Memo for Advanced Transportation Planning

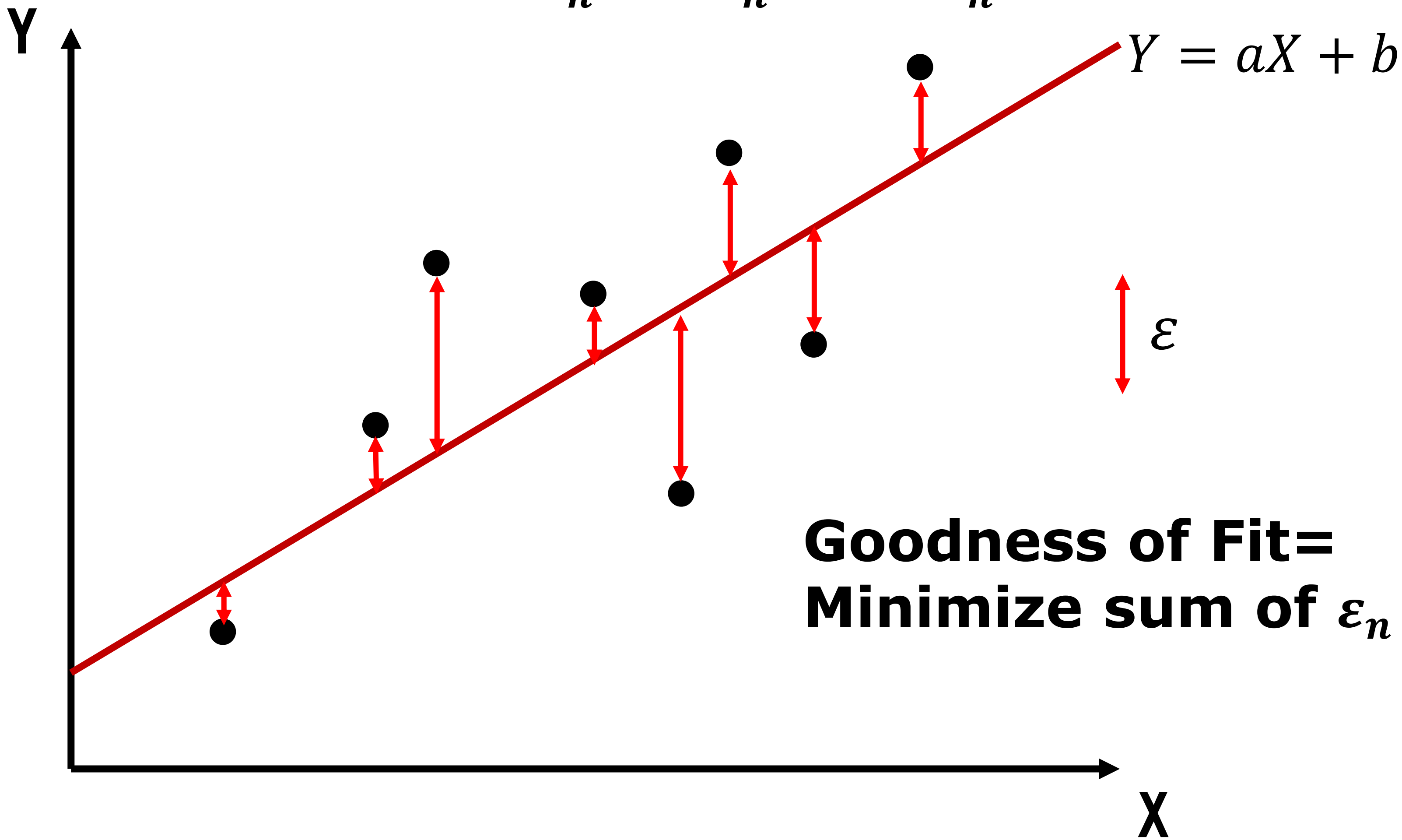
1) Multiple Regression Model

2) Ordinary Least Squares (OLS) and Maximum Likelihood (ML)

Regression Model & Ordinary Least Squares: OLS

Estimate unknown parameters a, b with

$$\boldsymbol{Y_n = aX_n + b + \varepsilon_n}$$



Goodness of Fit= Minimize sum of $\boldsymbol{\varepsilon_n}$

How to minimize sum of ε_n ?

$$S = \sum_{n=1}^{N} \varepsilon_n = \sum_{n=1}^{N} \left(Y_n - (aX_n + b)\right)$$

→ ε_n has positive & negative → Inconvenience

$$S = \sum_{n=1}^{N} |\varepsilon_n| = \sum_{n=1}^{N} \left|\left(Y_n - (aX_n + b)\right)\right|$$

→ Absolute value is difficult to solve → reject

$$S = \sum_{n=1}^{N} \varepsilon_n^2 = \sum_{n=1}^{N} \left(Y_n - (aX_n + b)\right)^2$$

→ This equation is easy to solve

This equation is called Ordinary Least Squares (OLS). Another names is Minimizing Sum of Residuals

How to estimate unknown parameters.
Target equation is as follows.

$$S = \sum_{n=1}^{N} \left(Y_n - (aX_n + b)\right)^2$$

$$\min_{a,b} S$$

Solving two partial differential equations, such as $\frac{\partial S}{\partial a} = 0$ **and** $\frac{\partial S}{\partial b} = 0$ **. These are linear binary simultaneous equation. So,**

$$\hat{a} = \frac{N\overline{X}\overline{Y} - \sum_{n=1}^{N} X_n Y_n}{N\overline{X}^2 - \sum_{n=1}^{N} X_n^2}$$

$$\hat{b} = \overline{Y} - \hat{a}\overline{X}$$

Please confirm the above derivation !

What is "good model" as for regression model?

-Goodness of Fit is evaluated by coefficient of determination or R squared.

$$R^2 = 1 - \frac{\sum_n \varepsilon_n^2 / N}{\sigma_Y^2} = 1 - \frac{\mathit{Mean\ of\ residual\ square}}{\mathit{variance\ of\ Y}}$$

R^2 satisfies $0 <= R^2 <= 1$, and closing 1 means better fitness. Usually over 0.7 would be desirable.

-Statistical significance of parameters
Absolute value of t-value or t-statistics should be over 1.96 → 95% significant level

-Sign condition is also important

```
Call:
lm(formula = x$mileage ~ ., data = x)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.400  -1.919  -0.061   1.885  41.407 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.890e+01  4.126e-01  45.819  < 2e-16 ***
daily_km     3.330e-02  1.284e-03  25.939  < 2e-16 ***
hybrid       6.682e+00  9.721e-02  68.745  < 2e-16 ***
displace_cc -5.886e-01  1.205e-01  -4.884 1.05e-06 ***
weight_kg   -4.992e+00  2.297e-01 -21.732  < 2e-16 ***
age_month   -1.310e-02  8.659e-04 -15.126  < 2e-16 ***
temp_ave     4.627e-02  4.662e-03   9.926  < 2e-16 ***
gasprice     1.693e-03  2.492e-03   0.679    0.497    
pop_density -2.442e-04  3.069e-05  -7.959 1.92e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.81 on 9991 degrees of freedom
Multiple R-squared:  0.491,     Adjusted R-squared:  0.4906 
F-statistic:  1205 on 8 and 9991 DF,  p-value: < 2.2e-16
```

What is “t-value” ?
We can judge whether the estimated parameter is statistically apart from “0”.
If the absolute t-value is more than 1.96, it is not “0” as 95% significant level.
https://en.wikipedia.org/wiki/T-statistic

“Multiple R-squared” → $R^2 = 1 - \frac{\sum_n \varepsilon_n^2 / N}{\sigma_y^2}$

“Adjusted R-squared” → $\bar{R}^2 = 1 - \frac{\frac{\sum_n \varepsilon_n^2}{N-K-1}}{\frac{\sigma_y^2}{N-1}}$

Explanation of “Degree of Freedom”:
https://statisticsbyjim.com/hypothesis-testing/degrees-freedom-statistics/

Example by "Vehicle Fuel Consumption Survey" by Ministry of Land, Infrastructure, Transport & Tourism

- Starting 2007, every month survey
- "gas_10k.csv" is passenger car data sampled 10,000 record
- It contains the following 11 variables

1 year: survey year
2 month: survey month
3 mileage: [km/L]
4 daily_km: vehicle km_per day during survey
5 hybrid: hybrid dummy variable
6 displace_cc:engine displacement [cc]
7 weight_kg: vehicle weight [kg]
8 age_month: age of vehicle [month]
9 temp_ave: average temperature of registered place [Celsius degree]
##10 gasprice: gasoline price on surveyed month & place [yen/L]
##11 pop_density: population density at surveyed place [persons/km^2]

- Please estimate good model which explains mileage

mileage = β_1*daily_km + β_2*hybid + β_3*weight_kg …

```
rm(list=ls())
setwd("d:/usr/doc/dropbox/daigakuin/")

dt <- read.csv("gas_10k.csv",header=T)
str(dt)
summary(dt)
dt[,6] <- dt[,6]/1000  ## displacement unit to "litter"
dt[,7] <- dt[,7]/1000  ## weight unit to "ton"

x <- dt[,3:11]

par(ask=T)

pairs(x)

hist(x$mileage)
plot(density(x$mileage))
boxplot(x$mileage)
boxplot(x$mileage~x$hybrid)

plot(x$weight_kg,x$mileage)
round(cor(x),digits=3)
```

```
res1 <- lm(x$mileage~.,data=x)

res2 <- lm(x$mileage~(x$daily_km + x$hybrid + x$displace_cc +
           x$weight_kg + x$age_month + x$temp_ave +
           x$gasprice + x$pop_density)^2)
```

The relationship between mileage and temperature would be non-linear.
→ How to introduce this relationship ?

By using “res4”

$$
\begin{aligned}
&-0.00492 \times temp^2 + 0.187 \times temp = \\
&-0.00492(temp^2 - 38.01temp) \\
&-0.00492(temp - 19.0)^2 + 1.78
\end{aligned}
$$

Other method to improve goodness of fit is
- Dividing parameters “>= 2000 kg” and “<2000 kg”
- Introducing “log transform”
- BoxCox transformation !
  → https://www.youtube.com/watch?v=vGOpEpjz2Ks

OLS vs. Maximum Likelihood (ML)

Ex.1) I rolled a dice 10 times, “1” comes out 3 times. Please estimate the unknown parameter θ, probability of “1” comes out

Of course the result is 3/10.

How to estimate by ML

$\boldsymbol{L}^{*} = {}_{10}C_3\theta^3(1-\theta)^{10-3}$ → **Likelihood, solve** $\max_{\boldsymbol{\theta}} \boldsymbol{L}^{*}$

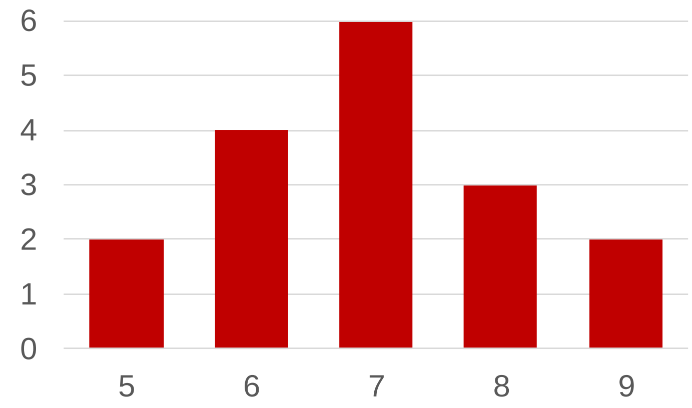
$\ln L^{*} = \boldsymbol{L} = \ln {}_{10}C_3 + \mathbf{3}\ln\boldsymbol{\theta} + \mathbf{7}\ln(1-\theta)$

→ **Log-Likelihood**

$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{\theta}} = \frac{\mathbf{3}}{\boldsymbol{\theta}} - \frac{\mathbf{7}}{\mathbf{1}-\boldsymbol{\theta}} = \mathbf{0}$ → $\boldsymbol{\theta} = \frac{\mathbf{3}}{\mathbf{10}}$ →**Maximum log-Likelihood**

Ex.2) The histogram of a paper test was as right graph. 10 points and $2+4+6+3+2=$ 17 students.

We assume normal distribution, and estimate mean value by ML.



$$\boldsymbol{f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma}}}$$

$$\boldsymbol{L^* = f(5)^2 \times f(6)^4 \times f(7)^6 \times f(8)^3 \times f(9)^2}$$

$$\ln(L^*) = L = 2\ln f(5) + 4\ln f(6) + 6\ln f(7) + 3\ln f(8) + 2\ln f(9)$$

$$\ln \boldsymbol{f(x)} = -\frac{\ln(2\pi)}{\boldsymbol{2}} - \ln \boldsymbol{\sigma} - \frac{\boldsymbol{(x-\mu)^2}}{\boldsymbol{2\sigma}}$$

Here, we assume σ as any constant value and omit it from the previous equation, so…

$$L = -2(5-\mu)^2 - 2(6-\mu)^4 - 2(7-\mu)^6 - 2(8-\mu)^3 - 2(9-\mu)^2$$

$$\frac{\partial L}{\partial \mu} = 4(5-\mu) + 8(6-\mu) + 12(7-\mu) + 6(8-\mu) + 4(9-\mu) = 0$$

$$\rightarrow 236 - 34\mu = 0 \quad \rightarrow \quad \mu = \frac{236}{34} = 6.94$$

Similarly, we apply ML to regression model

$y_n = X\beta + \varepsilon_n$ ***X→ (N*K), β → (K)***

We assume ε_n as normal distribution with $N(0, \sigma^2)$

$$f(\varepsilon_n) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y_n - X\beta)^2}{2\sigma}}$$

$$L^* = \prod_{n=1}^{N} f(\varepsilon_n) \quad \rightarrow \quad \ln L^* = L = \sum_{n=1}^{N} \ln f(\varepsilon_n)$$

$$L = \sum_{n=1}^{N} \left[-\frac{\ln(2\pi)}{2} - \ln \sigma - \frac{(y_n - X\beta)^2}{2\sigma} \right]$$

As well, σ is assumed constant (actually, by estimated β, we can calibrate σ easily).

$$L = \sum_{n=1}^{N} -(y_n - X\beta)^2 \quad \rightarrow \quad \max_{\beta} L = \min_{\beta} -L$$

So, if we assume normal distribution as probabilistic distribution of residual ε_n, the final equation of ML equals OLS equation.

However, there are many cases which ε_n is not normally distributed → OLS is not adequate !

ML can be applied even this case → ML can cover wide range as for modelling !

Homework 💀

Please estimate better model for “gas_10k.csv” data.

And summarize the result (only one case) on 1 page Power Point file. Next week you would introduce your result !