Lesson8

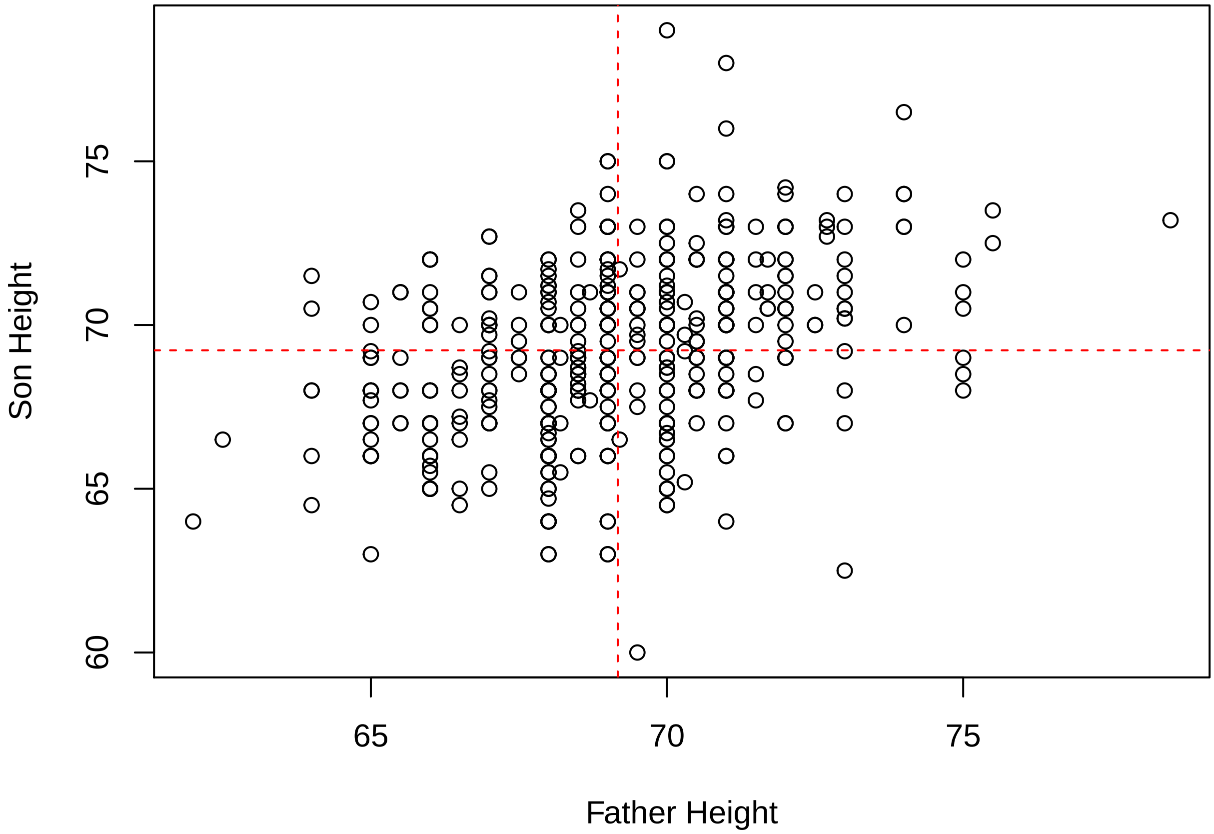
Sung Won Kang

2016년 12월 일

Data 준비

```
#hf=read.csv("http://www.math.uah.edu/stat/data/Galton.csv",header=T,stringsAsFactors = FALSE)
#save(hf,file="Fatherson_o.rdata")
#load("Fatherson_o.Rdata")
#str(hf)
#str(hf$Gender)
#hf$Gender=factor(hf$Gender,levels=c("M","F"))
#str(hf$Gender)
#str(hf)

#hf.son=subset(hf,Gender=="M")
#str(hf.son)
#hf.son=hf.son[c("Father","Height")]
#str(hf.son)
#save(hf.son, file="Fatherson.Rdata")
load("Fatherson.rdata")
par(mar=c(4,4,1,1))
plot(hf.son$Father,hf.son$Height,xlab="Father Height",ylab="Son Height",mani="FS height")
abline(v=mean(hf.son$Father),col=2,lty=2)
abline(h=mean(hf.son$Height),col=2,lty=2)
```



상관계수

1. 공분산, 표본공분산

$$\sigma_{xy} = E[(x-\mu_x)(y-\mu_y)], \qquad S_{xy} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{N-1}$$

2. 상관계수, 표본상관계수

$$\rho_{xy} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sqrt{E(X-\mu_x)^2}\sqrt{E(X-\mu_y)^2}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$r_{xy} = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X_i-\bar{X})^2}\sqrt{\Sigma(Y_i-\bar{Y})^2}} = \frac{S_{xy}}{S_x S_y} S_x^2 = \frac{\Sigma(X_i-\bar{X})^2}{N-1} \qquad S_y^2 = \frac{\Sigma(Y_i-\bar{Y})^2}{N-1}$$

$$|\rho_{xy}| < 1, |S_{xy}| < 1$$

-1, 1에 가까우면 강한 상관관계, 0에 가까우면 약한 상관관계.

3. 복수의 확률변수 간의 상관계수: 분산-공분산 행렬

$$X = \{x_1, x_2, \cdots x_k\}$$

$$E(X) = E\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \mu$$

$$Cov(X) = E((X-\mu)(X-\mu)^T] = E\begin{bmatrix} (X_1-\mu_1)(x_1-\mu_1) & (x_1-\mu_1)(x_2-\mu_2) & \cdots & (x_1-\mu_1)(x_n-\mu_n) \\ (X_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)(x_2-\mu_2) & \cdots & (x_2-\mu_1)(x_n-\mu_n) \\ & & \vdots & \\ (X_n-\mu_n)(x_1-\mu_1) & (x_n-\mu_n)(x_2-\mu_2) & \cdots & (x_n-\mu_n)(x_n-\mu_n) \end{bmatrix}$$

행렬 A에 확률변수 벡터 X를 곱하면 또 다른 학률변수 AX 가 되고, 그 분산 공분산 행렬은 다음과 같다.

$$E(AX) = AE(X) = A\mu$$

$$Cov(AX) = E[(AX-A\mu)(AX-A\mu)^T] = E[A(X-\mu)(X-\mu)^T A^T] = AE[(X-\mu)(X-\mu)^T]A^T = ACov(X)A^T$$

rcorr 함수의 p value는 $H_0 : r = 0$ 에 대한 t-test의 p-value 인데 참고로만 쓰인다. (가정이 너무 제한적)

corrgram package의 corrgram 함수는 여러 변수의 상관계수의 방향 및 값에 대한 시각 정보를 제공해 준다. 빨간색은 음, 파란색은 양의 상관관계를 나타내며, 색이 짙으면 강한 상관관계를 나타낸다.

회귀분석

추정식

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, 2, \cdots, n \quad \epsilon_i \sim N(0, \sigma^2)$$

$$Y_{(n\times1)} = [1, X]_{(n\times2)}\beta_{(2\times1)} + \epsilon_{(n\times1)} \quad \epsilon \sim N(0, \sigma^2 I_{(n\times n)})$$

$$Y_{(n\times1)} = X_{(n\times k)}\beta_{(k\times1)} + \epsilon_{(n\times1)} \quad \epsilon \sim N(0, \sigma^2 I_{(n\times n)})$$

1. Y : 종속변수
2. X : 독립변수
3. ϵ : 오차
4. β_k : 회귀계수

$$\beta_k = \frac{\Delta Y}{\Delta X_k} = \frac{\partial Y}{\partial X}$$

- X_k 1단위 변화에 따른 Y 의 변화
- 다른 변수를 모두 동일하게 고정하고 X_k 만 변화시켰을 경우의 Y 의 변화 (net effect)

기본가정

1. i.i.d

$$Cov(\epsilon) = E(\epsilon\epsilon^T) = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & \dots \\ 0 & 1 & 0 & \dots \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 I$$

2. X 와 ϵ은 독립

$$E[\epsilon|X] = 0, \qquad E[X^T\epsilon|X] = X^T E[\epsilon|X] = 0$$

회귀계수의 추정

- 잔차(residual)의 제곱의 합을 가장 작게 하는 β 를 찾음

$$\min_\beta \Sigma(Y_i - X_i^T\beta)^2 = \min_\beta (Y - X\beta)^T(Y - X\beta)$$
$$\hat{\beta} = (X^TX)^{-1}(X^TY)$$

- $X = [1, x]$ 일 경우

$$(X^TX) = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \times \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x \\ \sum x & \sum x^2 \end{bmatrix} \qquad (X^TY) = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}$$

$$(X^TX)^{-1} = \frac{1}{[n\sum x^2 - (\sum x)^2]} \begin{bmatrix} \sum x^2 & -\sum x \\ -\sum x & n \end{bmatrix}$$
$$(X^TX)^{-1}(X^TY) = \frac{1}{[n\sum x^2 - (\sum x)^2]} \begin{bmatrix} \sum x^2 & -\sum x \\ -\sum x & n \end{bmatrix} \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix}$$

$$= \frac{1}{[n\sum x^2 - (\sum x)^2]} \begin{bmatrix} \sum x^2 \sum y - \sum x \sum xy \\ -\sum x \sum y + n \sum xy \end{bmatrix}$$

$$= \frac{1}{[n\sum x^2 - n^2(\bar{x})^2]}\begin{bmatrix} n\sum x^2\bar{y} - n\bar{x}\sum xy \\ -n^2\bar{x}\bar{y} + n\sum xy \end{bmatrix}$$

$$= \frac{1}{[\sum x^2 - n(\bar{x})^2]}\begin{bmatrix} \sum x^2\bar{y} - \bar{x}\sum xy \\ -n\bar{x}\bar{y} + \sum xy \end{bmatrix}$$

$$= \frac{1}{[\sum x^2 - n(\bar{x})^2]}\begin{bmatrix} \sum x^2\bar{y} - n\bar{x}^2\bar{y} + n\bar{x}^2\bar{y} - \bar{x}\sum xy \\ -n\bar{x}\bar{y} + \sum xy \end{bmatrix}$$

$$= \frac{1}{[\sum x^2 - n(\bar{x})^2]}\begin{bmatrix} \bar{y}[\sum x^2 - n\bar{x}^2] - \bar{x}[\sum xy - n\bar{x}\bar{y}] \\ \sum xy - n\bar{x}\bar{y} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\bar{y}[\sum x^2 - n\bar{x}^2] - \bar{x}[\sum xy - n\bar{x}\bar{y}]}{[\sum x^2 - n(\bar{x})^2]}, & \frac{\sum xy - n\bar{x}\bar{y}}{[\sum x^2 - n(\bar{x})^2]} \end{bmatrix}^T$$

$$= \begin{bmatrix} \bar{y} - \bar{x}\frac{S_{xy}}{S_x^2}, & \frac{S_{xy}}{S_x^2} \end{bmatrix}^T$$

```r
n=dim(hf.son)[1]
x=as.matrix(cbind(rep(1,n),hf.son$Father))
y=as.matrix(hf.son$Height)
bhat=(solve(t(x)%*%x))%*% (t(x)%*%y)
bhat
```

```
##             [,1]
## [1,] 38.2589122
## [2,]  0.4477479
```

```r
(h.reg=lm(Height~Father,data=hf.son))
```

```
##
## Call:
## lm(formula = Height ~ Father, data = hf.son)
##
## Coefficients:
## (Intercept)       Father
##     38.2589       0.4477
```

```r
summary(h.reg)
```

```
##
## Call:
```

```
## lm(formula = Height ~ Father, data = hf.son)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3774 -1.4968  0.0181  1.6375  9.3987
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.25891    3.38663   11.30   <2e-16 ***
## Father       0.44775    0.04894    9.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.424 on 463 degrees of freedom
## Multiple R-squared:  0.1531, Adjusted R-squared:  0.1513
## F-statistic: 83.72 on 1 and 463 DF,  p-value: < 2.2e-16
```

- 회귀분석은 수직거리가 가장 짧아지는 직선(평면)을 찾는다.

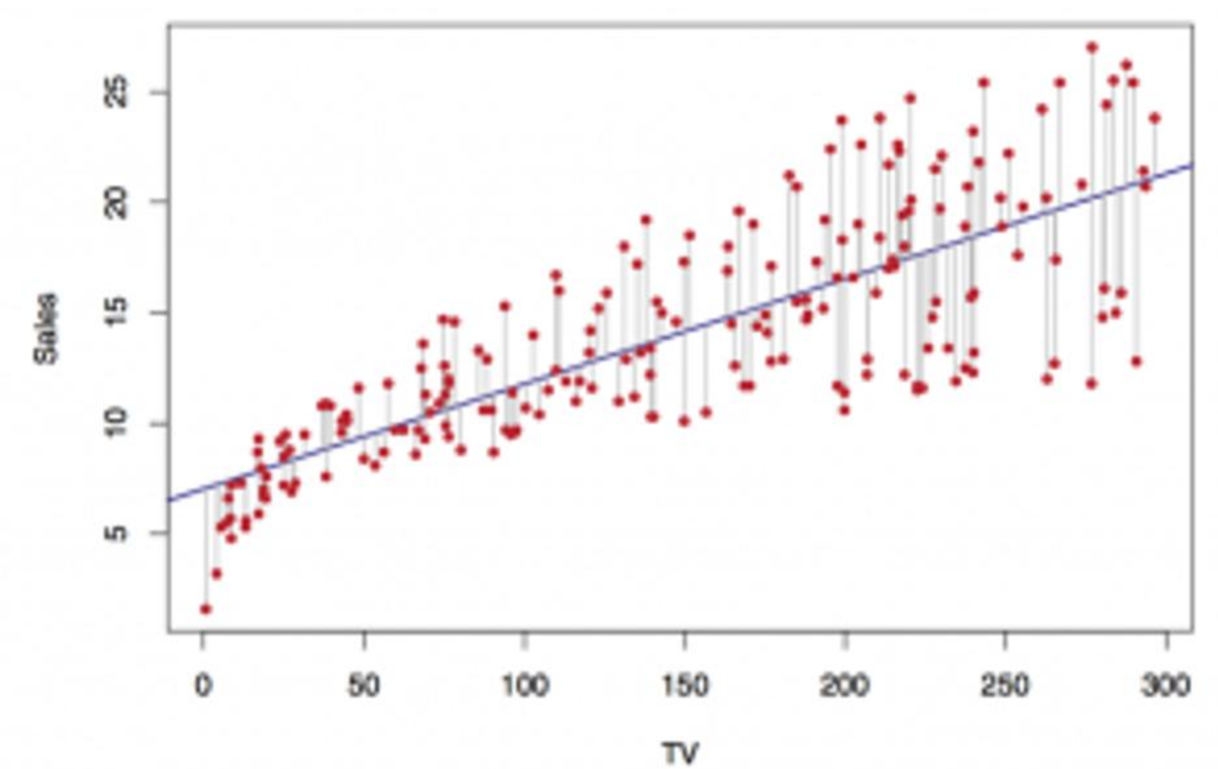


그림 1: 거리 최소화

회귀계수의 특징

$$\hat{\beta} = (X^TX)^{-1}(X^TY) = (X^TX)^{-1}(X^TX)\beta + (X^TX)^{-1}X^T\epsilon = \beta + (X^TX)^{-1}X^T\epsilon$$

평균: 불편성 (unbiasedness)

$$E[\hat{\beta}|X] = E[(X^TX)^{-1}(X^TY)|X]$$
$$= E[(X^TX)^{-1}(X^TX)\beta + (X^TX)^{-1}X^T\epsilon|X]$$
$$= \beta + E[(X^TX)^{-1}X^T\epsilon|X]$$
$$= \beta + (X^TX)^{-1}X^TE[\epsilon|X] = \beta$$

분산

$$Cov(\hat{\beta}|X) = E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T|X]$$
$$= E[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}|X]$$
$$= (X^TX)^{-1}X^T \times E[\epsilon\epsilon^T|X] \times X(X^TX)^{-1}$$
$$= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$$

분포

회귀계수 : 정규분포

$$\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$$
$$\hat{\beta}_k \sim N(\beta_k, \sigma_k^2) \qquad \sigma_k^2 = \sigma^2(X^TX)_k^{-1} = \text{kth diagonal of } (X^TX)^{-1}$$

잔차항(residual) : χ^2

- fitted value

$$\hat{Y} = X\hat{\beta}$$

- 잔차항 : residual

$$e \equiv Y - \hat{Y} = Y - X\hat{\beta}$$
$$e^Te/\sigma^2 = \sum e_i^2/\sigma^2 \sim \chi^2(n-k)$$
$$s^2 \equiv e^Te/(n-k), \qquad E(s^2|X) = E(e^Te/(n-k)|X) = \sigma^2$$

회귀분석

가설검정

모형의 유의성 : F test

$$\sum(y-\bar{y})^2 = \sum(y-\hat{y})^2 + \sum(\hat{y}-\bar{y})^2$$

- $\sum(y-\bar{y})^2$: Total Sum of Squres (TSS)
- $\sum(y-\hat{y})^2$: Sum of Squred errors (SSE)
- $\sum(\hat{y}-\bar{y})^2$: Regression Sum of Squres (RSS)

1. 가설

$$H_0 : \beta_{j\neq 0} = 0 \qquad .vs \qquad H_1 : \exists j \neq 0 \quad \beta_j \neq 0$$

2. 검정통계량

$$\sum(y-\hat{y})^2 \sim \chi^2(n-k), \qquad \sum(\hat{y}-\bar{y})^2 \sim \chi^2(k-1,\lambda) \quad \text{independent}$$

$$F = \frac{\sum(\hat{y}-\bar{y})^2/(k-1)}{\sum(y-\hat{y})^2/(n-k)} \sim F(k-1, n-k, \lambda)$$

귀무가설 하에서는 $\lambda = 0$ (Central F)

3. 기각역 : 단측검정 (오른쪽)

$$P(f > C_u) = \alpha, (C_u, \infty]$$

$$pvalue = P[f > F]$$

4. 기각/채택 결정 : p-value < 2.2e-16 <0.05 → 귀무가설 기각

```
(h.reg=lm(Height~Father,data=hf.son))
```

```
##
## Call:
## lm(formula = Height ~ Father, data = hf.son)
##
## Coefficients:
## (Intercept)       Father
##     38.2589       0.4477
```

```
summary(h.reg)
```

```
##
## Call:
## lm(formula = Height ~ Father, data = hf.son)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3774 -1.4968  0.0181  1.6375  9.3987
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.25891    3.38663   11.30   <2e-16 ***
## Father       0.44775    0.04894    9.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.424 on 463 degrees of freedom
## Multiple R-squared:  0.1531, Adjusted R-squared:  0.1513
## F-statistic: 83.72 on 1 and 463 DF,  p-value: < 2.2e-16
```

```r
s.h.reg=summary(h.reg)
s.h.reg$fstatistic
```

```
##     value     numdf     dendf
##  83.71863   1.00000 463.00000
```

```r
F=s.h.reg$fstatistic[1]
df1=s.h.reg$fstatistic[2]
df2=s.h.reg$fstatistic[3]

(p.value=1-pf(F,df1,df2))
```

```
## value
##     0
```

```r
(c.u=qf(1-0.05,df1,df2))
```

```
## [1] 3.861621
```

```r
ifelse(F>c.u,"Reject H0","Not reject H0")
```

```
##       value
## "Reject H0"
```

```r
#p.value<0.05

ifelse(p.value<0.05,"Reject H0","Not reject H0")
```

```
##        value
## "Reject H0"
```

개별 계수의 유의성 : t test

1. 가설

$$H_0 : \beta_k = 0 \quad .vs \quad H_1 : \beta_k \neq 0$$

2. 검정통계량

$\hat{\beta}$,e 는 독립이다. 따라서

$$\begin{aligned} T &= \frac{(\hat{\beta}_k - \beta_k)/\sigma\sqrt{(X^TX)_k^{-1}}}{\sqrt{e^Te/\sigma^2(n-k)}} \\ &= \frac{\hat{\beta}_k - \beta_k}{\sigma\sqrt{(X^TX)_k^{-1}}\sqrt{e^Te/\sigma^2(n-k)}} \\ &= \frac{\hat{\beta}_k - \beta_k}{\sqrt{s^2/(X^TX)_k^{-1}}} \sim t(n-k) \end{aligned}$$

3. 기각역, 임계치 (유의수준 α), p value

$$P[|t| > c_{n-k,\alpha/2}] = \alpha, \quad (-c_{n-k,\alpha/2}, c_{n-k,\alpha/2})$$

$$P[t > |T|]$$

```r
s.h.reg$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 38.2589122 3.38663400 11.297032 2.642076e-26
## Father       0.4477479 0.04893533  9.149788 1.824016e-18
```

```r
(beta=s.h.reg$coefficients[,1])
```

```
## (Intercept)      Father
##  38.2589122   0.4477479
```

```r
(sdx=s.h.reg$coefficients[,2])
```

```
## (Intercept)      Father
##  3.38663400  0.04893533
```

```r
(tstats=s.h.reg$coefficients[,3])
```

```
## (Intercept)      Father
##   11.297032    9.149788
```

```r
(pvalue.t=s.h.reg$coefficients[,4])
```

```
##  (Intercept)       Father
## 2.642076e-26 1.824016e-18
```

모형의 평가 : 모형의 Fit (Goodnesss of fit) R^2

$$\sum(y-\bar{y})^2 = \sum(y-\hat{y})^2 + \sum(\hat{y}-\bar{y})^2$$
$$1 = \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} + \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}$$
$$R^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}$$

- R^2 는 변수를 n 개를 사용하면 잔차항이 0 이 되어서 최대값인 1 이 된다. 이 경우에는 모든 sample 마다 그 sample을 설명하는 변수를 하나씩 갖게 되기 때문에 종속변수의 변화를 일으키는 주요 요인을 파악할 수 없게 된다. 하지만 R^2 는 이러한 '변수가 지나치게 많은 모형'을 더 좋은 모형으로 판단하는 단점이 있다.
- 이러한 단점을 극복하기 위해서 R^2 를 개량하려는 시도는 많이 이루어졌지만, 특히 어떤 방법이 더 좋다는 결론은 내려져 있지 않다. 조정 R^2 (adjusted rsqure) 역시 이러한 시도의 한 예이다.

$$adjR^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

n 이 아주 크면 R^2 에 근접하지만, k (변수의 수) 가 증가하면 값이 하락한다.

추정 $E(Y|X)$

1. 회귀분석의 경우 주된 추정의 대상은 unknown parameter β 및 종속변수 Y의 기대값 $E(y|X)$ 이다.

$$E(Y|X) = E(X^T\beta + \epsilon|X) = X\beta$$

2. 특히 회귀분석은 새로운 독립변수 X 에 대응하는 '아직 관찰되지 않은' Y 값의 기대값$E(Y|X)$ 를 추정하는 수단으로 자주 활용된다. 예를 들어 '배출규제 위반 벌금을 10% 올렸을 경우 배출량' 을 추정하는 전형적인 정책 simulation의 경우 '벌금 10% 인상'은 새로운 독립변수, '배출량'은 그에 대응하는 종속변수의 값이다.

$$Y_{n+1} = X_{n+1}\beta + \epsilon$$

$$E(Y_{n+1}|X) = X_{n+1}\beta$$

3. $E(Y|X)$ 는 보이지 않는 값이기 때문에 추정량을 사용하여 추정하며, 따라서 신뢰구간이 존재한다.

$$E(Y|X) = X\beta \Leftarrow X\hat{\beta}$$

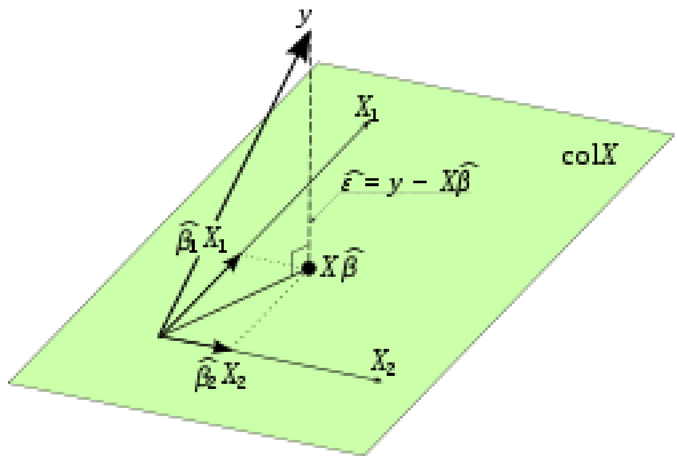


그림 2: Regression and Projection

$E(Y|X)$ 의 점추정

$$\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY = P_xY$$

$E(Y|X)$ 의 구간추정

$$E(Y|X) = X\beta$$
$$\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY = X(X^TX)^{-1}X^T[X\beta + \epsilon] = X\beta + X(X^TX)^{-1}X^T\epsilon = X\beta + P_x\epsilon$$
$$\hat{Y} \sim N(X\beta, \sigma^2 P_x)$$
$$\hat{Y}_i \sim N(X_i\beta, \sigma^2 P_{x,i}) \quad P_{x,i} : \text{ith diagonal of } P_x$$
$$T = \frac{\hat{Y}_i - X_i\beta}{s\sqrt{P_{x,i}}} \sim t(n-k) \qquad s = \sqrt{\sum(y-\hat{y})^2/(n-k)}$$

- 95% 신뢰구간은?

$$C.I = (\hat{Y}_i - t_{n-k,0.025}s\sqrt{P_{x,i}}, \quad \hat{Y}_i + t_{n-k,0.025}s\sqrt{P_{x,i}})$$

여기서 $X = [1, \quad x]$ 인 경우

$$P_{x,i} = \begin{bmatrix} 1 & x_i \end{bmatrix} \frac{1}{[n\sum x^2 - (\sum x)^2]} \begin{bmatrix} \sum x^2 & -\sum x \\ -\sum x & n \end{bmatrix} \begin{bmatrix} 1 \\ x_i \end{bmatrix}$$
$$= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x - \bar{x})^2}$$
$$C.I = \left(\hat{Y}_i - t_{n-2,0.025}s\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x-\bar{x})^2}}, \quad \hat{Y}_i + t_{n-2,0.025}s\sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum(x-\bar{x})^2}}\right)$$

이러한 신뢰구간은 평균값에서 가장 작고, 평균에서 멀어질수록 값이 커진다.

관측되지 않은 $E(Y_{n+1}|X)$ 추정

- 점추정치

$$E(Y_{n+1}|X) = X_{n+1}\beta \leftarrow X_{n+1}\hat{\beta}$$

$$\hat{Y}_{n+1} = X_{n+1}\hat{\beta} = X_{n+1}(X^TX)^{-1}X^TY = X_{n+1}(X^TX)^{-1}X^T[X\beta+\epsilon] = X_{n+1}\beta + X_{n+1}(X^TX)^{-1}X^T\epsilon$$

$$E(\hat{Y}_{n+1}) = X_{n+1}\beta$$

$$V(\hat{Y}_{n+1}) = E[X_{n+1}(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}X_{n+1}^T]$$

$$= \sigma^2[X_{n+1}(X^TX)^{-1}X^TX(X^TX)^{-1}X_{n+1}^T] = \sigma^2X_{n+1}(X^TX)^{-1}X_{n+1}^T$$

$$\hat{Y}_{n+1} \sim N(X_{n+1}\beta, \sigma^2X_{n+1}(X^TX)^{-1}X_{n+1}^T)$$

- 95% 신뢰구간

$$C.I = \left(\hat{Y}_{n+1} - t_{n-k,0.025}s\sqrt{X_{n+1}(X^TX)^{-1}X_{n+1}^T},\quad \hat{Y}_{n+1} + t_{n-k,0.025}s\sqrt{X_{n+1}(X^TX)^{-1}X_{n+1}^T}\right)$$

$X = [1, x]$ 인 경우

$$V(\hat{Y}_{n+1}) = \begin{bmatrix}1 & x_{n+1}\end{bmatrix} \frac{1}{[n\sum x^2 - (\sum x)^2]} \begin{bmatrix}\sum x^2 & -\sum x \\ -\sum x & n\end{bmatrix} \begin{bmatrix}1 \\ x_{n+1}\end{bmatrix}$$

$$= \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum(x-\bar{x})^2}$$

$$C.I = \left(\hat{Y}_{n+1} - t_{n-2,0.025}s\sqrt{\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum(x-\bar{x})^2}},\quad \hat{Y}_{n+1} + t_{n-2,0.025}s\sqrt{\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum(x-\bar{x})^2}}\right)$$

Prediction interval: Y_{n+1} 이 추정 대상일 경우

$$Y_{n+1} \sim N(X_{n+1}\beta, \epsilon) \quad \hat{Y}_{n+1} = X_{n+1}\hat{\beta}$$

$$\hat{Y}_{n+1} - Y_{n+1} = X_{n+1}\hat{\beta} - X_{n+1}\beta + \epsilon_{n+1}$$

$$X_{n+1}\hat{\beta} - X_{n+1}\beta = X_{n+1}(\hat{\beta}-\beta) \sim N(0, \sigma^2X_{n+1}(X^TX)^{-1}X_{n+1}^T)$$

$$\epsilon_{n+1} \sim N(0, \sigma^2)$$

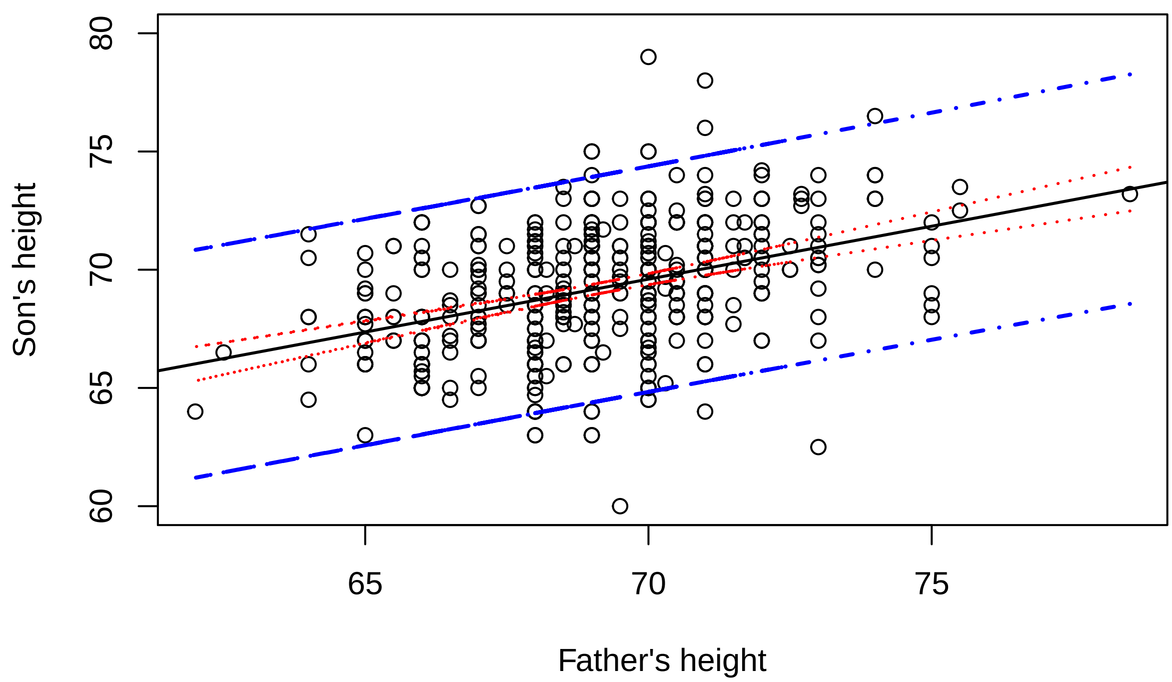
$$\hat{Y}_{n+1} - Y_{n+1} \sim N(0, \sigma^2\left[1 + X_{n+1}(X^TX)^{-1}X_{n+1}^T\right])$$

이 경우의 신뢰구간을 Prediction interval 이라고 한다.

$$P.I = \left(\hat{Y}_{n+1} - t_{n-2,0.025}s\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n(x-\bar{x})^2}},\quad \hat{Y}_{n+1} + t_{n-2,0.025}s\sqrt{1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n(x-\bar{x})^2}}\right)$$

아래 그림에서 확인할 수 있듯이, Y 에 대한 신뢰구간은 $E(Y|X)$ 에 대한 신뢰구간보다 훨씬 넓다. 추정치의 불확실성과 오류항의 불확실성이 모두 반영되기 때문이다.

```r
# 그림 9-5
#no <- par(no.readonly = TRUE)
#par(mar=c(2,2,2,1))
plot(Height~Father, data=hf.son, main="",
     xlab="Father's height", ylab="Son's height",
     ylim=c(60, 80)
     )
abline(h.reg, lwd=1.5)
ci <- predict(h.reg, interval="confidence")
prd <- predict(h.reg, interval="predict")
lines(hf.son$Father, ci[,2], lty=3, lwd=1.5, col="red")
lines(hf.son$Father, ci[,3], lty=3, lwd=1.5, col="red")
lines(hf.son$Father, prd[,2],lty=4, lwd=2,col="blue")
lines(hf.son$Father, prd[,3],lty=4, lwd=2,col="blue")
```




```r
#par(no)
```