R：Statistical Programming Methods
R：程式、機率與統計

Simple Linear Regression

Simple Linear Regression

- Simple linear regression provides a model of the relationship between the magnitude of one variable and that of a second
- To measure the relationship – correlation also does the same trick!
- The difference is that while correlation measures the *strength* of an association between two variables, regression quantifies the *nature* of the relationship.

Regression Equation

- $y = \beta_0 + \beta_1 x_1$
  - y: Dependent variables 應變數
  - x: Independent variables 自變數
  - β_0: Intercept 截距
  - β_1: Coefficients 係數 (in this case, the slope 斜率)

```
df <- read.csv("behavior.csv", header=TRUE) #from week 3

#correlation between score and sleep
cor(df$sleep,df$sport)
## [1] -0.1134812
# simple linear regression
model <- lm(sleep~sport, data=df)
```

→ $Sleep = \beta_0 + \beta_1 sport$

What does it mean?

```
summary(model)
## Call:
## lm(formula = sleep ~ sport, data = df)
## Residuals:
##      Min      1Q  Median      3Q     Max
## -5.1150 -1.5085  0.1166  1.3313  5.6690
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.28238    1.08820   7.611  1.7e-11 ***
## sport       -0.10713    0.09475  -1.131    0.261
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.079 on 98 degrees of freedom
## Multiple R-squared:  0.01288,    Adjusted R-squared:  0.002805
## F-statistic: 1.279 on 1 and 98 DF,  p-value: 0.2609
```

$$sleep = 8.28238 - 0.10713 \times sport$$

But the result is NOT statistically significant

Fitted Values and Residuals

- Fitted value
  - Prediction (Based on the model to predict y)
  - $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1$

```
#fitted value
fitted <- predict(model)
```

- Residuals
  - Prediction errors (difference between prediction and actual value)
  - $\hat{e} = \mathrm{y} - \hat{y}$

```
#residual
residual <- residuals(model)
```

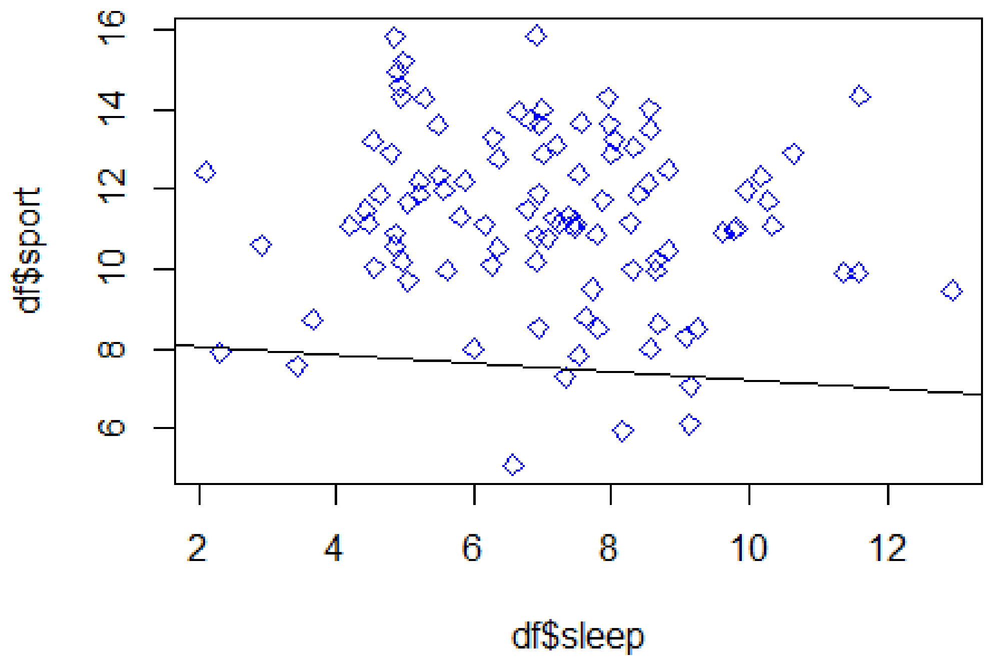
Least squares / Residual sum of squares最小平方法

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i)^2$$

- Linear regression is to minimize the sum of squared residual values

Scatterplot

```
plot(df$sleep, df$sport,
        col="blue",
        pch=23)
abline(lm(sleep~sport, data=df))
```



Prediction v.s. Explanation (profiling)

- Conclusions about causation must come from a broader understanding about the relationship.
- Which one is the cause and which one is the outcome?

- With the advent of big data, regression is widely used to form a model to predict individual outcomes for new data (i.e., a predictive model) rather than explain data in hand.

Multiple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + e$$

age: house age (numeric, in year)

MRT:distance to the nearest MRT station (numeric)

stores: number of convenience stores (numeric)

Latitude: latitude (numeric)

Longitude: longitude (numeric)

price: unit price per area (numeric)

```
df2 <- read.csv("house.csv", header=TRUE)
house_lm <- lm(unitprice~age+stores+MRT, data=df2)
summary(house_lm)
```

Multiple Linear Regression

- Root mean squared error (RMSE)
  - The square root of the average squared error of regression
- R-squared
  - The proportion of variance explained by the model, from 0 to 1.

```
## Call:
## lm(formula = unitprice ~ age + stores + MRT, data = df2)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -37.304  -5.430  -1.738   4.325  77.315
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.977286   1.384542  31.041  < 2e-16 ***
## age         -0.252856   0.040105  -6.305 7.47e-10 ***
## stores       1.297443   0.194290   6.678 7.91e-11 ***
## MRT         -0.005379   0.000453 -11.874  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.251 on 410 degrees of freedom
## Multiple R-squared:  0.5411, Adjusted R-squared:  0.5377
## F-statistic: 161.1 on 3 and 410 DF,  p-value: < 2.2e-16
```

Practice

- Life Expectancy (WHO) | Kaggle
- Check the following regression and identify which factor would have effect on life expectancy (and how is the effect)

$$life\ expectancy = \beta_0 + \beta_1 Adult\ Mortality + \beta_2 infant\ deaths + \beta_3 Alcohol + \beta_4 \mathrm{BMI} + \beta_5 \mathrm{GDP} + \beta_6 Schooling + \beta_7 Population + e$$