고급회귀분석론

Ch5. Transformations and weighting to correct model inadquacy

양성준

데이터 변환

- 모형에 대한 가정이 위배되는 경우, 변수에 적절한 변환을 취한 후 선형모형을 적합하는 것이 하나의 해결책이 될 수 있다.
- 데이터 변환은 변수들간의 함수관계를 바꿀 뿐 아니라, 분포 및 분산 등에 모두 영향을 미치게 된다.
- 보통 반응변수의 변환에 일차적으로 초점을 맞추지만 예측변수들의 변환도 필요할 수 있다.
- 어떤 변환을 취할 것인지는, 해당 domain의 지식을 이용하여 주관적으로 결정하는 것이 좋을 수 있다. 하지만, 이런 사전지식이 없는 경우 등에는 기술적으로 변환함수를 결정하게 된다.

분산안정화변환

- 오차항의 등분산 가정이 위배되는 경우를 가정
- 등분산 가정의 위배는 반응변수가 분산이 평균과 연결성을 가지는 분포를 따르기 때문에 발생하는 경우가 종종 있다.
- 이 경우 분산과 평균의 함수관계에 따라 변환함수를 적절히 결정하면 등분산 가정을 만족하도록 할 수 있다.

분산안정화변환

- 이론적 근거 : Delta method

$$Y' = g(Y) \implies Var(Y') \approx g'(\mu)^2 Var(Y)$$

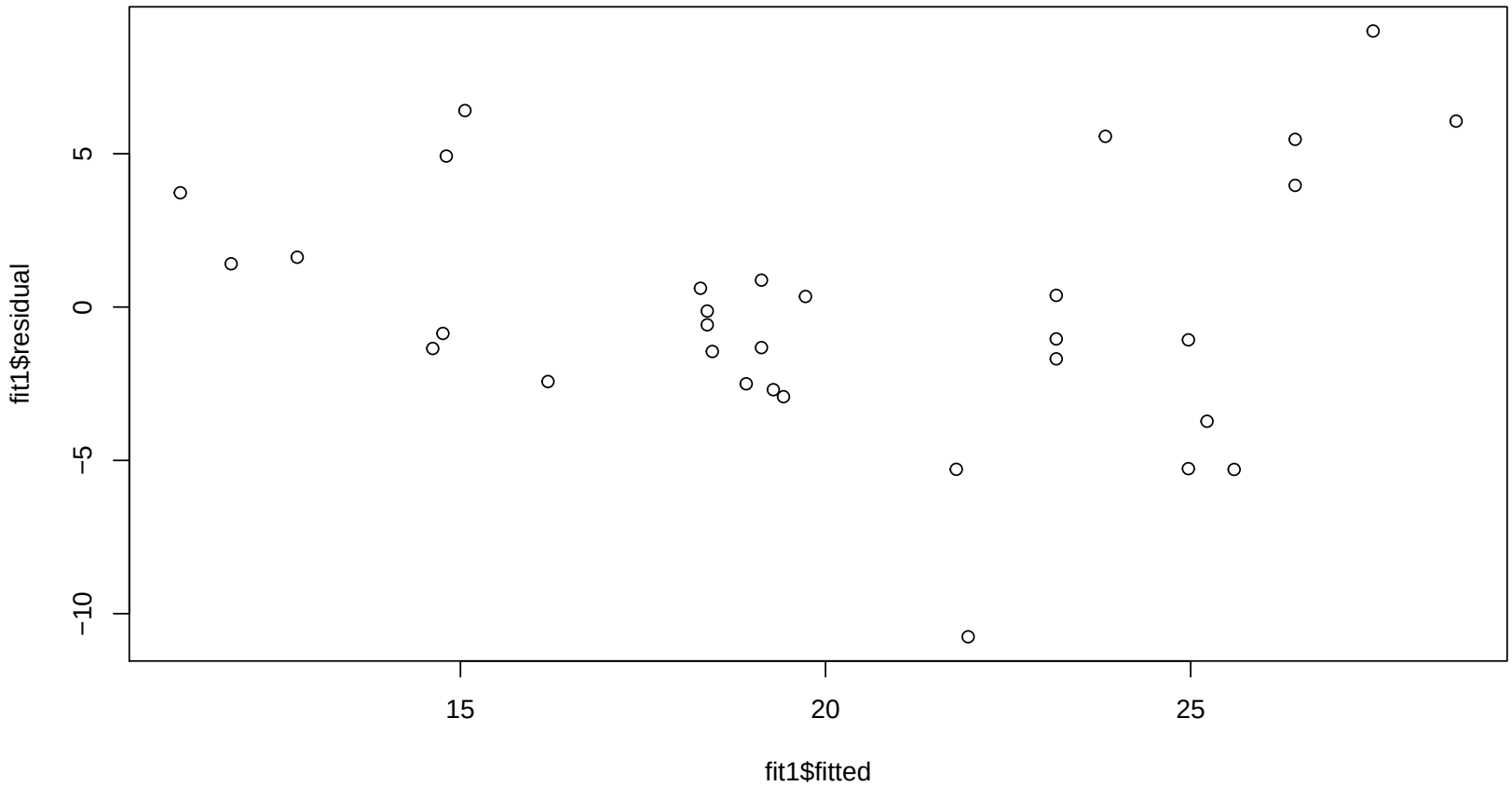
- Ex] $Y \sim Poission(\lambda) \rightarrow E(Y) = var(Y)$. 즉, $Y' = \sqrt{Y}$로 변환하면, $Var(Y') \approx 1/4$.

분산안정화변환

- Ex] $\sigma^2 = var(Y) \propto E(Y)$ 이면 $Y' = \sqrt{Y}$로 변환.
- Ex] $\sigma^2 = var(Y) \propto E(Y)^2$ 이면 $Y' = \log Y$로 변환.
- Ex] $\sigma^2 = var(Y) \propto E(Y)^3$ 이면 $Y' = Y^{-1/2}$로 변환.
- Ex] $\sigma^2 = var(Y) \propto E(Y)^4$ 이면 $Y' = Y^{-1}$로 변환.
- Ex] $\sigma^2 = var(Y) \propto E(Y)(1 - E(Y))$ 이면 $Y' = arcsin(\sqrt{Y})$로 변환.

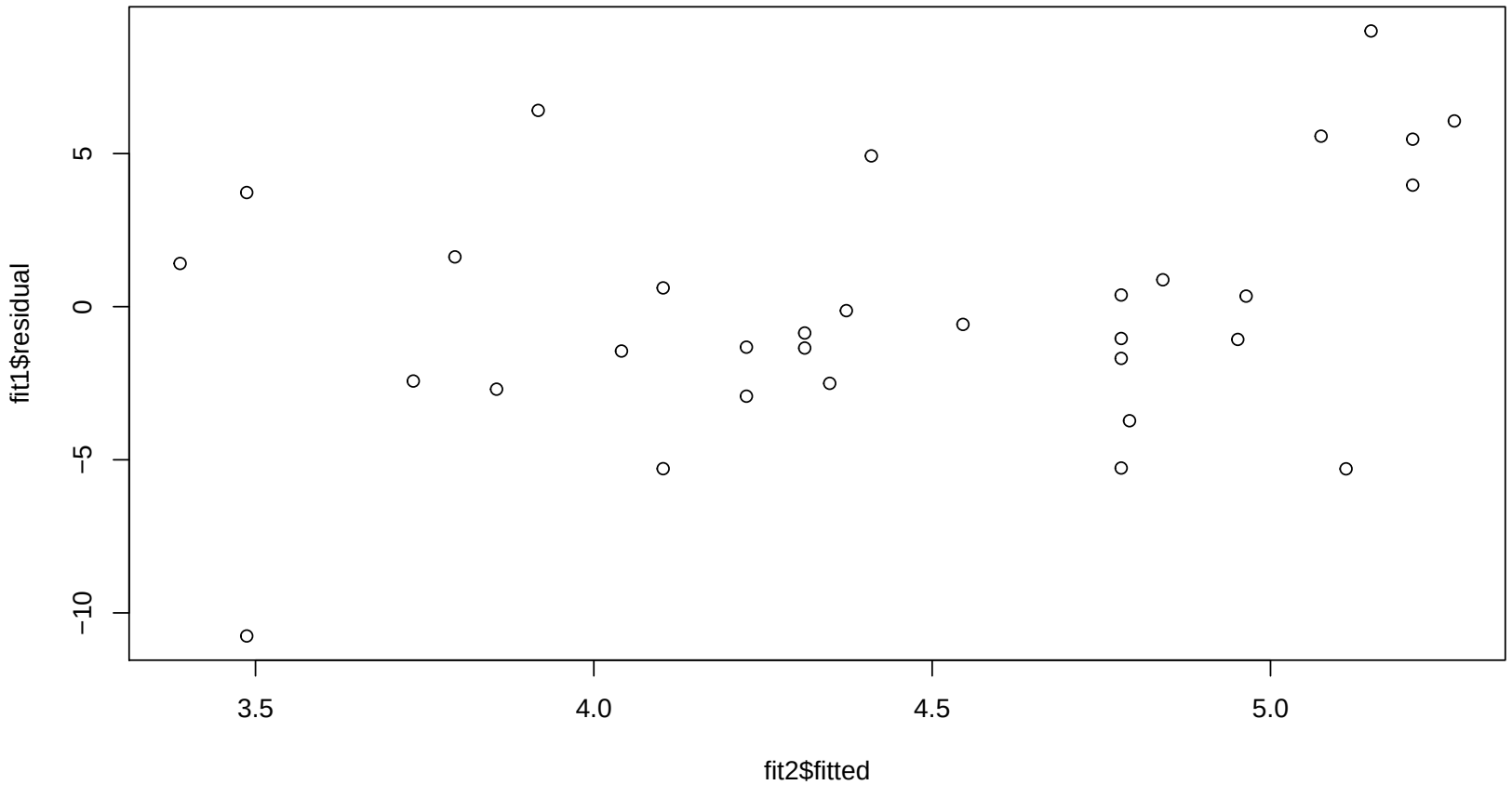
Example : mileage data (Table B.3)

```
library(MPV)
fit1 = lm(y~x8,data=table.b3) # mile/gallon ~ length
plot(fit1$fitted,fit1$residual)
```



Example : mileage data (Table B.3)

```
table.b3$y2 =(table.b3$y)^{1/2} # sqrt transformation
fit2 = lm(y2~x2,data=table.b3)
plot(fit2$fitted,fit1$residual)
```



Example : mileage data (Table B.3)

```r
summary(fit1)$r.squared
```

```
## [1] 0.5694537
```

```r
summary(fit2)$r.squared
```

```
## [1] 0.6849133
```

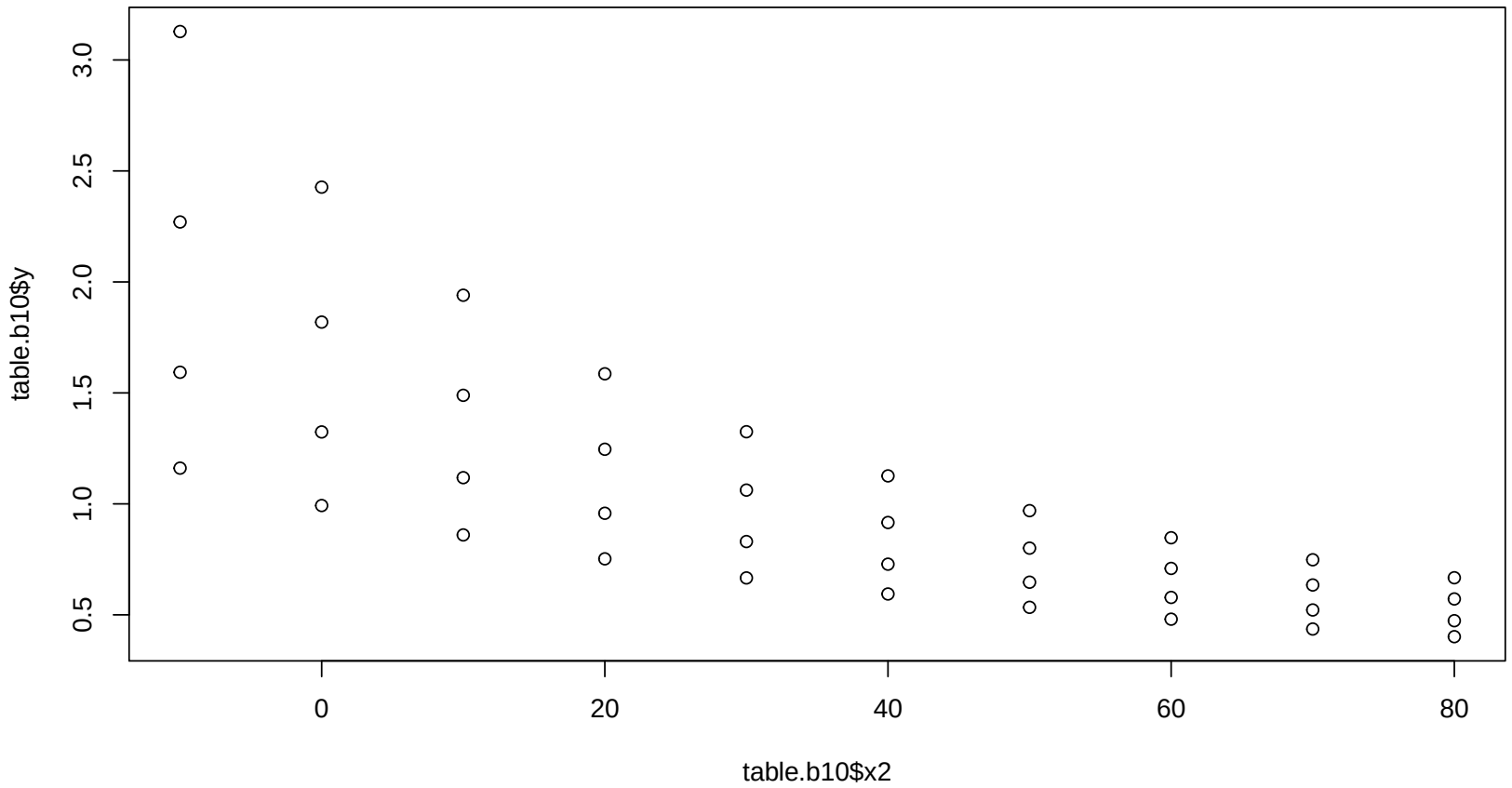
선형화변환

- 변수들간의 함수관계가 비선형일 때 적절한 변환을 통해 선형모형화 할 수 있다.
- 보통 특정 domain에서 축적된 지식을 통해 변환함수가 결정되는 경우가 많다.
- Ex] $y = \beta_0 e^{\beta_1 x} \epsilon$은 로그변환을 통하여

$$\log(y) = \log(\beta_0) + \beta_1 x + \log(\epsilon)$$

Example : Kinematic viscosity data (Table B.10)

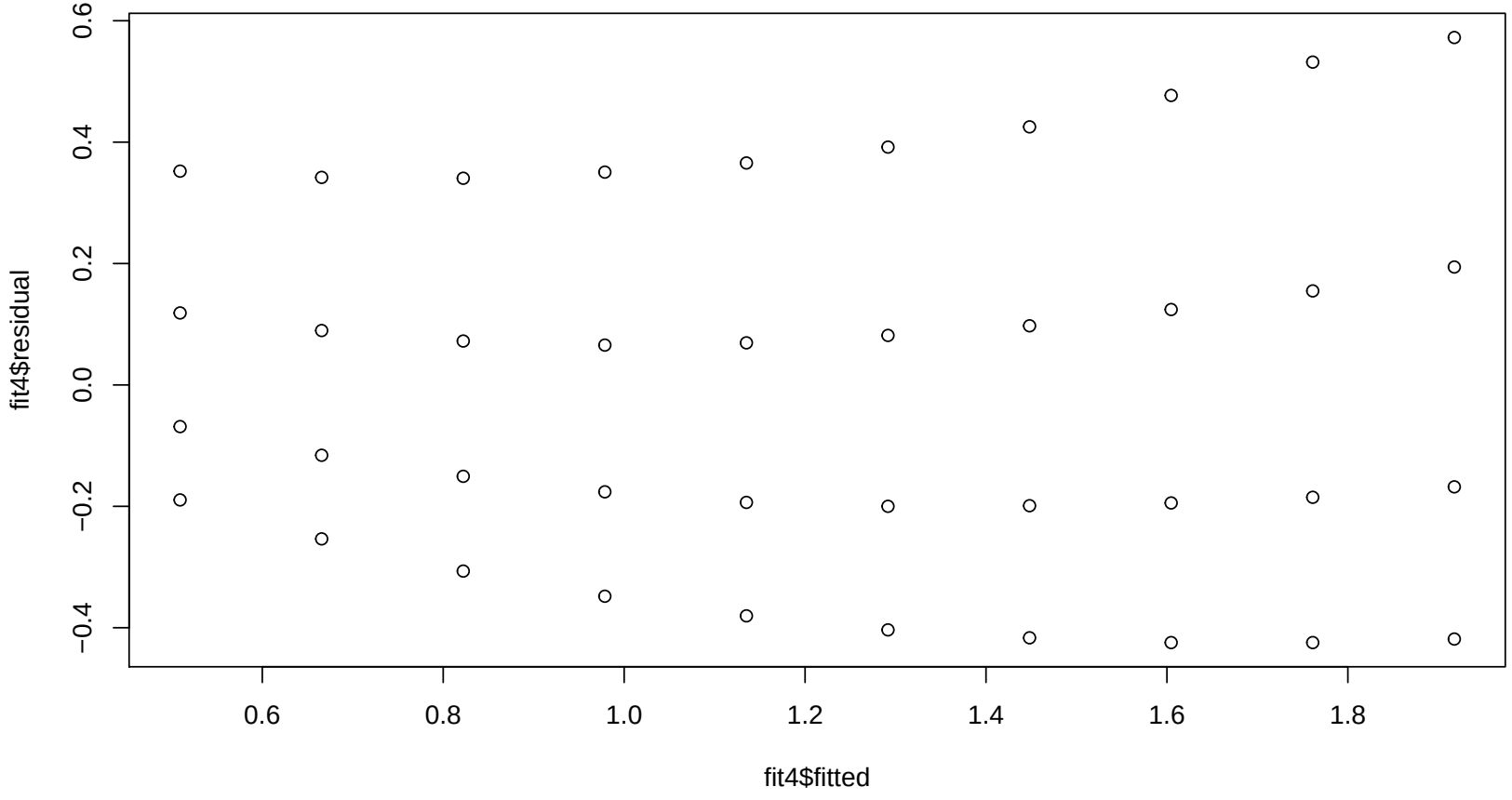
```
plot(table.b10$x2,table.b10$y) # linear?
```




```
fit3 = lm(y~x2,data=table.b10)
```

Example : Kinematic viscosity data (Table B.10)

```
table.b10$y2 =(table.b10$y)^{-1}
fit4 = lm(y2~x2,data=table.b10)
plot(fit4$fitted,fit4$residual)
```



Example：Kinematic viscosity data (Table B.10)

```
summary(fit3)$r.squared
```

```
## [1] 0.5764172
```

```
summary(fit4)$r.squared
```

```
## [1] 0.6956141
```

Box-Cox 변환

- 변환은 경험적으로 혹은 데이터의 형태를 보고 결정할 수 있다. 하지만, 좀 더 객관적인 변환을 찾는 방법이 필요할 수 있다.
- 반응변수의 이분산문제/비정규성 문제 등의 해결을 위해 종종 사용되는 다음과 같은 변환(power transform)에서 지수 λ를 결정해 주는 방법

$$y' = \frac{y^\lambda - 1}{\lambda}$$

- 단, $\lambda = 0$이면 변환은 $y' = \log(y)$로 정의한다.

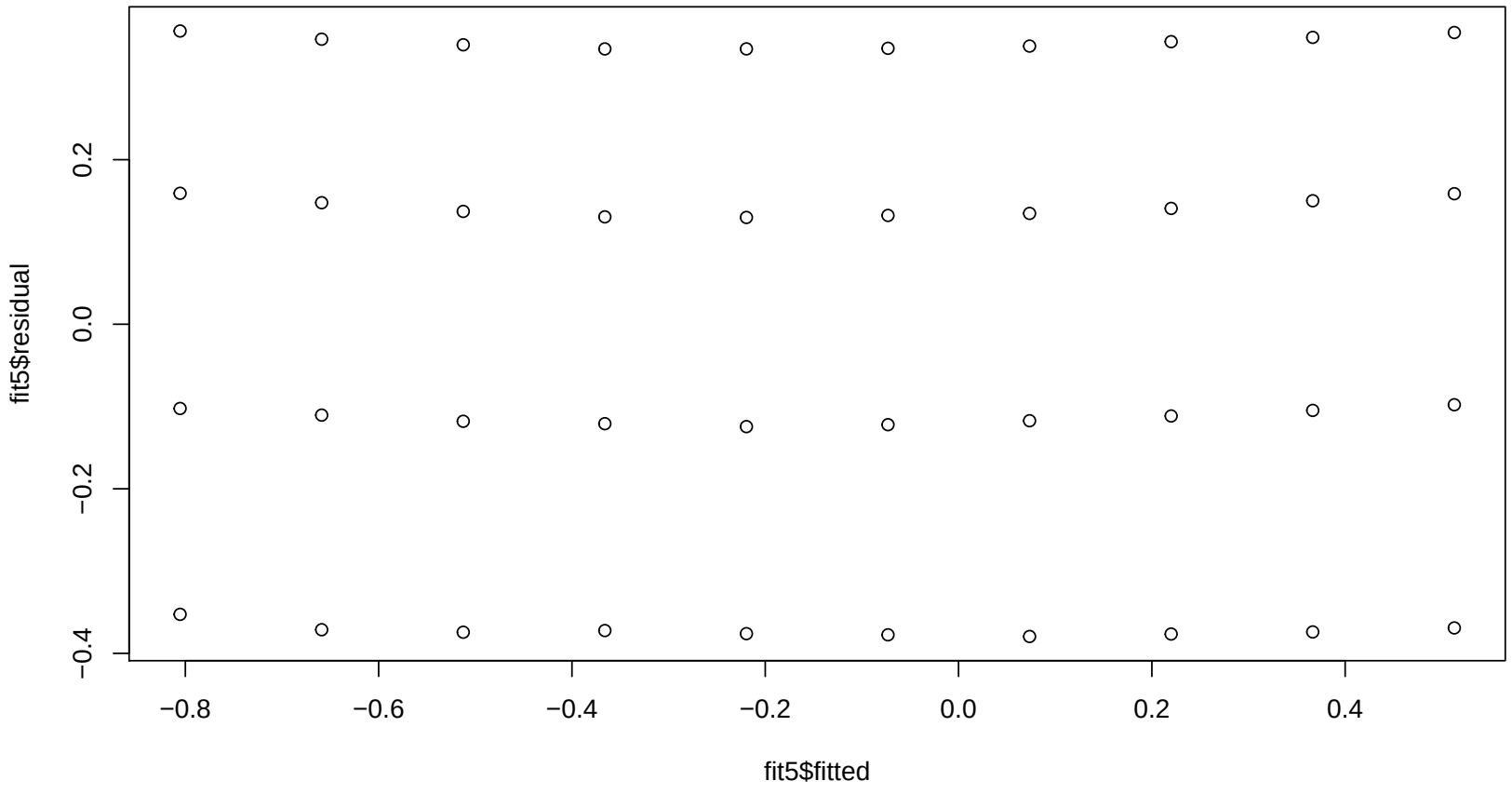
Example：Kinematic viscosity data (Table B.10)

```
library(EnvStats)
(l1=boxcox(table.b10$y,optimize=T))
```

```
##
## Results of Box-Cox Transformation
## ---------------------------------
##
## Objective Name:                  PPCC
##
## Data:                            table.b10$y
##
## Sample Size:                     40
##
## Bounds for Optimization:         lower = -2
##                                  upper =  2
##
## Optimal Value:                   lambda = -0.503665
##
```

Example : Kinematic viscosity data (Table B.10)

```
table.b10$y3 = (table.b10$y^l1$lambda - 1)/l1$lambda
fit5 = lm(y3~x2, data=table.b10)
plot(fit5$fitted,fit5$residual)
```



Example : Kinematic viscosity data (Table B.10)

```
(l2=boxcox(fit3,optimize=T))
```

```
##
## Results of Box-Cox Transformation
## ---------------------------------
##
## Objective Name:                  PPCC
##
## Linear Model:                    fit3
##
## Sample Size:                     40
##
## Bounds for Optimization:         lower = -2
##                                  upper =  2
##
## Optimal Value:                   lambda = 0.264594
##
## Value of Objective:              PPCC = 0.9879378
```

Example：Kinematic viscosity data (Table B.10)

```
summary(fit3)$r.squared
```

```
## [1] 0.5764172
```

```
summary(fit4)$r.squared
```

```
## [1] 0.6956141
```

```
summary(fit5)$r.squared
```

```
## [1] 0.7096884
```

```
summary(fit6)$r.squared
```

```
## [1] 0.6734184
```

Some issues

- Box-Cox 변환은 본질적으로 변환된 변수가 정규분포를 따르는 것을 전제함.
- `boxcox`함수에 numeric vector가 들어가는 경우 해당 변수의 정규성을, `lm` object가 들어가는 경우 적합된 선형모형에서 추출된 잔차벡터의 정규성을 바탕으로 λ를 결정한다.
- 정규성/등분산성 중 어떤 부분에 문제가 있느냐에 따라 반응변수/예측변수 어느 쪽에나 변환을 취할 수 있다. 정규성/등분산성에 문제가 없는 경우 반응변수의 변환은 일차적 고려대상이 아니다.
- 보통 반응변수의 변환을 일차적으로 고려하는 경우가 많다. 보통 반응변수는 1개, 예측변수는 다수이기 때문이다. 예측변수의 경우 지나치게 skewed된 분포를 가지는 것을 해소하기 위해 로그- 혹은 루트-변환을 실시하거나, 선형성을 벗어나는 것을 해소하기 위해 적절한 변환을 실시하는 경우가 많다. 하지만, 선형성을 벗어나는 경우는 예측변수의 변환에 의존하는 것 말고도 다항회귀, 비선형 회귀 등을 고려함으로써 해결할 수도 있다.

가중최소제곱추정량 (weighted least squares)

- 오차항의 등분산성이 만족되지 않을 때, 변환이 아닌 추정 단계에서 문제를 해결.
- 다음과 같은 가중최소제곱추정량이 최소제곱추정량보다 좋은 성질을 가진다.

$$\hat{\beta} = (X'WX)^{-1}X'Wy$$

여기서

$$W = diag(w_i),\ var(y_i) = \sigma^2/w_i$$

- $W^{1/2}X$, $W^{1/2}y$로 각각 데이터를 변환한 후 최소제곱추정량을 고려한 것으로 볼 수 있다.

일반화최소제곱추정량 (generalized least squares)

- 일반적으로 다음과 같은 모형에서

$$y = X\beta + \epsilon,\ E(\epsilon) = 0,\ var(\epsilon) = \sigma^2 V$$

일반화최소제곱추정량이 최소제곱추정량보다 좋은 성질을 가진다.

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y$$

- V가 대각 행렬인 경우, 즉 오차항의 독립성은 만족되는 경우 $V = W^{-1}$로 생각할 수 있다. 즉, 가중최소제곱추정량은 일반화최소제곱추정량의 특수한 경우이다. 만약 V가 대각행렬인 동시에 모든 대각원소들의 값이 같다면 최소제곱추정량과, 가중-일반화최소제곱추정량은 모두 일치하게 된다.

Practical issue

- 위 추정량들은 회귀계수의 추정을 위해 분산행렬에 대한 지식이 필요하다. 이는 사전 경험이나 이론모형으로부터 결정이 가능한 경우도 있으나 그러지 않은 경우에는 세심한 고려와 결정이 필요하며 쉬운 문제가 아니다.
- 어떤 경우에는 y_i가 특정 x_i에서의 n_i 관측치의 평균으로 주어지기도 하는데 만약 각 관측치의 분산이 같다면 $var(y_i) \propto 1/n_i$이 되어 $w_i = n_i$로 잡을 수 있다.

Example：clathrate formation data (Table B.8)

```
w1=1/var(table.b8$y[table.b8$x1==0])
w2=1/var(table.b8$y[table.b8$x1==0.02])
w3=1/var(table.b8$y[table.b8$x1==0.05])
w=w1*(table.b8$x1==0)+w2*(table.b8$x1==0.02)+w3*(table.b8$x
fit8=lm(y~x1,data=table.b8, weight=w)
fit7=lm(y~x1,data=table.b8)
```

Example：clathrate formation data (Table B.8)

```
par(mfrow=c(1,2))
plot(fit7$fitted.values,fit8$residuals)
plot(fit8$fitted.values*sqrt(w),fit8$residuals*sqrt(w))
```