제 6 장 Transformation of Variables

▷ **<u>Objectives</u>** : to ensure **linearity**, to achieve **normality**, to **stabilize the variance**

We will illustrate using simple regression, multiple regression requires more effort and care

▷ <u>Linearity</u> : linear model : parameters occur linearly

Examples : p.164

linear model : $Y = \beta_0 + \beta_1 \log X + \epsilon,$

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon,$$

nonlinear model : $Y = \beta_0 + e^{\beta_1 X} + \epsilon$

▷ Transformations may be necessary for several reasons :

1. Theoretical consideration :

   eg. T_i : the time taken to perform a task on the i-th occasion

$$T_i \approx \alpha\beta^i, \qquad \alpha > 0, \quad 0 < \beta < 1 \qquad (\textit{decrease exponentially})$$

   $\Rightarrow$ Log transformation : $\log T_i = \log\alpha + i\log\beta + \epsilon$: linear model

   Transformations to achieve linearity : See table 6.1 on p.165, Fig 6.1 - Fig 6.4

2. Probability distribution of Y is not normal, or $Var(Y)$ depends on X :

   non-normality : invalidate the standard test of significance (*does not cause major problem with large sample*)

   unequal variance : estimators are unbiased but no longer best... estimators will have large se ...

   $\Rightarrow$ Need variance stabilizing transformations : are also good normalizing transformation.

3. neither prior theoretical nor probabilistic reasons to suspect,

   evidence comes from examining the residuals.

◎ **Example** : bacteria deaths due to X-ray radiation

data : table 6.2 on p.168 : 200 k-volt X-rays for period $t = 1, 2, \ldots, 15$

n_t : no. of surviving bacteria(in thousands) after exposure time t.

Theory : $n_t = n_0 e^{\beta_1 t}$, $t \geq 0$, n_0 : no. of bacteria at the start of experiment

β_1 : decay rate

Scatter plot : n_t $vs.$ t (Fig 6.5 on p.169) : non-linearity

model ① : $n_t = \beta_0 + \beta_1 t + \epsilon_t$

Result : Table 6.3 on p.169 $R^2 = 0.823$

Residual plot (Fig 6.6 on p.169) shows non-linearity

⇒ log transformation : $\log n_t = \log n_0 + \beta_1 t = \beta_0 + \beta_1 t$

additive error term : $\log n_t = \beta_0 + \beta_1 t + \epsilon_t$

($n_t = n_0 e^{\beta_1 t} \epsilon'_t$, $\epsilon_t = \log \epsilon'_t$, *multiplicative in the original model*)

Scatter plot : $\log n_t$ $vs.$ t (Fig 6.7 on p.170) : linear model ② : $\log n_t = \beta_0 + \beta_1 t + \epsilon_t$

Result : Table 6.4 on p.170 $R^2 = 0.988$ (original scale : $R^2 = 0.9689$)

$\widehat{n_t} = \exp(\widehat{\log n_t})$: biased estimator!!

Residual plot(Fig 6.8 on p.171) : ideal !

Implementation :

$\widehat{\beta_1} = -0.218$, $se(\widehat{\beta_1}) = 0.0066$, 95% CI : (-0.232, -0.204)

$\widehat{\beta_0} = 5.973$, $\widehat{n_0} = e^{\widehat{\beta_0}} = 392.68$

<<< *이하의 내용은 어려우므로 상황 봐서...* >>>

$E(\widehat{n_0}) \geq e^{E(\widehat{\beta_0})} = e^{\beta_0} = n_0$: biased estimator (*Jensen's inequality*)

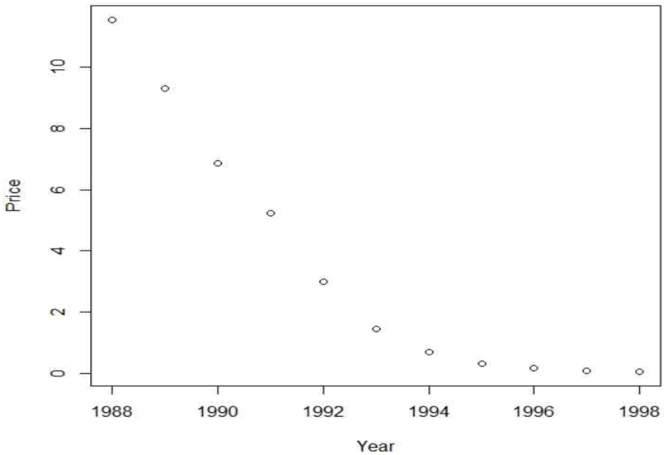
$\widehat{n_0}^* = \exp\left(\widehat{\beta_0} - \frac{1}{2} Var(\widehat{\beta_0})\right) = 381.11$: nearly unbiased

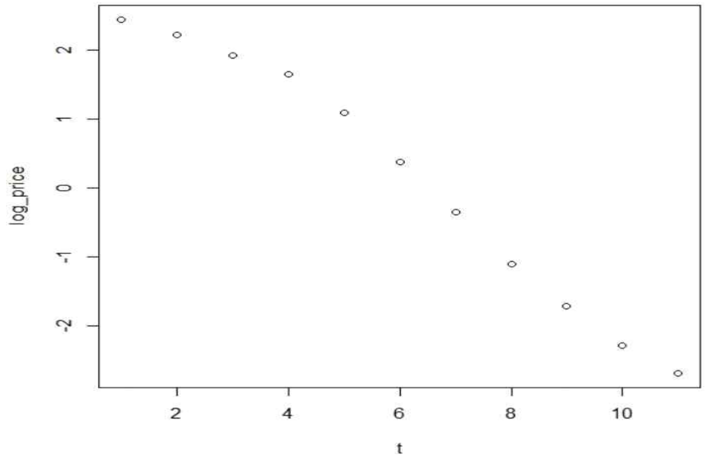
(*no theoretical statements...*)

Note : *the bias in estimating* n_0 *has no effect on the test of the theory or the estimation of the decay rate.*

$P(\log X \leq \log x) = P(X \leq x) \ldots$

◎ **Example** : #6.8 : The cost of storage : average price per megabyte in dollars 1988-1998





model ① : $\log(P_t) = \beta_0 + \beta_1 t + \epsilon$ $R^2 = 0.9791$

1988~1991과 1992~1998의 기울기가 다르게 보인다.

model ② : $\log(P_t) = \beta_0 + \beta_1 t + \beta_2 z + \beta_3 zt + \epsilon,$ $R^2 = 0.9972$ ($z=1$ if year 1988~1991, 0 if year 1992_1998)

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.21053	0.18380	22.908	7.66e-08	***
t	-0.64548	0.02229	-28.959	1.51e-08	***
z	-1.47691	0.23377	-6.318	0.000397	***
tz	0.37763	0.05726	6.595	0.000306	***

1988~1991 : 메가바이트당 가격이 매년 $e^{-0.26785}$배로 감소하고,

1992~1998 : 메가바이트당 가격이 매년 $e^{-0.64548}$배로 감소한다.

▷ Transformations to stabilize variance

equal variance : homoscedasticity,

unequal variance : heteroscedasticity.

eg. Fig 6.9 on p. 172

Probability distribution of Y : $Var(Y)$ depend on $E(Y) = \beta_0 + \beta_1 X$

⇒ Transformations to stabilize variance (*and normalizing*) : See Table 6.5 on p. 173

◎ **Example** : Injury incidents in Airlines

Data : Table 6.6 on p.174,

N : proportion of total flights from NY among 9 major US airlines for a single year,

Y : no of injury incidents

$$n_i = \frac{f_i}{\sum f_i}, \qquad f_i \text{ : no of total flights of the } i\text{-th airlines}$$

Scatter plot : y_i $vs.$ n_i Fig 6.10 on p.174

If all the airlines are equally safe,

model ① : $y_i = \beta_0 + \beta_1 n_i + \epsilon_i$,

Result : Table 6.7, $R^2 = 0.4872$

residual plot : Fig 6.11 : *increasing variance* (*Poisson distribution*)

⇒ transformation as suggested in Table 6.5

model ② : $\sqrt{y_i} = \beta'_0 + \beta'_1 n_i + \epsilon_i,$

Result : Table 6.8, $R^2 = 0.483$ (original scale $R^2 = 0.4893$)

residual plot : Fig 6.12 : ideal

Only 48% of the total variability of the injury incidents is explained by the variation in the no of flights.
It appears that for a better explanation of injury incidents other factors have to be considered.

◎ **Example** : Industrial data (Detection of heteroscedastic errors)

data : Table 6.9 on p.176,

Scatter plot : $y_i \ vs. \ x_i$ (Fig 6.13 on p. 177) : *increasing variance*

X : no. of supervised workers

Y : no. of supervisors

model ① : $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$,

Result : Table 6.10, $R^2 = 0.776$

residual plot : Fig 6.14 on p. 177 : *increasing variance*

▷ Removal of heteroscedasticity

$Var(\epsilon_i) = k^2 x_i^2, \quad k > 0$

$$\Rightarrow \quad \frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + \frac{\epsilon_i}{x_i}, \qquad y_i{}' = \frac{y_i}{x_i}, \quad x_i{}' = \frac{1}{x_i}, \quad \beta_0{}' = \beta_1, \quad \beta_1{}' = \beta_0, \quad \epsilon_i{}' = \frac{\epsilon_i}{x_i}$$

$\Rightarrow$ model ② : $y_i{}' = \beta_0{}' + \beta_1{}' x_i{}' + \epsilon_i{}'$,

Result : Table 6.11, ($y_i{}'$에 대한 $R^2 = 0.027$ .. 이유 설명)

residual plot : Fig 6.15 : ideal

$$\widehat{\left(\frac{y_i}{x_i}\right)} = 0.121 + 3.803\left(\frac{1}{x_i}\right) \qquad \Rightarrow \qquad \hat{y_i} = \hat{y_i}{}' x_i = \hat{\beta}_1{}' + \hat{\beta}_0{}' x_i = 3.803 + 0.121\, x_i$$

model ① : $se(\widehat{\beta_1}) = 0.011$

model ② : $se(\widehat{\beta_1}) = se(\widehat{\beta_0'}) = 0.009$

$\Rightarrow$ 33% reduced

model ① : $R^2 = 0.776, \quad \hat{\sigma} = 21.73$,

model ② : $R^2 = 0.758, \quad \hat{\sigma} = 22.577$

▷ Weighted least squares

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad Var(\epsilon_i) = \sigma_i^2$

$\Rightarrow \quad \frac{y_i}{\sigma_i} = \frac{\beta_0}{\sigma_i} + \frac{\beta_1}{\sigma_i} + \frac{\epsilon_i}{\sigma_i}, \qquad (\sigma_i^2 = k^2 x_i^2$: special case)

$\Rightarrow$ *more in detail in Ch 7.*

▷ Logarithmic transformation of data

reduce asymmetry and remove heteroscedasticity.

◎ **Example** : Industrial data

Use logarithmic transformation to remove heteroscedasticity.

model ③ : $\log\ y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Scatter plot : $\log\ y_i \quad vs. \quad x_i$ Fig 6.16 on p.180

Result : Talbe 6.12, $R^2 = 0.77$ (original scale $R^2 = 0.574$)

residual plot : Fig 6.17 : non-linearity

⇒ add the square term

model ④ : $\log\ y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$

Result : Talbe 6.13, $R^2 = 0.886$ (original scale $R^2 = 0.7967$)

residual plot : Fig 6.18-20 :

note : *two acceptable models : model ② and ④*

model ② : $R^2 = 0.758$, *easier to interpret*

model ④ : $R^2 = 0.886$, (original y에 대한 값 : $R^2 = 0.7967$)

▷ <u>Power Transformation</u>

Y^{λ} $\quad \lambda=-1$: reciprocal $\dfrac{1}{Y}$

$\lambda=0.5$: square root $\sqrt{Y}$

$\lambda=0$: logarithmic $\log Y$

cf : Box-Cox transformation : $\dfrac{Y^{\lambda}-1}{\lambda}$ ($\rightarrow \log Y$, as $\lambda \rightarrow 0$)

If λ cannot be determined by theoretical considerations,

$\Rightarrow$ try $\lambda=2,\ 1.5,\ 1,\ 0.5,\ 0,\ -0.5,\ -1,\ 1.5,\ -2$ (a ladder of transformation)

$\Rightarrow$ choose the best value

◎ **Example** : The brain data

Y : average brain weight

X : average body weight

Scatter plot : Fig 6.21 on p.184 : need transformation!

See Fig 6.22 on p.185 : $\log Y$, $\log X$: appropriate, 3 outliers

◎ #6.1, #6.2 읽어볼 것