีสุทธิยากร (2564)

# การคำนวณ correlation ด้วย R

```
cor.test(x, y,
         alternative = c("two.sided",
                         "less",
                         "greater"),
         method = c("pearson",
                    "kendall",
                    "spearman"))
```

# การคำนวณ correlation ด้วย R

```
1 dat %>%
2   select(salary, yrs.service, yrs.since.phd) %>%
3   cor()
```

```
                salary yrs.service yrs.since.phd
salary       1.0000000   0.3347447     0.4192311
yrs.service  0.3347447   1.0000000     0.9096491
yrs.since.phd 0.4192311   0.9096491     1.0000000
```

# Correlation Test

```
1 cor.test(dat$salary, dat$yrs.service)
```

```
    Pearson's product-moment correlation

data:  dat$salary and dat$yrs.service
t = 7.0602, df = 395, p-value = 7.529e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2443740 0.4193506
sample estimates:
      cor
0.3347447
```

# Correlation Test

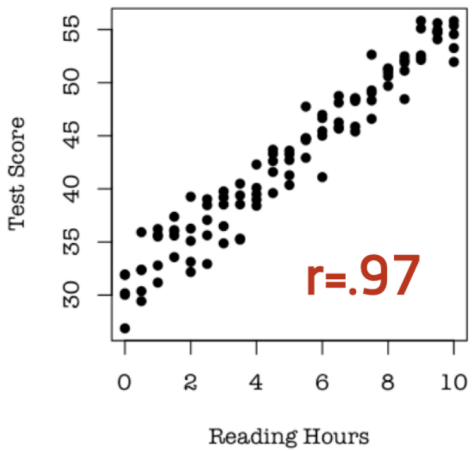
```
1 cor.test(dat$salary, dat$yrs.service, alternative = "greate
```

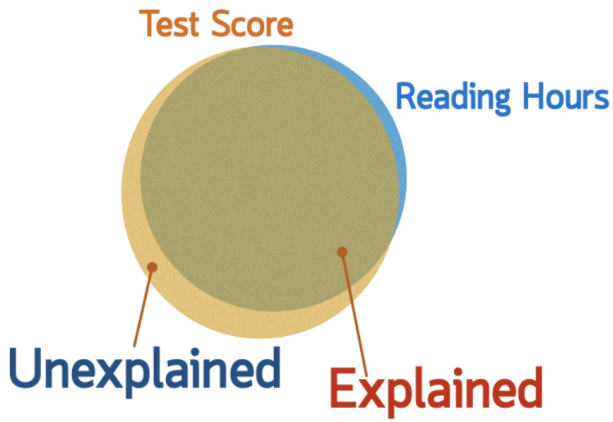
```
    Pearson's product-moment correlation

data:  dat$salary and dat$yrs.service
t = 7.0602, df = 395, p-value = 3.764e-12
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.259242 1.000000
sample estimates:
      cor
0.3347447
```

# Coefficient of Determination

$$R^2 = \frac{SSR}{SST} = \frac{\text{ความผันแปรในตัวแปรตาม } y \text{ ที่อธิบายได้ด้วยตัวแปรอิสระ } x}{\text{ความผันแปรรวมในตัวแปร } y}$$

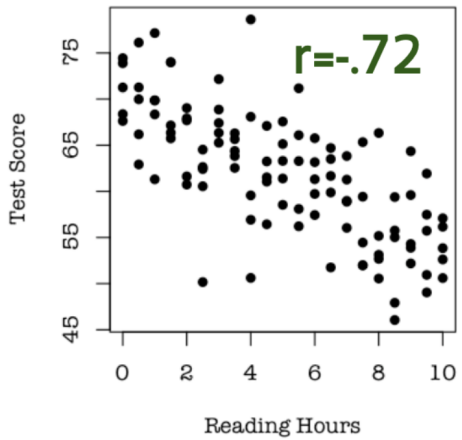
$R^2$ X100= ร้อยละของความผันแปรที่ร่วมกันระหว่างตัวแปร Y กับ X

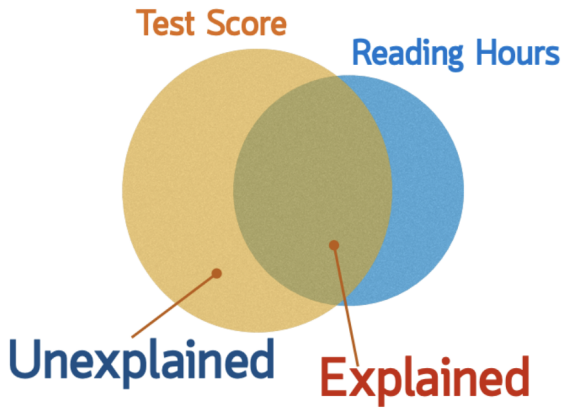



$$R^2 = \frac{\text{Explained}}{\text{Explained} + \text{Unexplained}} = (.97)^2 = .94$$

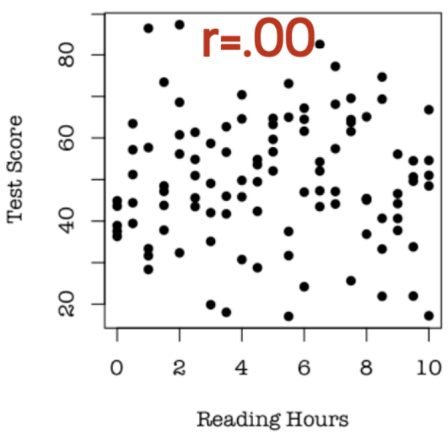
# Coefficient of Determination

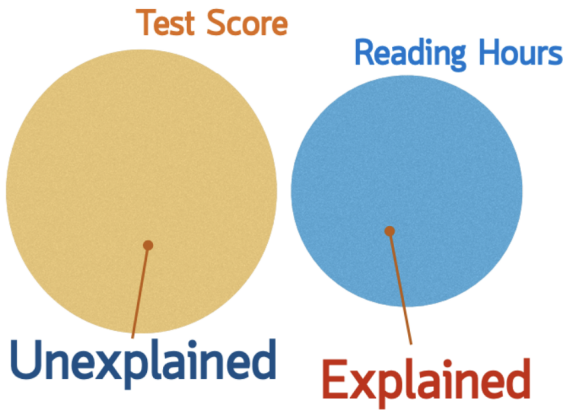



$$R^2 = \frac{\text{Explained}}{\text{Explained} + \text{Unexplained}} = (-.72)^2 = .518$$

# Coefficient of Determination



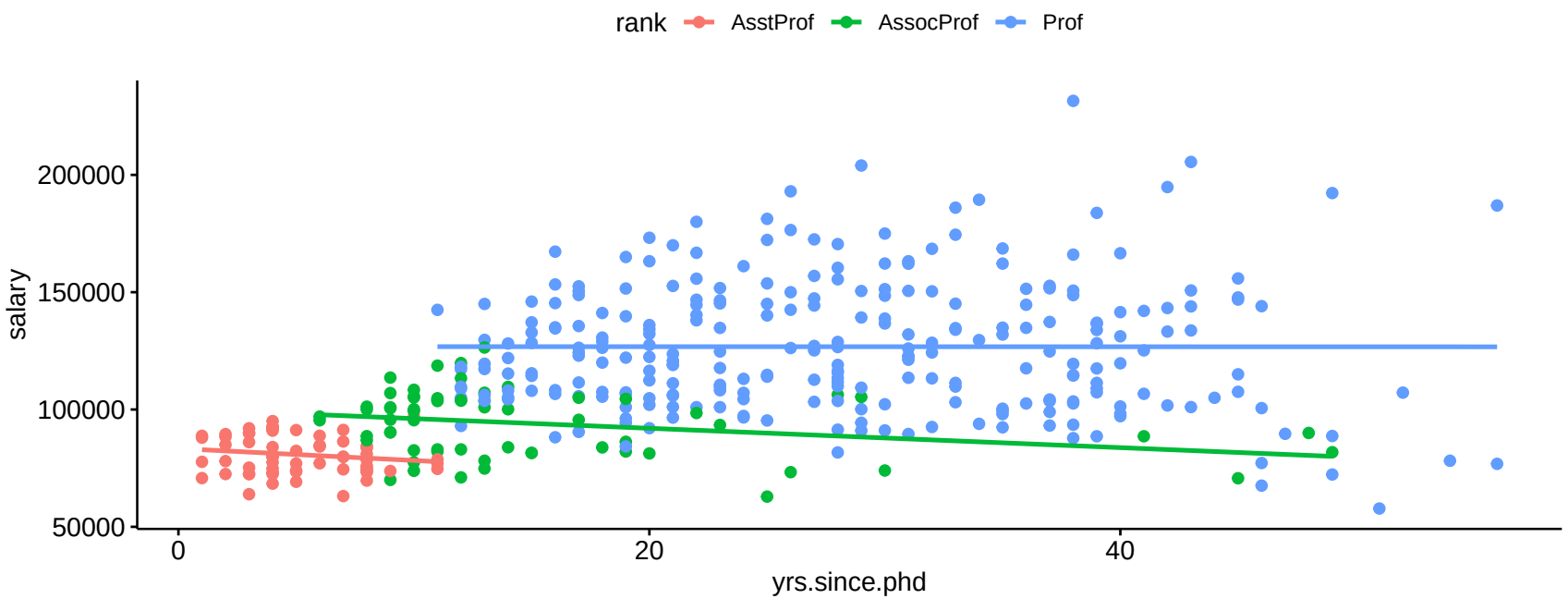




$$R^2 = \frac{\text{Explained}}{\text{Explained} + \text{Unexplained}} = (.00)^2 = .00$$

# Exploring Interaction Effects

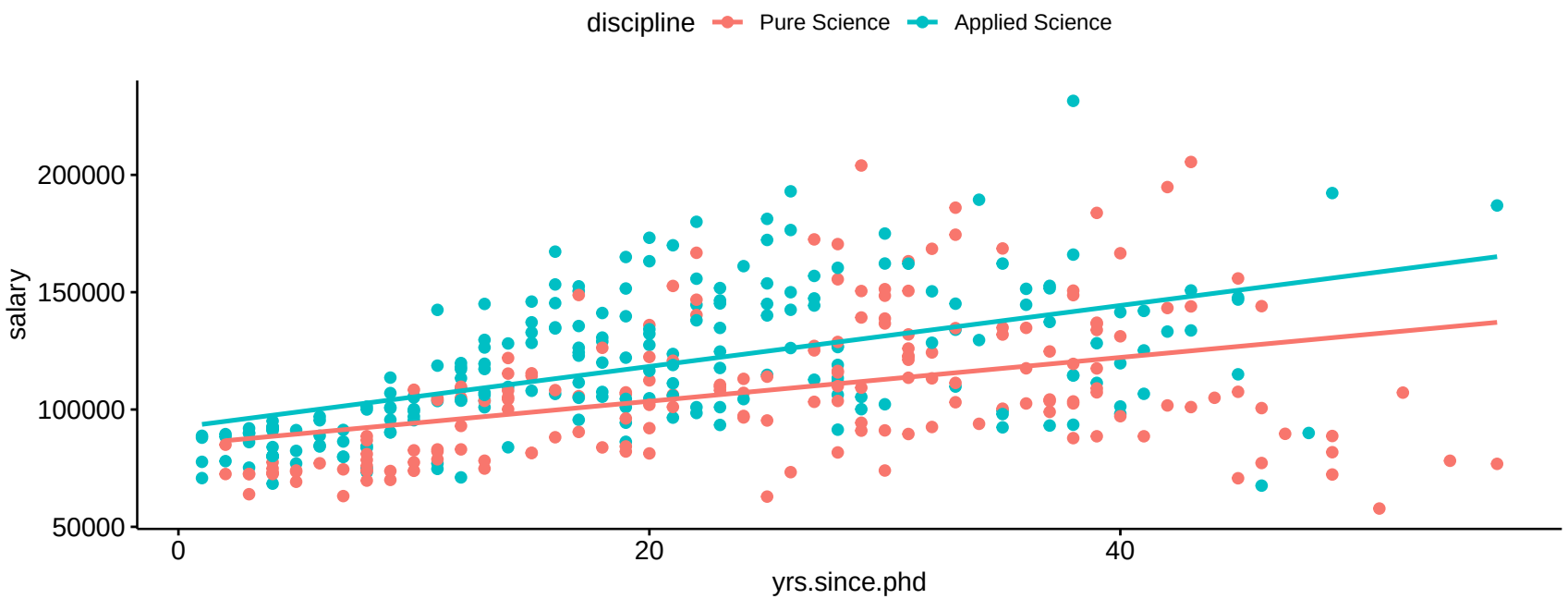
# Exploring Interaction Effects

```
ggscatter(dat, x = "yrs.since.phd",
                y = "salary",
          add = c("reg.line"), col= "rank")
```



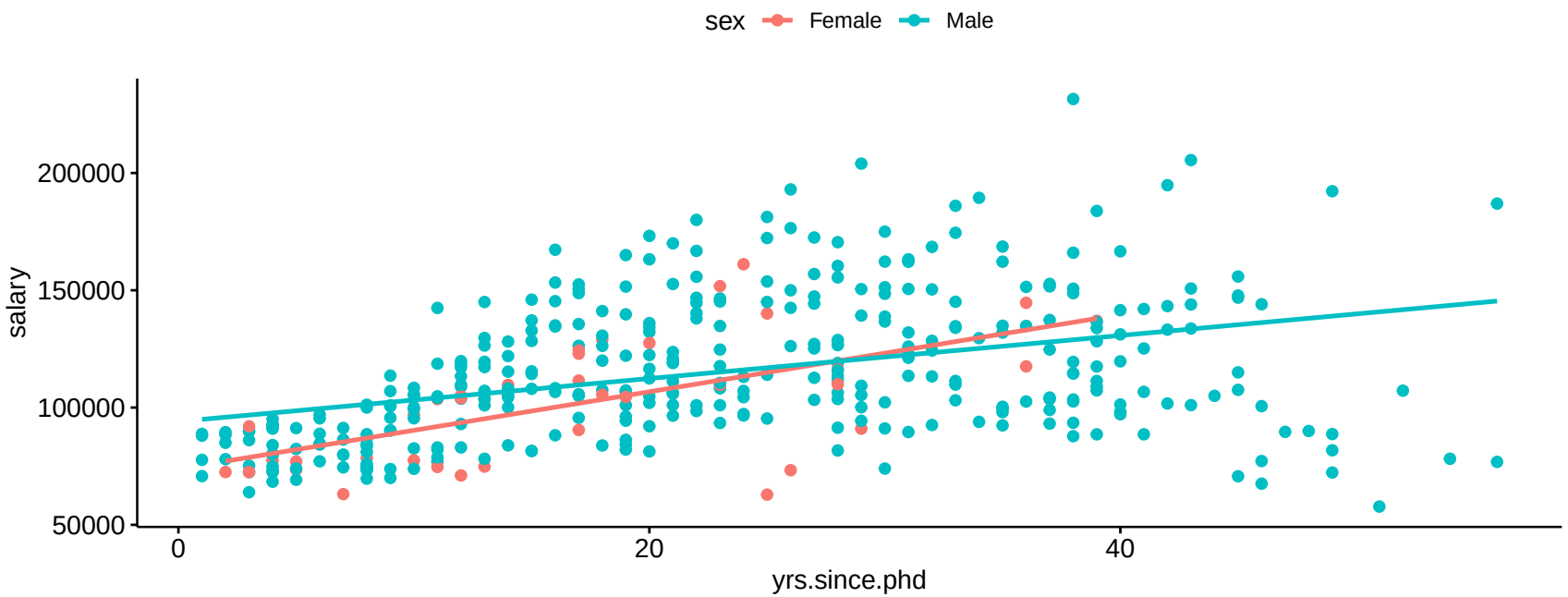
# Exploring Interaction Effects

```
1 ggscatter(dat, x = "yrs.since.phd",
2                y = "salary",
3          add = c("reg.line"), col= "discipline")
```



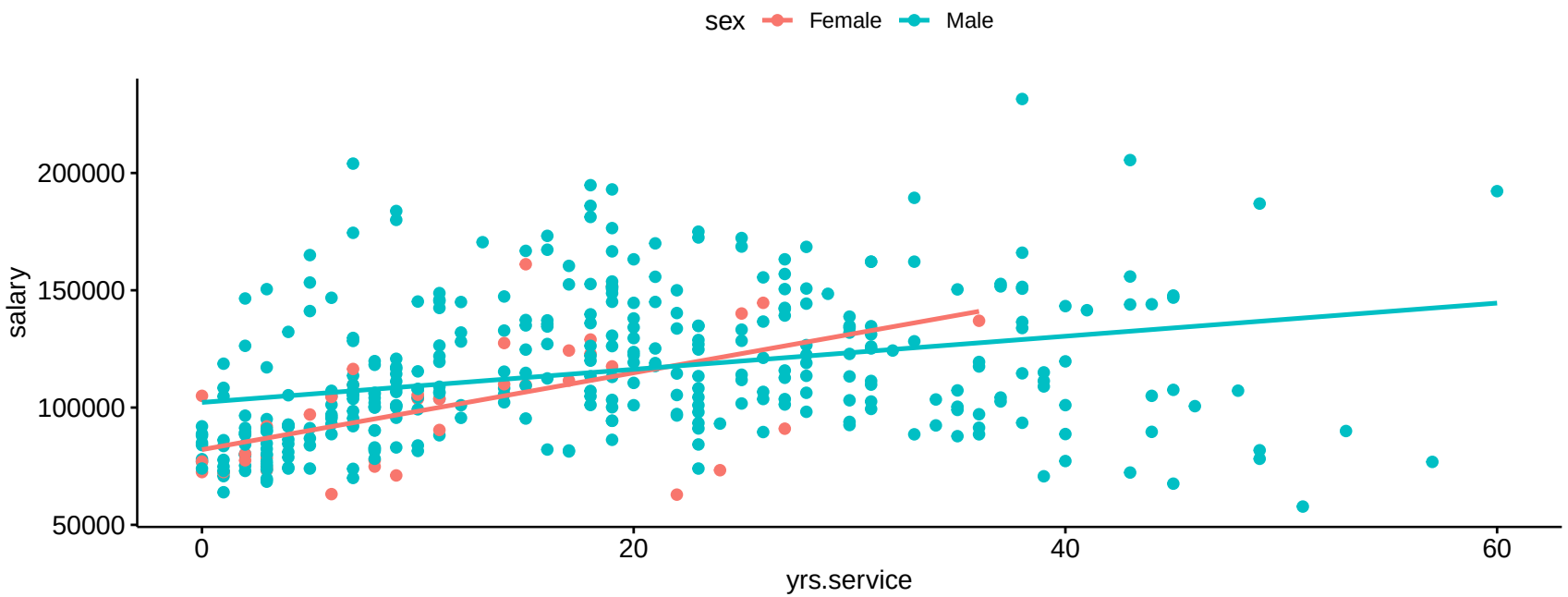
# Exploring Interaction Effects

```
ggscatter(dat, x = "yrs.since.phd",
                y = "salary",
          add = c("reg.line"), col= "sex")
```



# Exploring Interaction Effects

```
ggscatter(dat, x = "yrs.service",
                y = "salary",
          add = c("reg.line"), col= "sex")
```



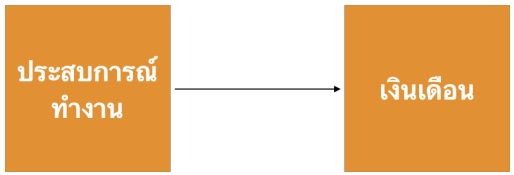
# สรุป

ผลการสำรวจข้อมูลข้างต้นพบว่า ...

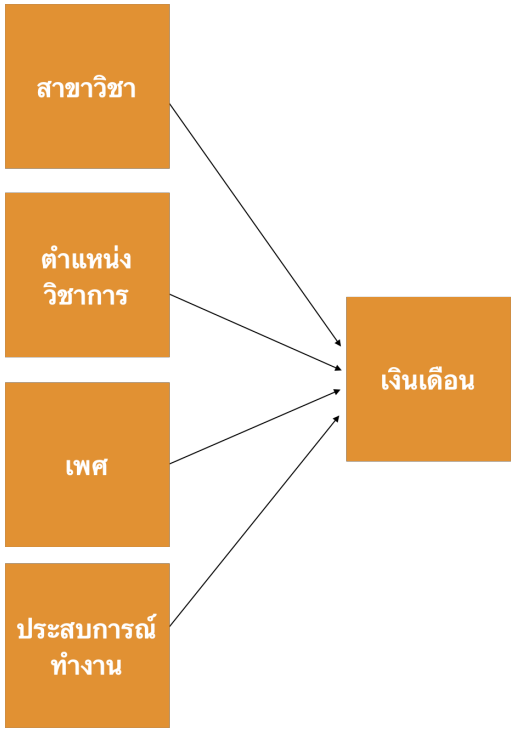
# Modelling

# Modelling: Regression analysis

## Job 1: Simple Regression




## Job 2: Multiple Regression

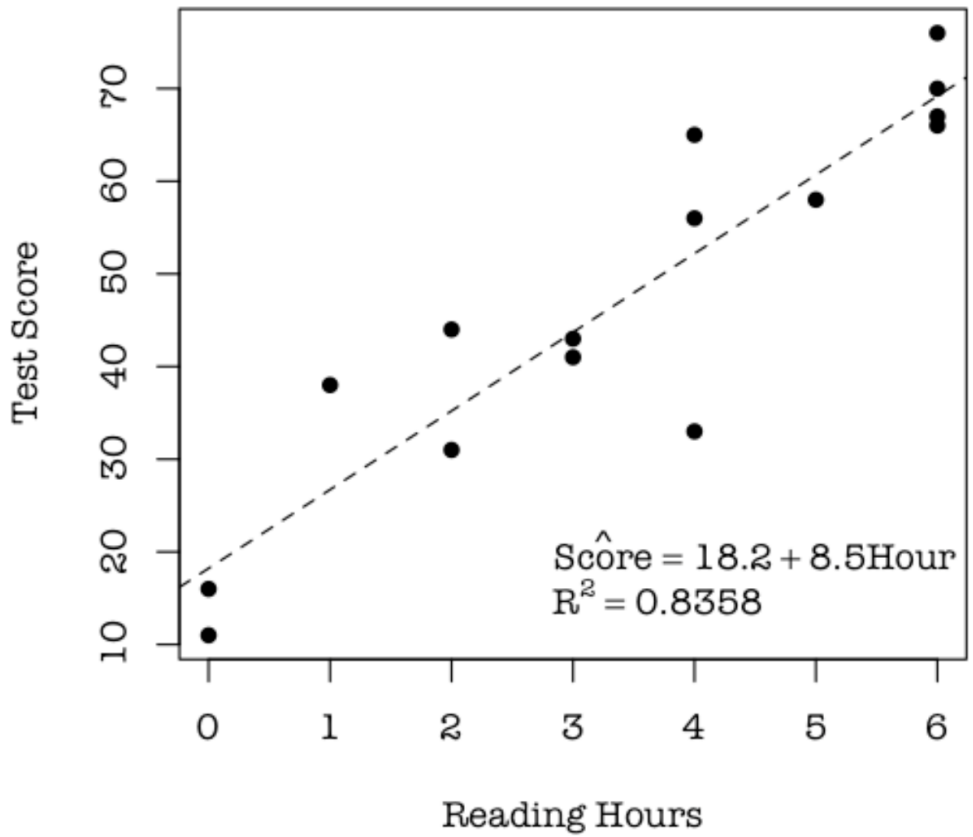


# วัตถุประสงค์ของการวิเคราะห์การถดถอย

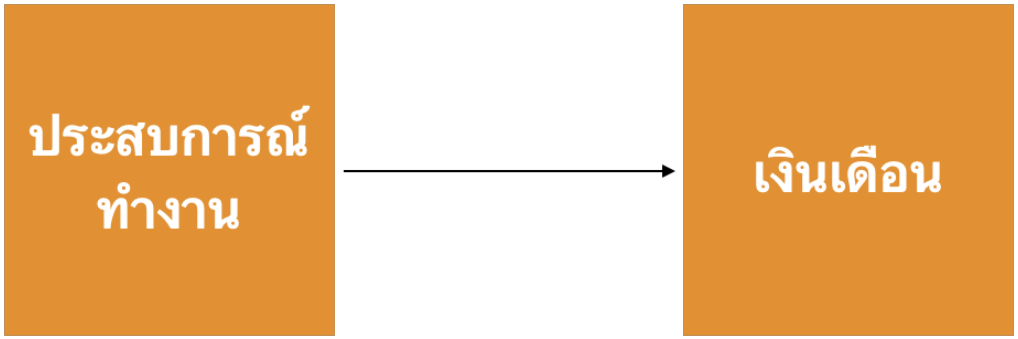
- เพื่ออธิบายความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ
- เพื่อทำนายแนวโน้มตัวแปรตามด้วยตัวแปรอิสระ

# Job 1: Simple Regression

# Basic Concept



# Simple Regression




$$\hat{salary} = b_0 + b_1 \times yrs.service$$

```
fit <- lm(salary ~ yrs.service, data = dat)
summary(fit)
```

# Calculate Simple Regression

```
Call:
lm(formula = salary ~ yrs.service, data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-81933 -20511  -3776  16417 101947

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  99974.7     2416.6   41.37  < 2e-16 ***
yrs.service    779.6      110.4    7.06 7.53e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

# แปลผลการวิเคราะห์ Simple Regression

## Linear Equation

**Slope-intercept form**

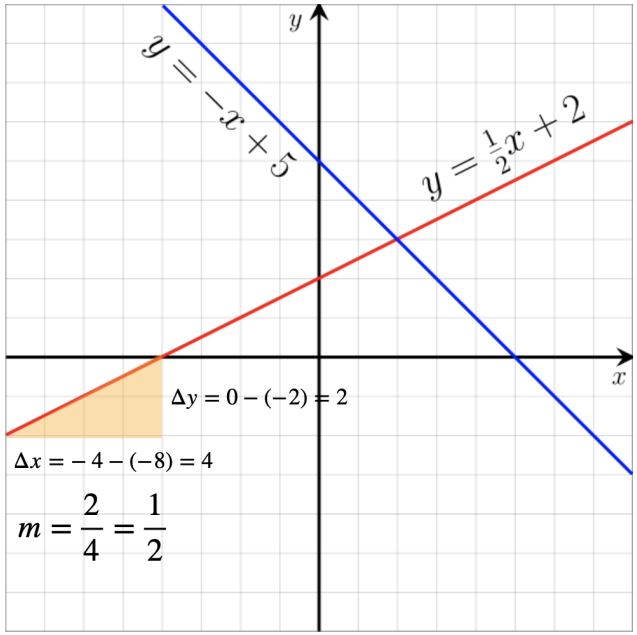
$$y = mx + c$$

Slope — Y-intercept (ค่าของ y เมื่อ x = 0)

- Slope
- Gradient
- Rate of change

$$m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

- อัตราการเปลี่ยนแปลงของ y เมื่อเทียบกับ x
- ถ้าตัวแปร x มีการเปลี่ยนแปลงเพิ่มขึ้น 1 หน่วย แล้วตัวแปร y จะมีการเปลี่ยนแปลง (เพิ่มหรือลด) $m$ หน่วย




https://en.wikipedia.org/wiki/Linear_equation#Slope–intercept_form_or_Gradient-intercept_form

**Figure 2:** ที่มา : สิวะโชติ ศรีสุทธิยากร (2564)

# แปลผลการวิเคราะห์ Simple Regression

เราสามารถแปลผลเพื่ออธิบายความสัมพันธ์
ระหว่างตัวแปรตามกับตัวแปรอิสระได้จากค่าสัมประสิทธิ์การถดถอย

- สัมประสิทธิ์ความชัน = 779.6
- สัมประสิทธิ์จุดตัดแกน y = 99974.7

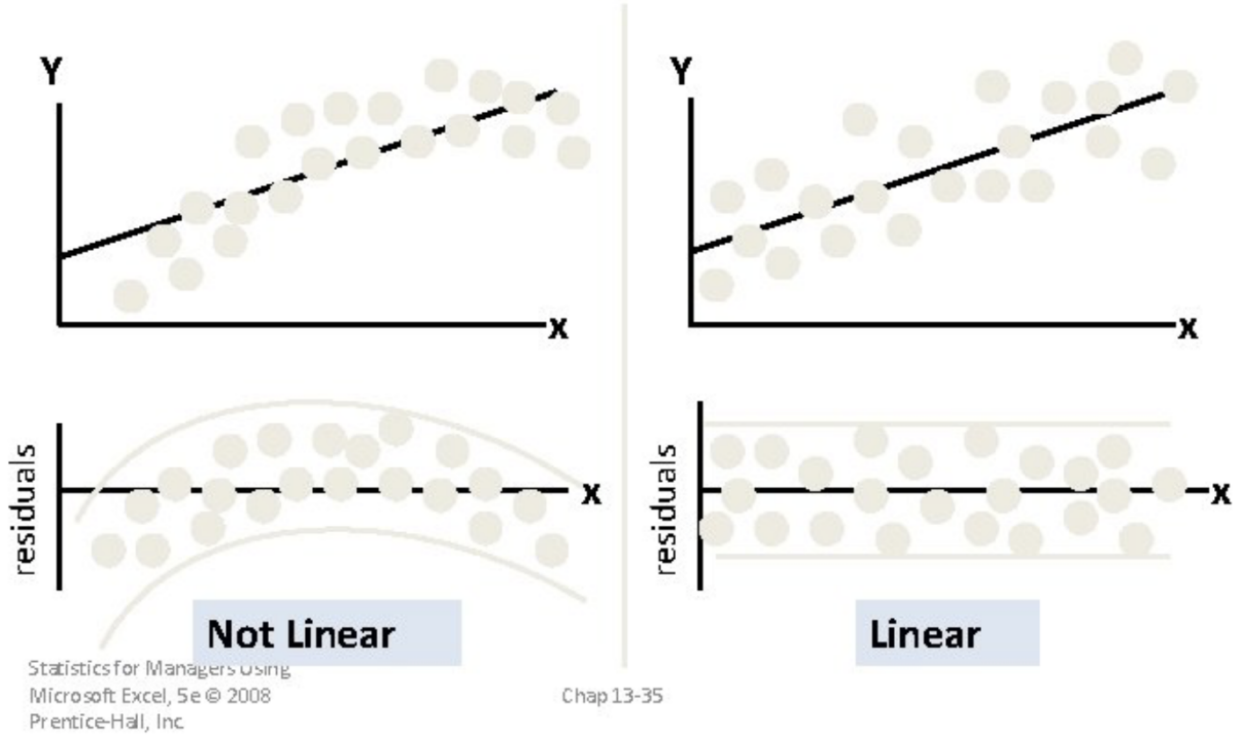
# Assumptions Checking

การตรวจสอบข้อตกลงเบื้องต้นของ regression จะใช้การวิเคราะห์เศษเหลือ (residual analysis)

- Linearity
- Normality
- Homoscedasticity

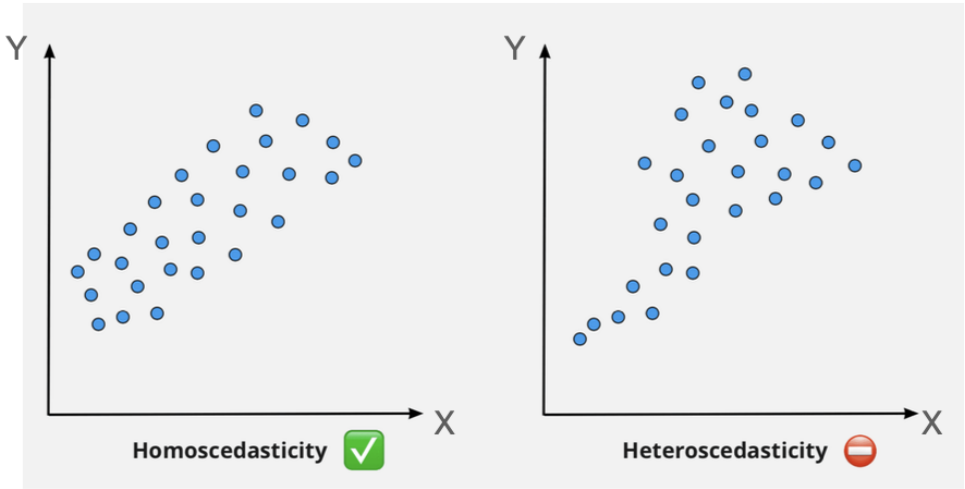
# Linearity

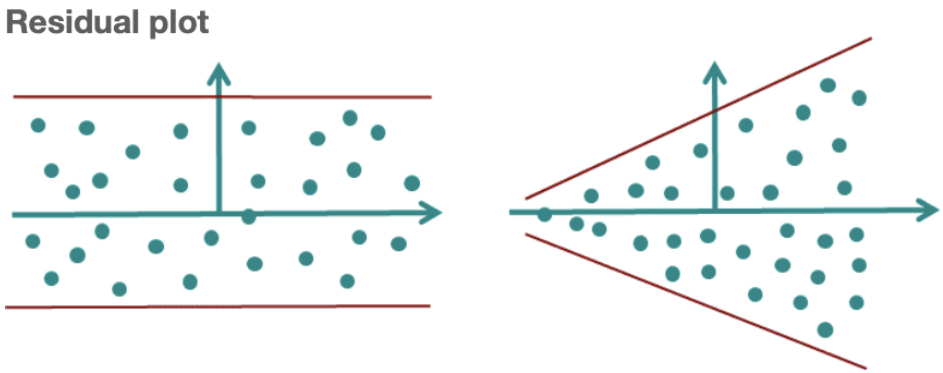
## Linearity




https://slidetodoc.com/linear-regression-example-data-house-price-in-1000/

# Homoscedasticity



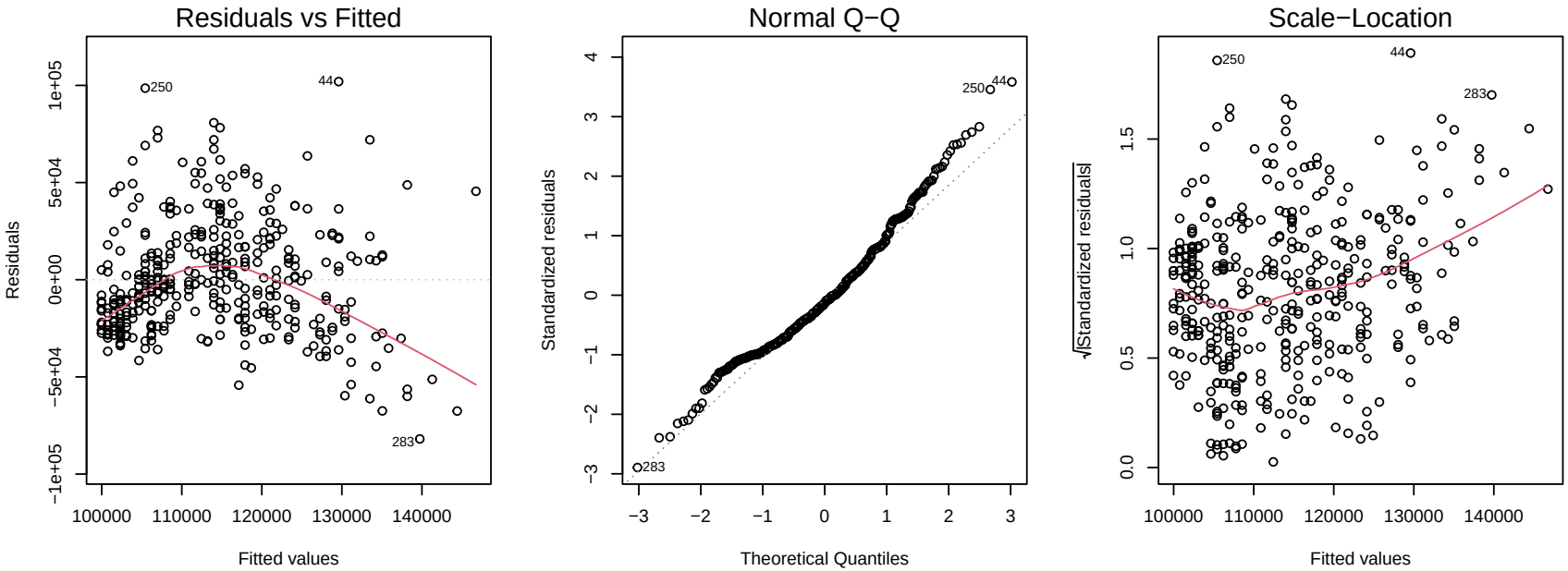

https://i1.wp.com/dataaspirant.com/wp-content/uploads/2020/12/10-Homoscedasticity-Vs-Heteroscedasticity.png?ssl=1

https://datatab.net/tutorial/linear-regression

# Calculate Residual Plots

```
1 par(mfrow=c(1,3))
2 plot(fit, 1:3)
```

# Calculate Residual Plots



# การคำนวณค่าทำนายจากสมการถดถอย

เราสามารถใช้ฟังก์ชัน `predict()`
เพื่อคำนวณค่าทำนายของตัวแปรตามจากสมการถดถอยที่ประมาณค่าได้

```
#first 6th predicted values
predict(fit) %>% head()
```

```
       1        2        3        4        5        6
114006.9 112447.8 102313.4 130377.8 131937.0 104652.1
```

# การคำนวณค่าทำนายจากสมการถดถอย

ถ้ามีชุดข้อมูลใหม่เราสามารถนำมาทำนายเงินเดือนของอาจารย์มหาวิทยาลัยด้วยฟังก์ชัน `predict()` เช่นเดียวกัน ดังนี้

```
1 new_dat<-data.frame(yrs.service = c(2,3,6,9,10))
2 new_dat
```

```
  yrs.service
1           2
2           3
3           6
4           9
5          10
```

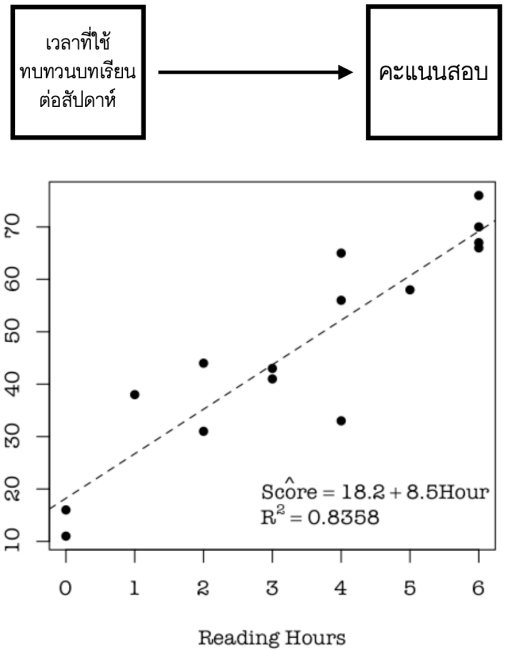
# การคำนวณค่าทำนายจากสมการถดถอย

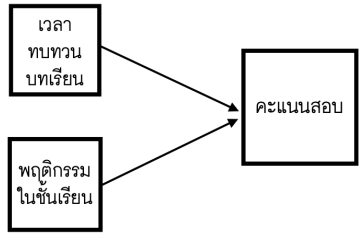
```
predict(fit, new_dat)
```

```
       1        2        3        4        5
101533.8 102313.4 104652.1 106990.8 107770.3
```

# Job 2: Multiple Regression

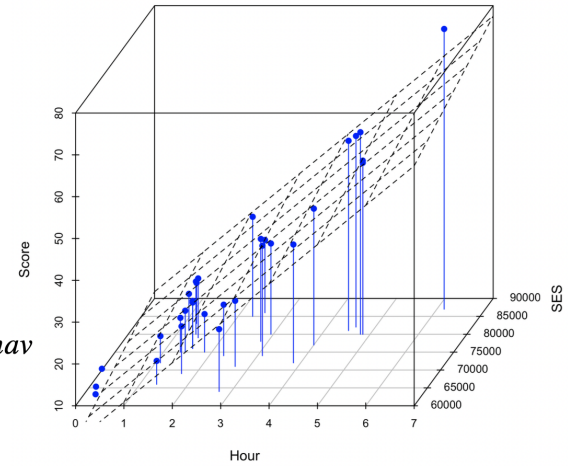
# Basic concept




**Regression equation**

$$\hat{Score} = b_0 + b_1 Hour + b_2 Class.Behav$$



# Calculate Multiple Regression

```
# multiple regression model specification
fit_multireg <- lm(salary ~ yrs.service + sex + rank, data
# all-in
fit_multireg <- lm(salary ~ ., data = dat)
summary(fit_multireg)
```

# Calculate Multiple Regression

```
Call:
lm(formula = salary ~ ., data = dat)

Residuals:
   Min      1Q  Median      3Q     Max
-65248 -13211  -1775   10384   99592

Coefficients:
                           Estimate Std. Error t value Pr(>
(Intercept)                 65955.2     4588.6  14.374  < 2e
rankAssocProf               12907.6     4145.3   3.114  0.00
rankProf                    45066.0     4237.5  10.635  < 2e
disciplineApplied Science   14417.6     2342.9   6.154 1.88e
yrs.since.phd                 535.1      241.0   2.220  0.02
```

# การแปลผลสมการถดถอย

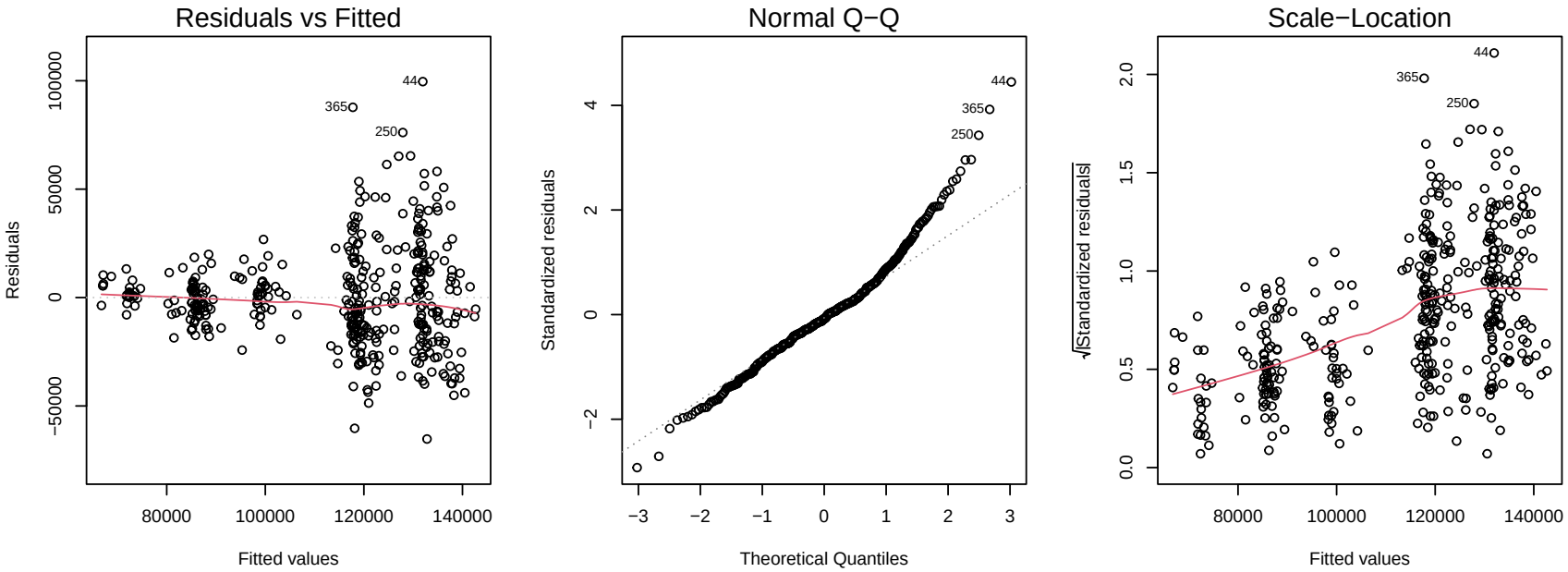
ในทำนองเดียวกับ การวิเคราะห์ความถดถอยอย่างง่าย
การอธิบายความสัมพันธ์ระหว่างตัวแปรตาม กับตัวแปรอิสระหลาย ๆ ตัว ใน multiple regression จะพิจารณาจากค่าสัมประสิทธิ์การถดถอย ดังนี้

# Assumptions Checking

- Linearity
- Normality
- Heteroscedasticity
- No Multicollinearity

```
par(mfrow=c(1,3))
plot(fit_multireg, 1:3)
```

# Residual Analysis



# Variance Inflation Factor (VIF)

```
1 library(DescTools)
2 VIF(fit_multireg)
```

```
                    GVIF Df GVIF^(1/(2*Df))
rank            2.013193  2        1.191163
discipline      1.064105  1        1.031555
yrs.since.phd   7.518936  1        2.742068
yrs.service     5.923038  1        2.433729
sex             1.030805  1        1.015285
```

# Refit the model

```
fit_multireg2 <- lm(salary~ . -yrs.since.phd, data=dat)
summary(fit_multireg2)
vif(fit_multireg2)
plot(fit_multireg2)
```

# Refit the model again

```
fit_multireg3 <- lm(log(salary)~ . -yrs.service, data=dat)
summary(fit_multireg3)
vif(fit_multireg3)
par(mfrow=c(1,3))
plot(fit_multireg3, 1:3)
```

# Refit the model again and again

```
fit_multireg4 <- lm(log(salary)~ . -yrs.service +
                      yrs.since.phd*sex,
                    data=dat)
summary(fit_multireg4)
vif(fit_multireg4)
plot(fit_multireg4)
head(dat)
```