Memo for Advanced Transportation Planning

1) Various Regression Models

2) Exercise for Multiple Regression Model by freight statistics

Various Regression Models

a) Cobb-Douglas production function

$Y = AL^{\alpha}K^{\beta}$ → product function （乗法関数）

Y : Production volume

A : Productivity factor

L : Labor, K : Investment, e.g. factory area

α, β : unknown parameters

Three parameters can be estimated easily

$\ln Y = A' + \alpha \ln L + \beta \ln K$ → summation function

（加法関数）

The importance of $\hat{\alpha}$ and $\hat{\beta}$.

Let's consider *K* and *L* are doubled.

$$Y' = A(2L)^{\alpha}(2K)^{\beta} = A(L)^{\alpha}(K)^{\beta} \times 2^{\alpha+\beta}$$
$$= Y \times 2^{\alpha+\beta}$$

$\alpha + \beta > 1$ → Scale of economics increasing

$\alpha + \beta = 1$ → Scale of economics neutral

$\alpha + \beta < 1$ → Scale of economics decreasing

We can judge the scale of economics by the estimated parameters.

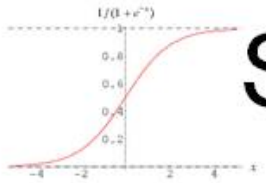
Developing era: $\alpha + \beta > 1$

Saturating era: $\alpha + \beta < 1$

b) Logistic Curve

Diffusion of durable consumer goods

Ex) Smartphone, 1970's CCC in Japan



Logistic curve function:

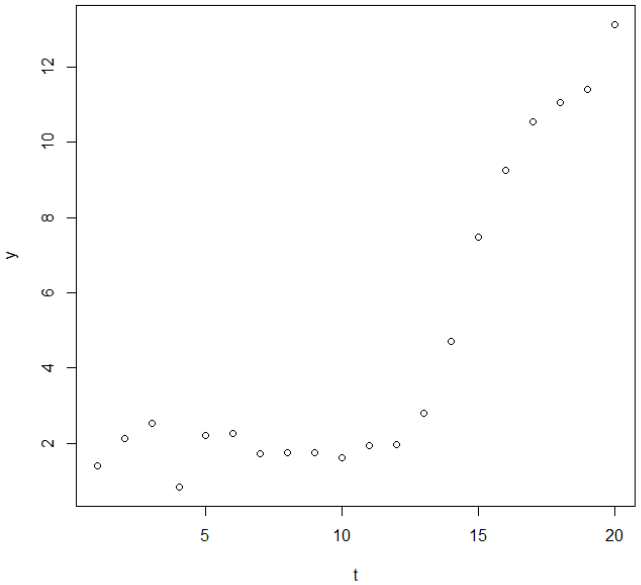
$$Y(t) = \frac{c}{1 + e^{-a+b}}$$ - at + b

$$Y(t) = \frac{c}{1+e^{-a+b}} \rightarrow \frac{c}{Y(t)} = 1 + e^{-a+b}$$

$$\rightarrow \frac{c-Y(t)}{Y(t)} = e^{-a+b} \rightarrow \ln\frac{c-Y(t)}{Y(t)} = -a + b$$

Let's estimate parameters:

```
set.seed(1234)
a <- 1; b <- 15; c <- 10
t <- seq(1,20)
e <- rnorm(20)/2
y <- c/(1+exp(-a*t+b)) + e + 2
plot(t,y)
```



```
for(n in 0:10){
  c <- ceiling(max(y)) + n
  yy <- log((c-y)/y)
  res <- lm(yy~t)
  cat("c value= ",c,"¥n")
  print(summary(res)$coefficients)
  print(summary(res)$adj.r.squared)
}
```

Best model is…

```
c value=  19
(Intercept)  3.247 (11.2)
t           -0.1723 (-7.1)
adj.R^2 0.722
```

To avoid the iterative method

→ Direct estimation as follows:

$$Y(t) = \frac{c}{1 + e^{-a+b}} + d + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

Insted of

$$\ln \frac{c - Y(t)}{Y(t)} = -a + b + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$$

```
dat <- data.frame(cbind(t,y))
res.nls <- nls(y~c/(1+exp(-a*t+b))+d, dat,
         start=list(a=1, b=10, c=10, d=1))
summary(res.nls)
```

nls → Nonlinear Least Squares

	Estimate	t value
a	0.948	(7.0)
b	14.21	(7.1)
c	10.39	(23.1)
d	1.755	(10.6)

$$Y(t) = \frac{19}{1+e^{-0.1723t+3.247}}$$ → logit transform case

$$Y(t) = \frac{10.39}{1+e^{-0.948t+14.21}} + 1.755$$ → nls case

よりoriginal equationに近い

$$Y(t) = \frac{10}{1+e^{-t+15}} + 2$$ → original equation

c) Generalized Least Squares (GLS)

Multivariate normal distribution:

$$X \sim N(\mu, \Sigma)$$

Σ：分散・共分散行列

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \cdots & \sigma_{1K}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{K1}^2 & \cdots & \sigma_{KK}^2 \end{pmatrix} = \begin{pmatrix} \sigma_{11}^2 & \cdots & \rho_{1K}\sigma_1\sigma_K \\ \vdots & \ddots & \vdots \\ \rho_{K1}\sigma_K\sigma_1 & \cdots & \sigma_{KK}^2 \end{pmatrix}$$

→ ***K*K*** **dimensional dispersion matrix**

$$f(X) = \frac{exp\left(-\frac{1}{2}(X-\mu)^t\Sigma^{-1}(X-\mu)\right)}{\sqrt{(2\pi)^K|\Sigma|}}$$

Multivariate normal distributionの

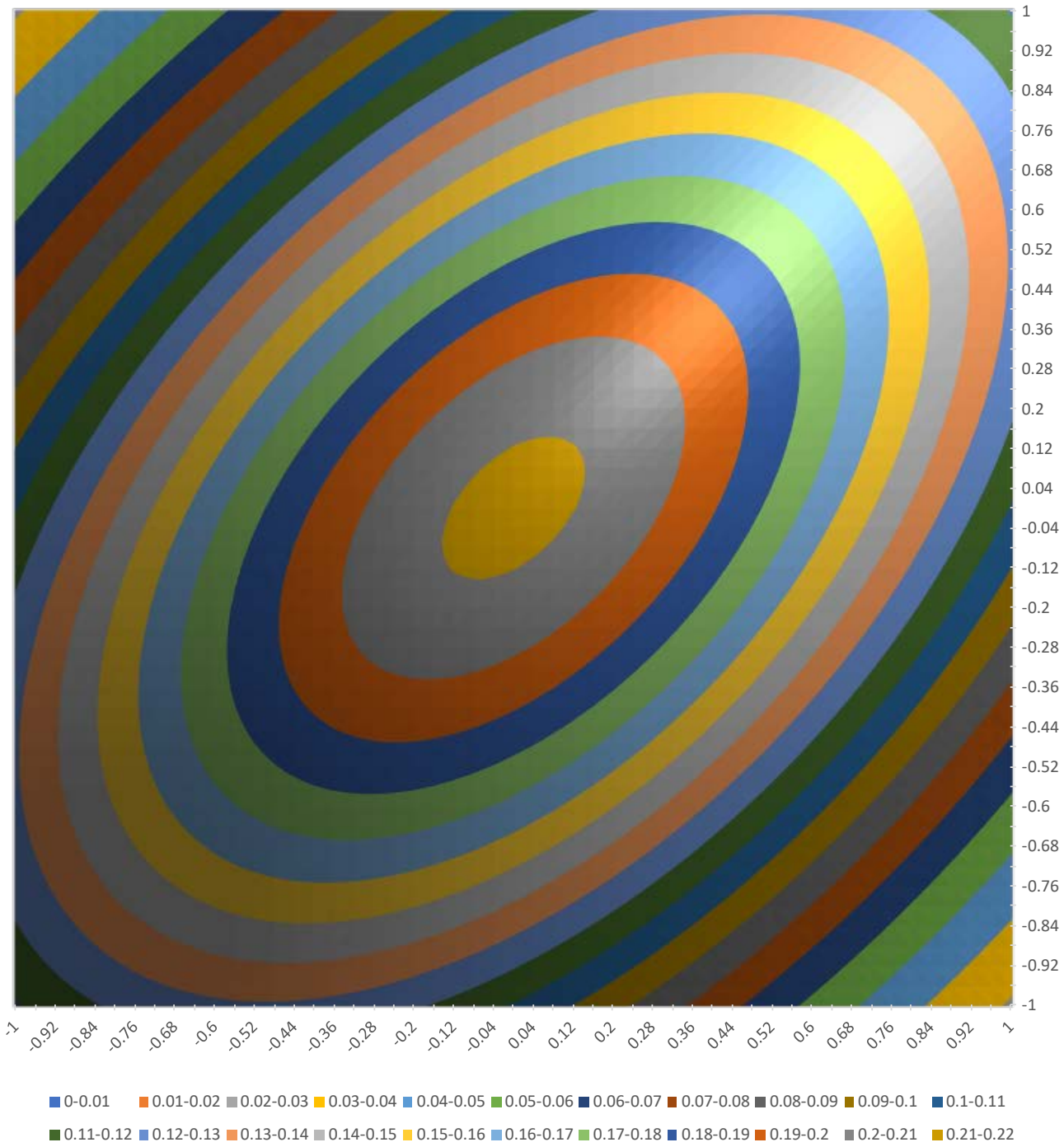
→ **Probabilistic density function**

In case of $K=2$,

$$X \sim N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right),\ |\Sigma| = 1-\rho^2$$

$$f(x_1, x_2) = \frac{exp\left(-\frac{1}{2}(x_1, x_2)\frac{1}{1-\rho^2}\begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}\right)}{2\pi\sqrt{1-\rho^2}}$$

$$f(x_1, x_2) = \frac{exp\left(-\frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{2(1-\rho^2)}\right)}{2\pi\sqrt{1-\rho^2}}$$


1
0.92
0.84
0.76
0.68
0.6
0.52
0.44
0.36
0.28
0.2
0.12
0.04
-0.04
-0.12
-0.2
-0.28
-0.36
-0.44
-0.52
-0.6
-0.68
-0.76
-0.84
-0.92
-1
-1 -0.92 -0.84 -0.76 -0.68 -0.6 -0.52 -0.44 -0.36 -0.28 -0.2 -0.12 -0.04 0.04 0.12 0.2 0.28 0.36 0.44 0.52 0.6 0.68 0.76 0.84 0.92 1
0-0.01
0.01-0.02
0.02-0.03
0.03-0.04
0.04-0.05
0.05-0.06
0.06-0.07
0.07-0.08
0.08-0.09
0.09-0.1
0.1-0.11
0.11-0.12
0.12-0.13
0.13-0.14
0.14-0.15
0.15-0.16
0.16-0.17
0.17-0.18
0.18-0.19
0.19-0.2
0.2-0.21
0.21-0.22

we assumed

$y_n = X\beta + \varepsilon_n$, $\varepsilon \sim N(0, \sigma^2)$

However,

a) ε **does not distribute constant** σ^2

→ **“heteroscedastic” dispersion （不均一分散）**

b) There are co-related ε **.** $E[\varepsilon_i \varepsilon_j] \neq 0$

→ **Don’t satisfy I.I.D. (Independent and Identically Distributed)（独立同一分布）**

Because of this violation of OLS assumption, OLS would derive biased parameters.

→ **GLS by ML is better way for parameter estimation.**

Again we assume,

$$y = X\beta + \varepsilon \ , \ \varepsilon \sim N(0, \Sigma^2)$$

$$f(y - X\beta) = \frac{exp\left(-\frac{1}{2}(y - X\beta)^t \Sigma^{-1}(y - X\beta)\right)}{\sqrt{(2\pi)^K |\Sigma|}}$$

$$L = \sum_{n=1}^{N} \ln f(y_n - X_n\beta) \rightarrow \max_{\beta,\Sigma} L$$

This ML estimation is quite complicated.

Practically, the following matrix would be applied sometimes.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\sigma_1^2} & \boldsymbol{\rho} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{\rho} & \boldsymbol{\sigma_2^2} & \boldsymbol{\rho} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\rho} & \boldsymbol{\sigma_3^2} & \boldsymbol{\rho} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\rho} & \boldsymbol{\sigma_4^2} & \boldsymbol{\rho} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\rho} & \boldsymbol{\sigma_5^2} \end{bmatrix}$$

パラメータσを推定しやすい

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\sigma_1^2} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\sigma_1^2} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\sigma_1^2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\sigma_2^2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\sigma_2^2} \end{bmatrix}$$

Exercise for Multiple Regression Model by freight statistics

Forecasting freight generation volume in 2045

Multiple Regression Model: $\boldsymbol{y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon}$

Summation of squared residual: $S = \boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{\beta X})^t(\boldsymbol{y} - \boldsymbol{\beta X})$

$= \boldsymbol{y}^t\boldsymbol{y} - 2\boldsymbol{\beta}^t\boldsymbol{X}^t\boldsymbol{y} + \boldsymbol{\beta}^t(\boldsymbol{X}^t\boldsymbol{X})\boldsymbol{\beta}$

Partial differential by parameters:

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^t\boldsymbol{y} + 2\,\boldsymbol{X}^t\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$$

Parameters to be estimated: $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{y}$

Data is statistics of Japanese freight generation weight by 47 prefectures, goods type and four years (2000, 2005, 2010, 2015). This data was downloaded “Commodity Flow Survey” homepage.
http://www.mlit.go.jp/sogoseisaku/transport/sosei_transport_fr_000074.html

“freight4y.csv” → table format, this PPT uses this.
“freight4yMELT.csv” → Melt format, sometimes easy to analyze

Data structure:

year	num	pref	pop.1000	GRP.mill	agriculture	wood	mine	machine	chemical	Smachine	miscind	special	total
2000	1	Hokkaido	5683	20471299	13062354	2100076	47400509	8638957	68047304	13457707	3209660	8445845	164362412
2000	2	Aomori	1476	4622235	3558257	481593	17733394	2025278	10073313	2344651	687354	2391114	39294954
2000	3	Iwate	1416	4893514	1725137	906590	15373502	2207847	14041095	2268808	2031276	690371	39244626
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2015	45	Miyazaki	1104	3643441	1500698	657421	3525008	597605	5702006	1702632	947604	2383239	17016213
2015	46	Kagoshima	1648	5330338	6676265	247527	9668769	781037	4873191	2611927	1684215	6186117	32729048
2015	47	Okinawa	1434	4051060	943789	14256	11519952	708861	6148572	1052887	131386	698150	21217853

Foreign students → please check the location of 47 prefectures

Item examples:

1. Agriculture: wheat, rice, fruits, vegetables, fish etc.
2. Wood: raw wood, lumber, firewood etc.
3. Mine: coal, iron ore, gravel, limestone, crude oil etc.
4. Machine: steel, metal products, industrial machinery, car, precision machine etc.
5. Chemical: cement, glass, china, heavy oil, LNG, chemical fertilizer etc.
6. Smachine: pulp, fabric, sugar, beverage etc.
7. Miscind: book, clothing, furniture, wood products, toy etc.
8. Special: container, drum, cardboard box, grass, barrel etc.

Unit:

Population [1000]
GRP [million yen]
Generation volume [ton/year]

Number of rows=4*47=188

1) Read csv file and make a graph of four years total weight.

```
setwd("c:/usr/")
dt <- read.csv("freight4y.csv",header=T)
z <- array(0,c(4,8)); y <- c(2000,2005,2010,2015)
for(i in 1:4){
  for(j in 1:8){
    z[i,j] <- sum( as.numeric( dt[(dt[,1]==y[i]),j+5] ) )
}}
rownames(z) <- y; colnames(z) <- colnames(dt)[6:13]
barplot(t(z),legend=T)
```

2) Check the basic statistics: Scatter plot between any variables:

```
par(ask=T)
y2015 <- dt[dt[,1]==2015,]
plot(log10(y2015[,4]),log10(y2015[,5]),type="n")
text(log10(y2015[,4]),log10(y2015[,5]),dt[1:47,3])
for(i in 1:4){
  a <- dt[dt[,1]==y[i],]
  plot(log10(a[,4]),log10(a[,14]),type="n")
  text(log10(a[,4]),log10(a[,14]),dt[1:47,3])
}
```

3) Try Multiple Regression model with population & economic data (GRP)

```
res.1 <- lm( log(dt[,6]) ~ log(dt[,4])+log(dt[,5]) )
res.2 <- lm( log(dt[,7]) ~ log(dt[,4])+log(dt[,5]) )
summary(res.1); summary(res.2)
```

This equations mean that the models are estimated by goods.

How about the sign condition ?
How about the fitness ?
Dummy variables may improve fitness (Tokyo dummy, 2010 Lehman dummy…)

Using all data simultaneously, but parameters are different by goods:

```
all <- array(0,c(47*4*8,17))
ns <- 0
for(i in 1:8){
  for(j in 1:(47*4)){ ns <- ns+1
    all[ns,1] <- log(dt[j,i+5])
    all[ns,i*2  ] <- log(dt[j,4])
    all[ns,i*2+1] <- log(dt[j,5])
}}
all <- data.frame(all)
res.all <- lm( X1 ~ . , data=all)
summary(res.all)

all <- array(0,c(47*4*8,17))
ns <- 0
for(i in 1:8){
  for(j in 1:(47*4)){ ns <- ns+1
    all[ns,1] <- log(dt[j,i+5])
    all[ns,i*2  ] <- log(dt[j,4])
    all[ns,i*2+1] <- log(dt[j,5]/dt[j,4])
}}
all <- data.frame(all)
res.al2 <- lm( X1 ~ . , data=all)
summary(res.al2)
```

4) Check the population data by National Institute of Population and Social Security Research

- Future population data is required for forecasting
- We forecast 2025, 2035, 2045 freight volume
- In 2015 outputs, observed and calibrated value should be same
  → Launch point adjustment
- Future GRPs are calibrated by constant growth rate assumptions
  → 0%/year, 0.5%/year, 1%/year

```
pop <- read.csv(file="futPOP2045.csv",header=T,row.names=1)
t <- apply(pop,2,sum)
barplot(t)
```

5) Forecast future freight volume in 2025, 2035, 2045 (National total volume)

- How to forecast decreasing demand ?
- Decreasing/increasing rate by prefectures should be illustrated

PLEASE make presentation Power Point file including estimated parameters and final forecasted volume

Presentation date by students: