

2장 회귀분석

산포도

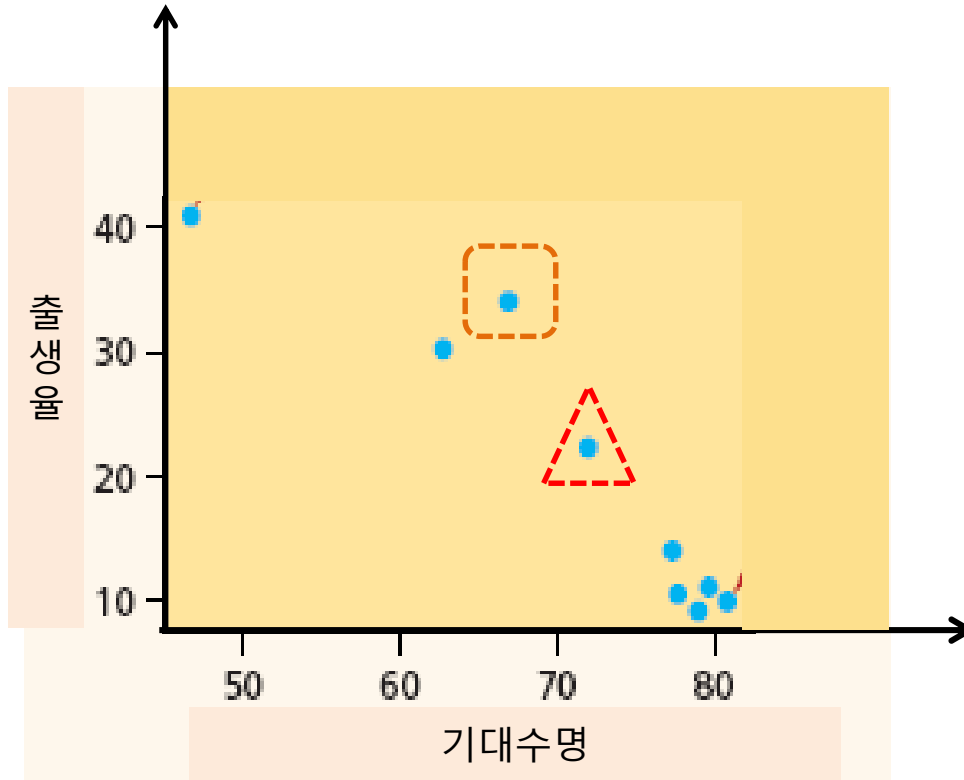
- 자료구조:

데이터 3요소(주체(subject), 변수(variable), 관찰값(observation)) 구분

국가명	출생율	기대수명
Afghanistan	41.03	46.60
Canada	11.09	79.70
Finland	10.60	77.80
Guatemala	34.17	66.90
Japan	10.03	80.90
Mexico	22.36	72.00
Pakistan	30.40	62.70
Spain	9.29	79.10
United States	14.10	77.40

산포도

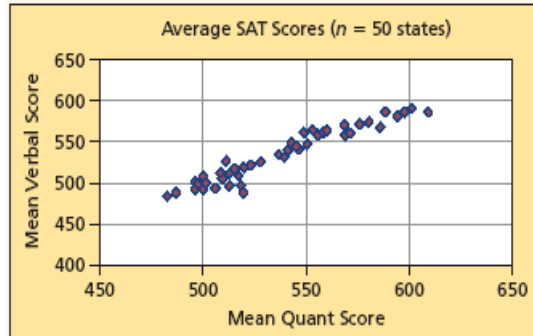
- 산포도가 제공하는 정보



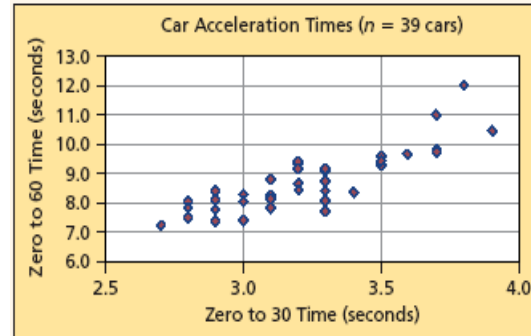
- Association(패턴)이 있나?
- Positive/ Negative?
- Strength(강한 정도)?
- 인과관계(cause & effect)?

산포도

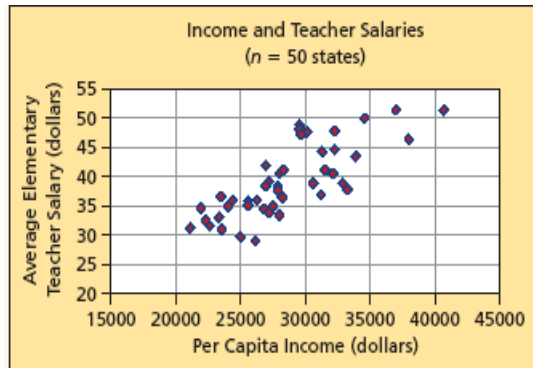
- Strength (강한 정도)



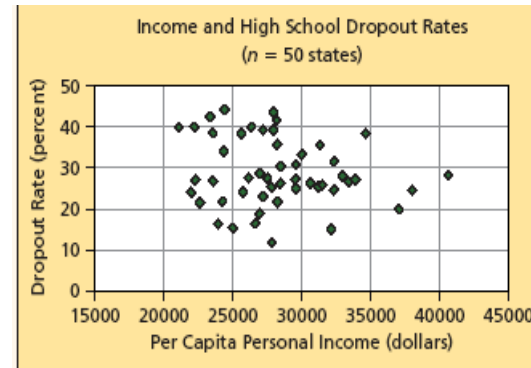
매우 강한 association



강한 association



보통의 association



작은 association

Correlation

- 상관계수(correlation): X와 Y의 선형관계(linear relation)의 강한 정도(strength)

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$(-1 \leq \rho \leq 1)$$

$\rho = 1$: X와 Y는 a perfect positive linear relationship.

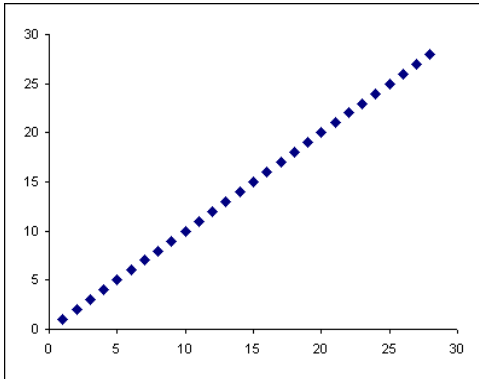
$\rho = -1$: X와 Y는 a perfect negative linear relationship.

$\rho = 0$: X와 Y는 a perfect no linear relationship.

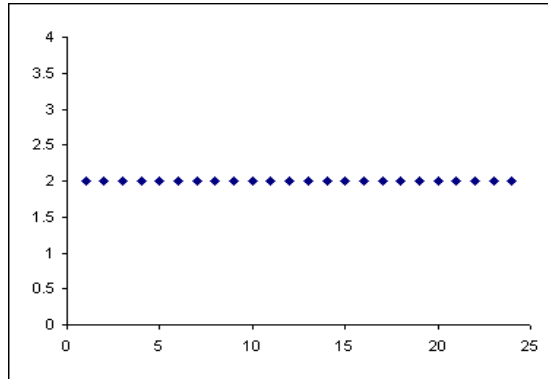
Correlation

- 상관계수 예

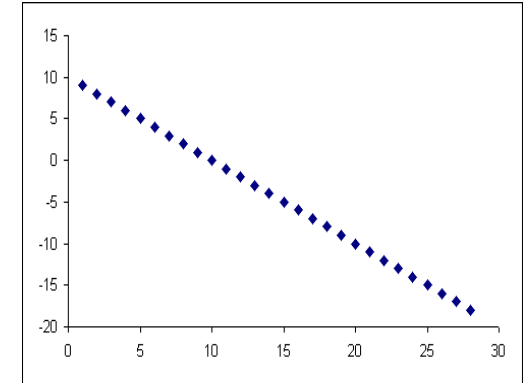
$$\rho = 1:$$



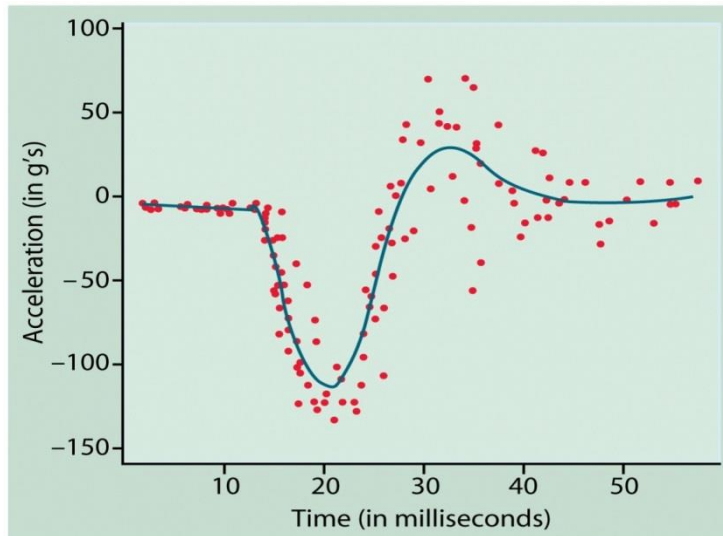
$$\rho = 0:$$



$$\rho = -1:$$



- 비선형 관계에서의 상관계수

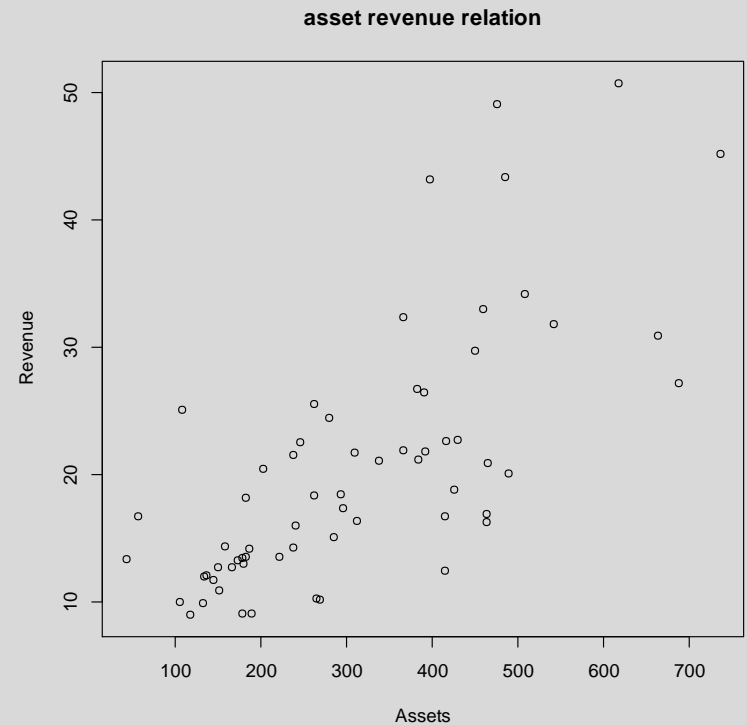


R code

■ 산점도 그리기

```
> library(lmtest); # call R packet
> bankdat=read.table("F:/bank.txt", header=T); # read the data set
> attach(bankdat);
> x11();
> plot(Assets, Revenue, main='asset revenue relation');

> cor(bankdat);
> with(bankdat, cor.test(Asset, Revenue));
```



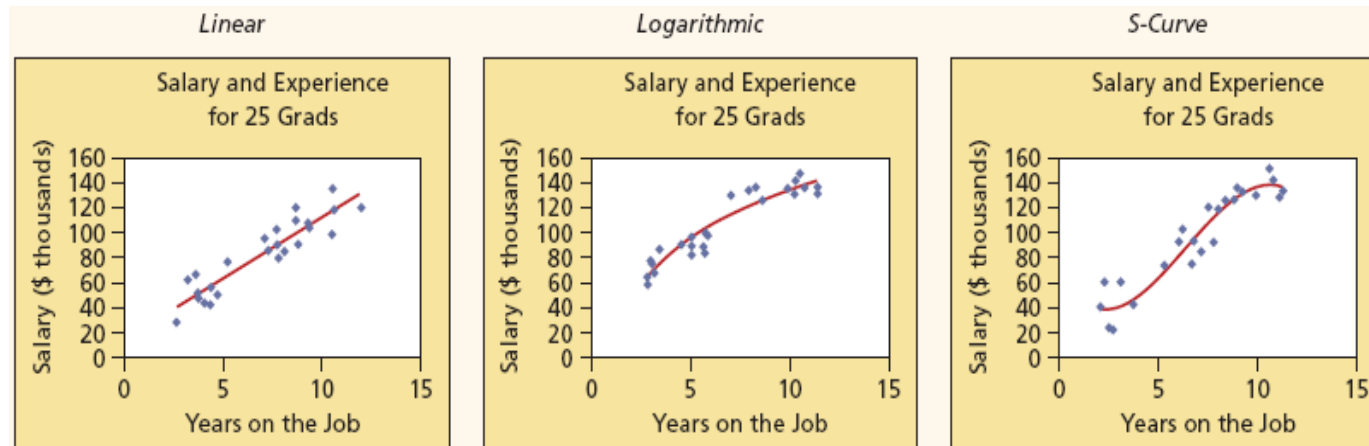
단순 선형 회귀 (simple linear regression)

- 단순 선형 회귀:

수평축에 X 변수: 설명변수, 예측변수, 독립변수

수직축에 Y 변수: 반응변수, 종속변수

- 선형 관계 모델의 장점



단순 선형 회귀 (simple linear regression)

- 단순 선형 회귀:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ 이고 } \varepsilon_i \sim i.i.d \ N(0, \sigma_\varepsilon^2)$$

절편 계수 (intercept coefficient) 기울기 계수 (slope coefficient) 오차항

* 모수: $\beta_0, \beta_1, \sigma_\varepsilon^2$ 표본 자료 (X와 Y)를 이용해서 추정해야 함.

* 관찰이 가능한 값: X와 Y 뿐

❖ IID(*i.i.d*: independent & identical) 또는 random sample의 의미는?

- 선형 회귀선 찾기:

각각의 X 값에 대한 선형관계의 Y 값 찾기

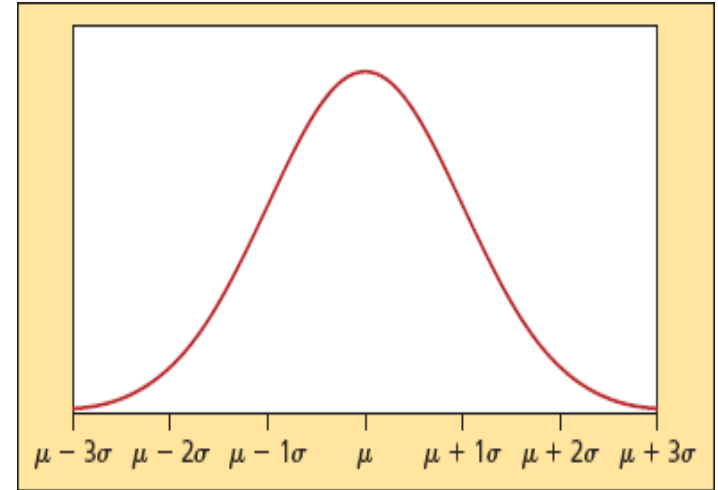
$$\hat{y}_i = b_0 + b_1 x_i \quad \leftarrow \quad \beta_0 + \beta_1 x_i \text{ 의 추정치(estimate)}$$

단순 선형 회귀 (simple linear regression)

- 정규분포(Normal) :

$$X \sim N(\mu, \sigma^2)$$

- 모수: $\mu = E(X), \quad \sigma^2 = V(X)$
- 분포식(확률밀도함수): $f(x) = \dots$
- 값의 범위: $-\infty < X < \infty$
- 특징: 1)경험룰(empirical rule)
 - 2) 종모양의 좌우 대칭 확률밀도함수



단순 선형 회귀 (simple linear regression)

- IID: Independent & Identical (또는 Random Sample)

표본을 추출할 때, 별다른 언급이 없으면 앞으로는 IID 방법으로 추출한 것으로 가정함.
Independent는 독립을 의미한다. 표본 집단 이 있을 때, 개개의 표본이 추출될 확률은 다른 표본의 추출에 의해서 영향을 받지 않는다는 의미이다.

Identical의 의미는 표본 집단의 개개의 표본이 모두 같은 모집단으로부터 나온 경우를 의미한다. 보다 구체적으로 표현하면, 1) 모두 **같은 확률분포**를 따르고, 이들의
2) **모수도 모두 같은 경우**이다.

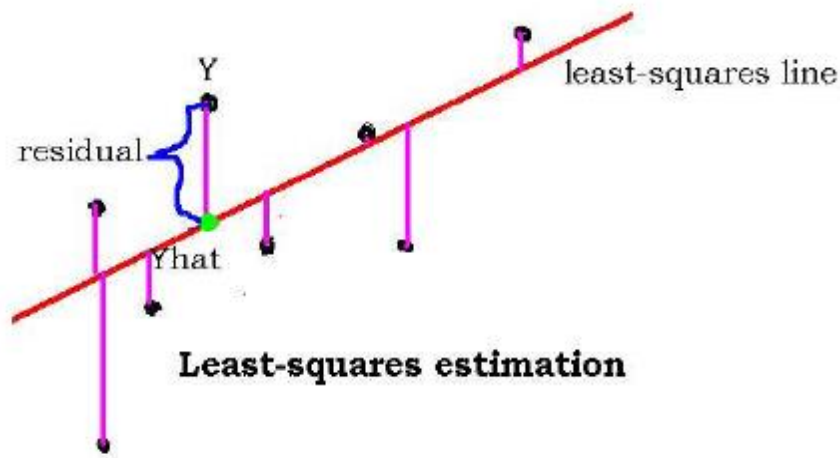
두 확률변수 X와 Y가 있는데 두 변수 모두 정규분포를 따르되, 모수가 다르면 Identical이 성립되지 않는다.

단순 선형 회귀 (simple linear regression)

- b_0 와 b_1 찾기: OLS(Ordinary Least Square) 방법이용

$$\min \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \min \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2$$

Least-squares estimation



$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Gauss-Markov Theorem

- The OLS estimates β_0, β_1 for $y = \beta_0 + \beta_1 x + \varepsilon$ is BLUE under the three conditions:

$$E(\varepsilon) = 0$$

$$V(\varepsilon) = \sigma^2$$

$$\text{Corr}(\varepsilon_i, \varepsilon_j) = 0$$

- BLUE is Best Linear Unbiased Estimator.

단순 선형 회귀 (simple linear regression)

- σ_ε^2 추정 : 잔차(residual) 이용

잔차: $e_i = y_i - \hat{y}_i$

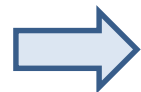
$$\hat{\sigma}_\varepsilon = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

단순 선형 회귀: Coefficient 검정

- Coefficients 검정: t 검정

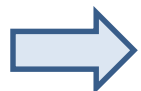
(t 검정을 위해선 회귀모형의 오차항이 정규분포를 따라야 함)

1) 절편 검정 $H_0 : \beta_0 = 0$ vs. $H_1 : \beta_0 \neq 0$

 p-value 이용

2) 기울기 검정

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0$$

 p-value 이용

단순 선형 회귀: 모델 검정 (ANOVA표와 R^2)

- 분산분해:
$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$\begin{array}{ccccc} \text{SST} & = & \text{SSE} & + & \text{SSR} \\ \text{(total variation} & & \text{(unexplained or} & & \text{(variation explained} \\ \text{around the mean)} & & \text{error variation)} & & \text{by the regression)} \end{array}$$

- X와 Y가 선형 관계 없으면...

$$b_1=0 \text{ 이면 } \hat{y} = \bar{y} \text{ (Why? } \hat{y} = b_0 + b_1x, \quad b_0 = \bar{y} - b_1\bar{x} \text{)}$$

- R^2 (결정계수):
$$R^2 = 1 - \frac{SSE}{SST} \quad \text{or} \quad R^2 = \frac{SSR}{SST}$$

$$0 \leq R^2 \leq 1$$

단순 선형 회귀: 모델 검정 (ANOVA표와 R^2)

- 모델 검정

H_0 : model is not valid vs. H_1 : model is valid

 $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

- ANOVA 표

Source	SS	d.f.	MS	F	P-value
Regression	SSR	1	MSR=SSR/1	MSR/MSE	
Residual	SSE	n-2	MSE=SSE/(n-2)		
Total	SST	n-1			

- F 검정

$$F = \text{MSR}/\text{MSE} = (\text{SSR}/1)/(\text{SSE}/(n-2)) \sim F_{1, (n-2)}$$

모델 진단 (잔차 분석 포함)

■ 모델 진단: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $\varepsilon_i \sim i.i.d \ N(0, \sigma_\varepsilon^2)$

1) x 와 y 는 선형 관계

2) $E(\varepsilon) = 0$

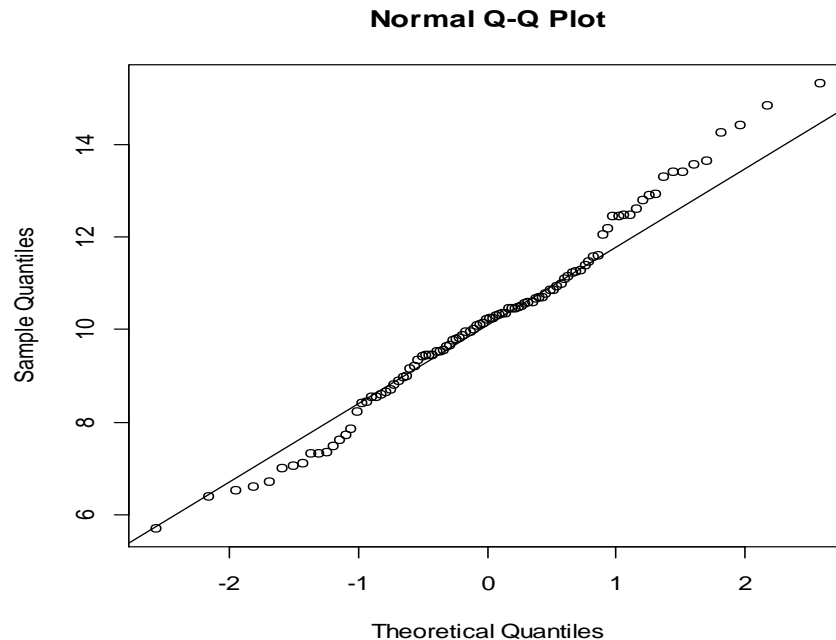
3) $V(\varepsilon) = \sigma_\varepsilon^2$ (동분산성: Breusch-Pagan 검정)

4) 오차항은 서로 독립 (상관관계 없는지 확인: Durbin-Watson 검정)

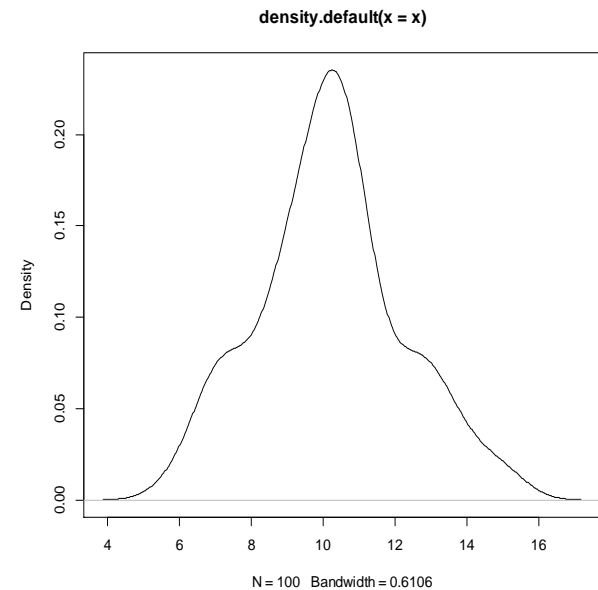
5) 오차항은 정규분포를 따름

모델 진단 (정규성 분석: Normal Q-Q plot)

- Normal Q-Q plot: (x축: 정규분포의 이론상 quantile, y축: sample quantile)
(정규성 판단) 산포도의 점이 Q-Q line 선상에 있으면 정규분포를 따른다고 판정(임의적)



[Normal Q-Q plot]



확률분포(pdf)

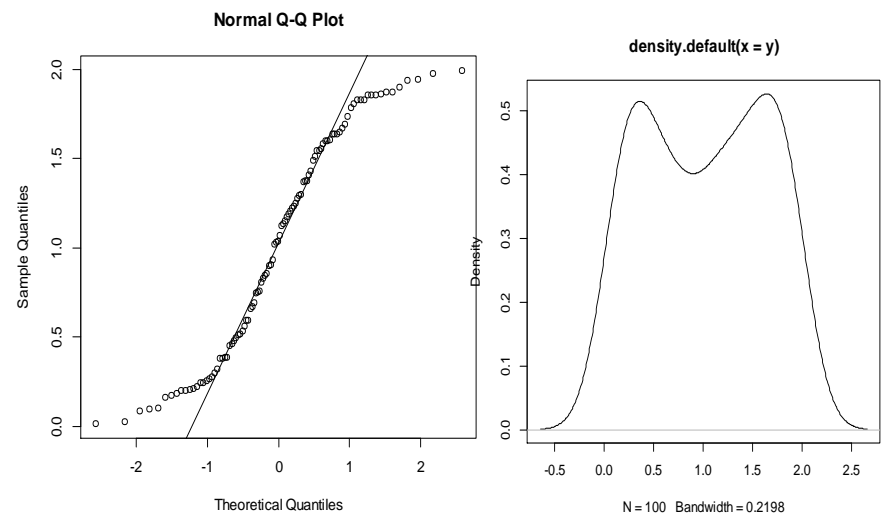
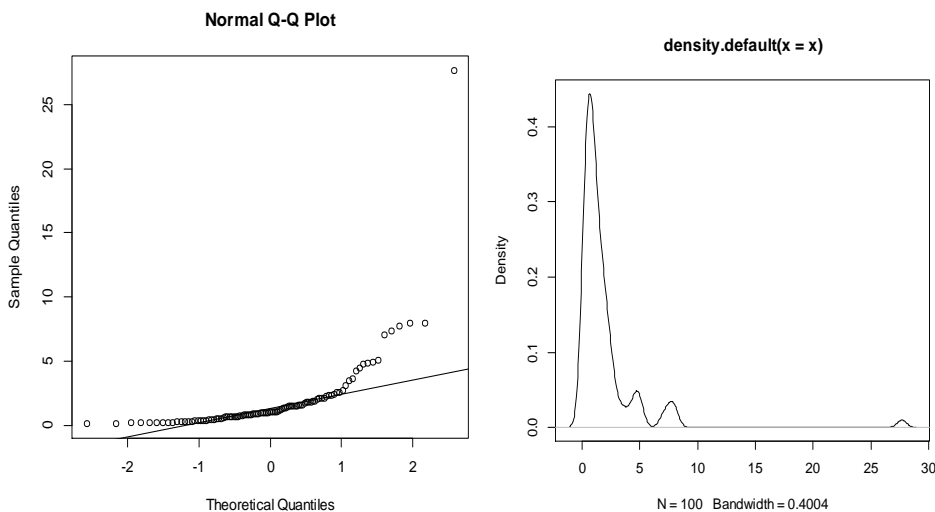
❖ pdf: probability density function

모델 진단 (정규성 분석: Normal Q-Q plot)

- Normal Q-Q plot의 산포도 모양으로 분포의 종류 추론 가능

Normal Q-Q plot의 x축 중간 값은 0이고, 이를 중심으로 좌측과 우측으로 구분함.

- 좌측에서 산포도 점들이 Q-Q line 보다 위 쪽에 있으면 해당 자료의 분포 좌측은 short-tail 형태
- 좌측에서 산포도 점들이 Q-Q line 보다 아래 쪽에 있으면 해당 자료의 분포 좌측은 long-tail 형태
- 우측에서 산포도 점들이 Q-Q line 보다 위 쪽에 있으면 해당 자료의 분포 우측은 long-tail 형태
- 우측에서 산포도 점들이 Q-Q line 보다 아래 쪽에 있으면 해당 자료의 분포 우측은 short-tail 형태



모델 진단 (정규성 분석: GoF 검정)

- Normal GoF (Goodness of Fit) 검정

- (1) Kolmogorov-Smirnov 방법

- (2) Jarque Bera 방법

- (3) Shapiro Wilk 방법

- ❖ 가설 정립 ($F_0 \sim$ 정규분포)

- $H_0: F = F_0$

- $H_1: F \neq F_0$

- ❖ P-value가 0.1 이하면 해당 자료는 정규분포를 따르지 않는다고 결론

모델 진단

- 이상치 판단:

1) X 방향 이상치: leverage 이용

⇒ leverage가 $3/n$ 보다 크면 이상치

다중 회귀 k 개 독립변수인 경우는 $2(k+1)/n$ 보다 크면 이상치

2) Y 방향 이상치: studentized deleted residual (sdr) 이용

⇒ sdr의 절대값이 2 이상이면 이상치

3) Influential point: Cook's distance 이용

⇒ Cook's D가 1 이상이면 influential point

모델 진단

- 이상치

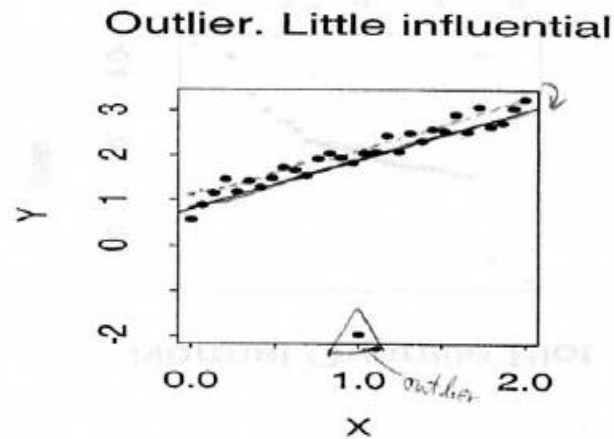
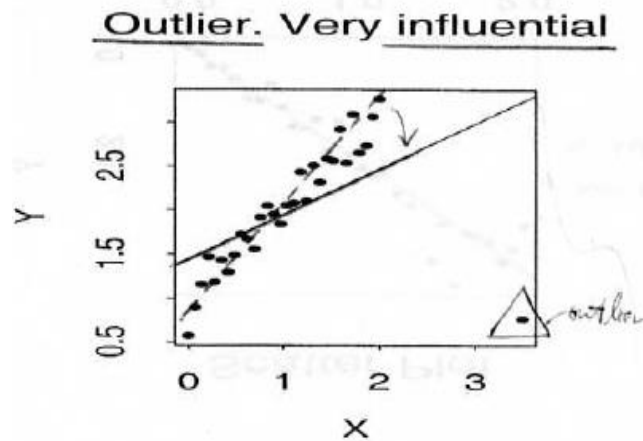
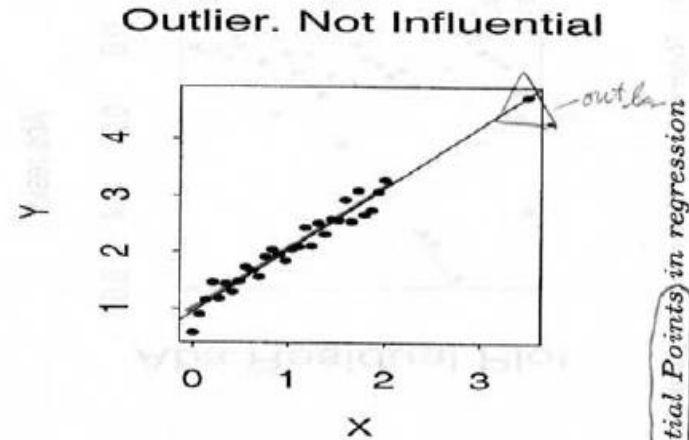
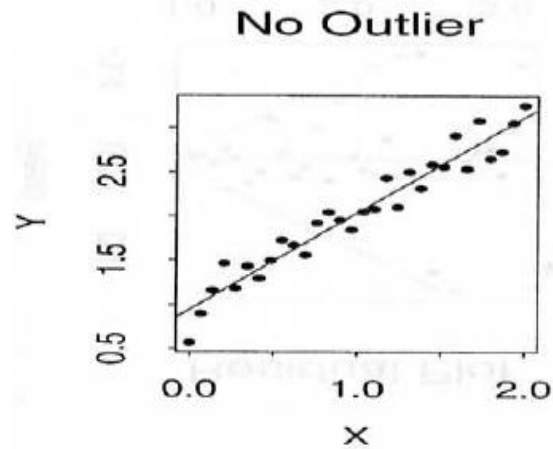


Figure 5: Outliers and Influential Points in regression

Lurking Predictor

- Some Predictor Variables are **lurking**

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \quad \Rightarrow \quad \text{Outliers are detected}$$

- Solution: add the lurking variables in the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- (Note) y : eye sight, x_1 : TV watching time, x_2 : Age

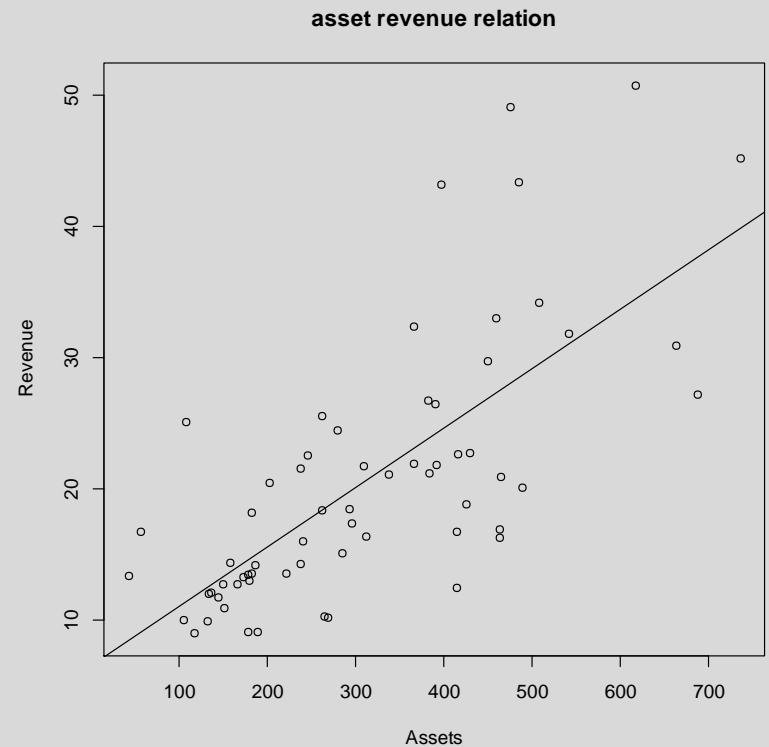
회귀분석 연습

회귀분석

```
> library(lmtest); # call R packet
> bankdat=read.table("F:/bank.txt", header=T); # read the data set
> attach(bankdat);
> out=lm(Revenue~Assets) ;# run linear regression model (Y: Revenue, X: Assets)
> summary(out);
> x11();
> plot(Assets, Revenue, main='asset revenue relation');
> abline(out);
```

```
# compute correlation btw two variables
```

```
> cor(Assets, Revenue);
> cor.test(Assets, Revenue);
> g=lm(Revenue~Assets)
> anova(g); # ANOVA Table
> vcov(g) ; # variance-covariance matrix
```



■ 회귀분석 연습

■ 잔차분석

```
> library(lmtest)
> bptest(Revenue~Assets, data=bankdat)    # Breusch-Pagan test (heteroscedasticity test)
> bptest(g)
> dwtest(Revenue~Assets, data=bankdat)    # Durbin-Watson test (correlation test)
> dwtest(g)
> plot(fitted(g), residuals(g), xlab="fitted", ylab="residuals") # constant variance?
> qqnorm(residuals(g), ylab="residuals") # normality?
> qqline(residuals(g))
```

■ 회귀분석 연습

■ 잔차의 정규분포 여부 확인

```
# Normal Q-Q plot
```

```
mu=10; sigma=2; k=100; x=rnorm(k, mu, sigma);
```

```
qqnorm(x); qqline(x);
```

```
plot(density(x));
```

```
# Normal Q-Q plot Exercises: check the shape of distribution
```

```
y1=runif(500); qqnorm(y1);qqline(y2);          # uniform distribution: (left: short, right: short)
```

```
y2=exp(rnorm(500)); qqnorm(y2);qqline(y2); # log-normal distribution: (left: short, right: long)
```

```
# Normal Goodness of Fit Test
```

```
ks.test(x, "pnorm", m=mean(x), sd=sd(x)); # K-S test
```

```
library(tseries);
```

```
jarque.bera.test(x); # J-B test
```

```
shapiro.test(x); # S-W test
```

■ 회귀분석 연습

■ 이상치 분석

```
ginf=influence(g) # leverage point
ginf$hat          # leverage point printout
plot(ginf$hat)
gs=summary(g)
gs$sig
r1=rstudent(g) # studentized deleted residuals
plot(r1)
cg=cooks.distance(g) # cooks distance
plot(cg)
par(mfrow=c(2,2)) # par(): 그림의 인수를 조정, mfrow=c(2,2): 한 화면에 4개의 plot 그리기
plot(bptest(g))
```

회귀모형을 이용한 예측

- 새로운 자료로 예측:

```
bankdat=read.table("F:/bank.txt", header=T); # read the data set

# 단순회귀함수
mbk=lm(Revenue~Assets, bankdat)
lm.error=mbk$residuals
(lmRMSE=sqrt(mean(lm.error^2)))

predict(mbk, new=bankdat[15:18,]); # 15~18번째 Assets 자료를 이용한 추정
fitted(mbk);

# 새로운 자료로 예측
library(dplyr);
str(bankdat);
dam=data.frame(Assets=c(305.2, 403.7), Revenue=rep(NA,2));
predict(mbk, new=dam);
```

회귀모형을 이용한 예측

■ 미래 예측:

Example

```
x=c(1:20); # x축: time 1~20까지 있음  
y=c(3,4,8,4,6,9,8,12,15,26,35,40,45,54,49,59,60,62,63,68);  
data1=data.frame(x,y); x11(); plot(data1,pch=16)
```

단순회귀함수 추정

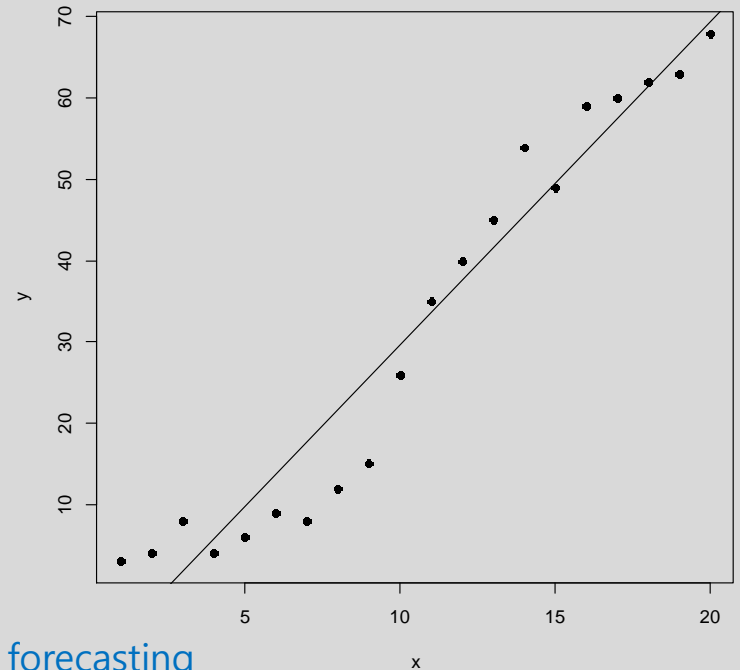
```
out=lm(y~x, data1); abline(out);  
lm.error=out$residuals;  
(lmRMSE=sqrt(mean(lm.error^2))); # 추정오차  
predict(out, new=data1[15:18,]); # x가 15~18일 때 추정  
fitted(out); # 모든 추정치
```

in-sample(1~15)/ out-of-sample(16~20) 구분

```
data2=data.frame(x=x, y0=c(y[1:15], rep(NA,5)));  
out0=lm(y0~x, data2[1:15,]); # in-sample fit  
p1=predict(out0, new=data2[16:20,]); # out-of-sample forecasting  
sqrt(mean((y[16:20]-p1)^2)) # rmse
```

future forecasting(21~25)

```
data3=data.frame(x=c(21:25),y=rep(NA,5));  
library(dplyr);  
all=bind_rows(data1, data3);  
predict(out, new=all[21:25,]);
```



회귀모형을 이용한 예측

```
#### initial in-sample(1~20) vs. out-of-sample(21~25) approach
# 1-time ahead forecasting using window shifts whose length is 20
x=c(1:25);
y=c(3,4,8,4,6,9,8,12,15,26,35,40,45,54,49,59,60,62,63,68,74,80,85,89,95);
y1=c(y[1:20], rep(NA, 5));
data4=data.frame(x, y1); # in-sample: 1~20, out-of-sample: 21~25 data
dim(data4);

# out-of-sample forecasting w/ 1-time ahead forecasting & window shift
for(i in 1:5)
{
  id1=i+1-1;
  id2=i+19;
  id3=i+20;
  out1=lm(y1~x, data4[c(id1:id2), ]);
  py=predict(out1, new=data4[id3, ])
  data4[id3,2]=py;
  print(i)
}

# 1-time ahead prediction result using window shift for the range 21 to 25
p2=data4[c(21:25),2]
sqrt(mean((y[21:25]-p2)^2)) # rmse
```

다중 회귀 분석을 위한 자료

- 자료

아파트 동	가격 (Y)	넓이(X1)	주차장 크기 (X2)	화장실개수(X3)
1	505.5	2192	16.4	2.5
2	784.1	3429	24.7	3.5
3	649.0	2842	17.7	3.5
4	689.8	2987	20.3	3.5
...
...
15	490.5	2134	13.4	2.5

다중 선형 회귀 (multiple linear regression)

- 다중 선형 회귀:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i \text{ 이고 } \varepsilon_i \sim i.i.d \ N(0, \sigma_\varepsilon^2)$$

➡ OLS 방법으로 모수 추정

➡ 오차항의 표준편차 추정:
$$\hat{\sigma}_\varepsilon = \sqrt{\frac{1}{n - (k + 1)} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

❖ 중요 조건

- 1) $n \gg k$
- 2) 독립변수들끼리는 서로 독립

다중 선형 회귀 (multiple linear regression)

- (독립) 변수 선택 논리

- 1) 부호의 논리성 판단
- 2) 통계적 유의성 고려
- 3) 가능하면 간단한 모델 선호
- 4) 안정성: 다중공선성 고려

- 예)

가격 (Y)	넓이(X1)	주차장 크기 (X2)	화장실개수(X3)
--------	--------	-------------	-----------

$$M1: \hat{Y} = 15.47 + 0.222X_1$$

$$M2: \hat{Y} = -23.21 + 0.187X_1 + 6.60X_2$$

$$M3: \hat{Y} = -28.85 + 0.171X_1 + 6.78X_2 + 15.54X_3$$

Q1) 각 변수에 대한 계수의 추정치는 왜 변하는가?

Q2) 어떤 모델을 최종적으로 선택할 것인가?

다중 선형 회귀: Coefficient 검정과 Adjusted R²

- Coefficients 검정: t 검정

1) 절편 검정 $H_0 : \beta_0 = 0$ vs. $H_1 : \beta_0 \neq 0$

➡ p-value 이용

2) 기울기 검정 $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$

➡ p-value 이용

- Adjusted R²

$$R_{adj}^2 = 1 - (1 - R^2) \left(\frac{n-1}{n-k-1} \right)$$

다중 선형 회귀: 모델 검정 (ANOVA표)

- 모델 검정

H_0 : model is not valid vs. H_1 : model is valid

↔ $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$ vs. H_1 : 적어도 하나는 0이 아니다

- ANOVA 표

Source	SS	d.f.	MS	F	P-value
Regression	SSR	K	MSR=SSR/k	MSR/MSE	
Residual	SSE	n-k-1	MSE=SSE/(n-k-1)		
Total	SST	n-1			

- F 검정

$$F = MSR/MSE = (SSR/k)/(SSE/(n-k-1)) \sim F_{k, (n-k-1)}$$

Partial F 검정을 이용한 모델검정

- [H_0 : Reduced model vs H_1 : Full model] 방법 이용

- Partial F 검정 이용

$$\begin{array}{ccc} H_0 : y = \beta_0 + \beta_1 x_1 + \varepsilon & \begin{array}{c} ? \\ \longleftrightarrow \end{array} & H_0 : \beta_2 = 0 \\ H_1 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon & & H_1 : \beta_2 \neq 0 \end{array}$$

```
home=read.table("c:/home.txt", header=T)
names(home)
reduced=lm(Price~SqFt+LotSize, data=home)
full=lm(Price~SqFt+LotSize+Baths, data=home)
anova(reduced, full);
```

AIC 기법을 이용한 독립 변수 자동 선택법

- Reduced model & full model 접근법에 AIC 이용

$$\begin{array}{ccc} H_0 : y = \beta_0 + \beta_1 x_1 + \varepsilon & \begin{array}{c} ? \\ \longleftrightarrow \end{array} & H_0 : \beta_2 = 0 \\ H_1 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon & & H_1 : \beta_2 \neq 0 \end{array}$$

- 종류

1) forward 2) backward 3) stepwise

```
home=read.table("i:/home.txt", header=T)
out=lm(Price~SqFt+LotSize+Baths)
null=lm(Price~1, data=home)
full=lm(Price~SqFt+LotSize+Baths, data=home)
step(null, scope=list(lower=null, upper=full), direction="forward")
step(full, data=Housing, direction="backward")
step(null, scope = list(upper=full), data=Housing, direction="both")
```

교호(interaction)작용

- 이항형태의 독립변수를 포함한 교호작용

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

- 예) Y: 연봉, X₁: 성별 (이항), X₂: 학력 (연속)

(남성이면 X₁ = 1, 여성이면 X₁ = 0)

(모델1) 여성인 경우: $Y = \beta_0 + \beta_2 X_2 + \varepsilon$

(모델2) 남성인 경우: $Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_2 + \varepsilon$

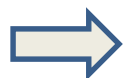
이항변수 형태의 여러 독립변수들 포함

- 이항형태의 독립변수가 아래와 같은 경우: disjoint & collectively exhaustive

이름	1학년	2학년	3학년	4학년
김길동	1	0	0	0
박순신	0	0	1	0
...				...
이태백	0	0	0	1

- 이항형태의 독립변수 가운데 임의의 하나의 변수를 제거한다

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \beta_k X_k + \varepsilon$$



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \varepsilon$$

이항변수 형태의 여러 독립변수들 해석

- 이항형태의 독립변수 해석 방법

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{k-1} X_{k-1} + \varepsilon$$

1) 집단1의 Y 평균: $\beta_0 + \beta_1$

2) 집단2의 Y 평균: $\beta_0 + \beta_2$

.....

3) 집단 k의 Y 평균: β_0

- β_j 의 의미



집단 j의 Y 평균 - 집단 k의 Y 평균

- 이항형태의 독립변수 모델

```
setwd("c:/")  
datay=read.table(file= "election1.txt", header=T)  
out1=lm(choice~Age65+Urban+ColGrad+Union+factor(Area), data=datay)  
summary(out1)
```

절편이 없는 회귀분석

```
out2=lm(choice~-1+Age65+Urban+ColGrad+Union+factor(Area), data=datay)  
summary(out2)
```

R 코드 분석 결과

```
lm(formula = choice ~ Age65 + Urban + ColGrad + Union + factor(Area), data = datay)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	100.22162	8.24377	12.157	2.42e-15 ***
Age65	-1.28688	0.40096	-3.210	0.002549 **
Urban	-0.09827	0.03476	-2.827	0.007158 **
ColGrad	-0.58151	0.19330	-3.008	0.004425 **
Union	-0.72808	0.13274	-5.485	2.18e-06 ***
factor(Area)Neast	-7.27780	1.91986	-3.791	0.000474 ***
factor(Area)Seast	-5.67146	2.03405	-2.788	0.007928 **
factor(Area)West	-1.92319	1.88728	-1.019	0.314025

Residual standard error: 4.287 on 42 degrees of freedom
Multiple R-squared: 0.7925, Adjusted R-squared: 0.758
F-statistic: 22.92 on 7 and 42 DF, p-value: 1.935e-12

```
lm(formula = choice ~ -1 + Age65 + Urban + ColGrad + Union + factor(Area), data = datay)
```

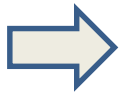
	Estimate	Std. Error	t-value	Pr(> t)
Age65	-1.28688	0.40096	-3.210	0.00255 **
Urban	-0.09827	0.03476	-2.827	0.00716 **
ColGrad	-0.58151	0.19330	-3.008	0.00443 **
Union	-0.72808	0.13274	-5.485	2.18e-06 ***
factor(Area)MidWest	100.22162	8.24377	12.157	2.42e-15 ***
factor(Area)Neast	92.94382	8.86144	10.489	2.65e-13 ***
factor(Area)Seast	94.55016	7.52537	12.564	8.12e-16 ***
factor(Area)West	98.29843	7.56673	12.991	2.64e-16 ***

Residual standard error: 4.287 on 42 degrees of freedom
Multiple R-squared: 0.9941, Adjusted R-squared: 0.993
F-statistic: 885.4 on 8 and 42 DF, p-value: < 2.2e-16

다중공선성

- VIF 이용: 10 이상이면 강한 다중공선성 존재

```
> setwd("c:/")  
> library(car)  
> datay=read.table(file= "election1.txt", header=T)  
> out1=lm(choice~Age65+Urban+ColGrad+Union+factor(Area), data=datay)  
> vif(out1)
```




- 1) 다중공선성이 가장 높은 변수부터 모델에서 제거한 뒤
나머지 변수로 다중 공선성 확인 후 추가 제거여부 결정
- 2) 모델에 반드시 필요한 변수가 다중공선성이 가장 높으면
다음으로 높은 변수를 제거한 뒤 다중공선성을 재확인함

- 새로운 자료로 예측

```
datay=read.table(file= "c:/election1.txt", header=T)
out1=lm(choice~Age65+Urban+ColGrad+Union+factor(Area), data=datay)
str(datay);
#new data
newone=data.frame(Age65=23, Urban=92, ColGrad=78, Union=20, Area="MidWest")
# prediction
predict(out1, new=newone)
```

- Centering of Predictor Variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$


$$y = \beta_0 + \beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_k(x_k - \bar{x}_k) + \varepsilon$$

- 다중 공선성(VIF) 약화 시킴.

- Standardization of Predictor Variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

➡
$$y = \alpha_0 + \alpha_1 \frac{(x_1 - \bar{x}_1)}{s_{x1}} + \alpha_2 \frac{(x_2 - \bar{x}_2)}{s_{x2}} + \cdots + \alpha_k \frac{(x_k - \bar{x}_k)}{s_{xk}} + \varepsilon$$

- Standardization transforms the **scales of variables** into same. Means are all 0, and variances are all 1.
- New **coefficients** $\{\alpha_1, \alpha_2, \dots, \alpha_k\}$ can be used to compare **the impacts of predictor variables** $\{x_1, x_2, \dots, x_k\}$ to **response variable**.

- Standardization of Predictor Variables & Response Variable

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

$$\Rightarrow \frac{(y - \bar{y})}{s_y} = \alpha_0 + \alpha_1 \frac{(x_1 - \bar{x}_1)}{s_{x1}} + \alpha_2 \frac{(x_2 - \bar{x}_2)}{s_{x2}} + \cdots + \alpha_k \frac{(x_k - \bar{x}_k)}{s_{xk}} + \varepsilon$$

$$\text{Here } \alpha_0 = 0, \quad \alpha_i = \text{Corr}(y^*, x_i^*) \times \frac{\sqrt{V(y^*)}}{\sqrt{V(x^*)}} = \text{Corr}(y, x_i)$$

- Note: $\beta_i = \text{Corr}(y, x_i) \times \frac{\sqrt{V(y)}}{\sqrt{V(x)}}$, $y^* = \frac{(y - \bar{y})}{s_y}$, $x_i^* = \frac{(x_i - \bar{x}_i)}{s_{xi}}$

Other Regression Problems

* *Spurious Correlation*

- In a *spurious correlation* two variables appear related because of the way they are defined.
- This problem is called the *size effect* or *problem of totals*.
- Example: see the next page

State	Total Population (millions)	Using Totals		Using Per Capita Data	
		K-12 Spending (\$ billions)	No. of Prisoners (thousands)	K-12 Spending per Capita (\$)	Prisoners per 1,000 Pop.
Alabama	4.447	4.52	24.66	1,016	5.54
Alaska	0.627	1.33	3.95	2,129	6.30
⋮	⋮	⋮	⋮	⋮	⋮
Wisconsin	5.364	8.48	20.42	1,580	3.81
Wyoming	0.494	0.76	1.71	1,543	3.47

Other Regression Problems

