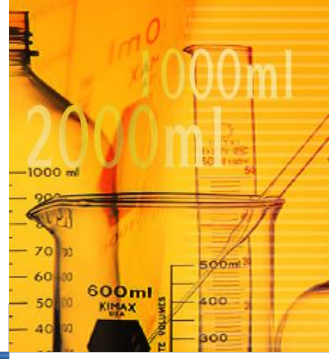
Machine learning



Chapter 8

Regression

Sejong Oh

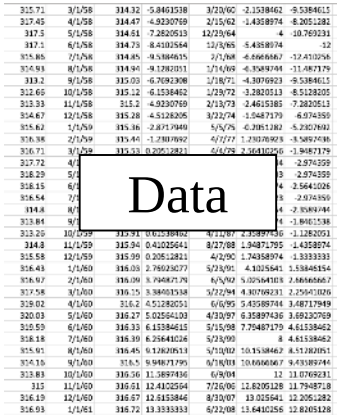
Bio Information Technology Lab.

Contents

- 회귀분석 개요
- 단순회귀분석
- 다중선형회귀

회귀 분석 개요

- Classification

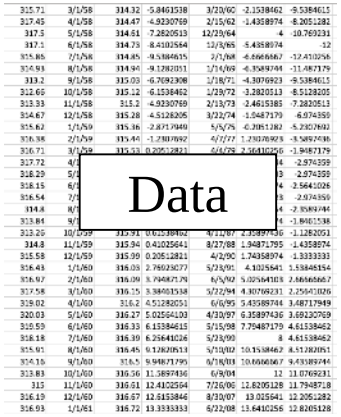


Data

Predict class

Normal/Patient
Pass/Fail

- Regression



Data

Predict value

Degree of pollution
score

회귀 분석 개요

- 회귀 분석
  - 영국의 유전학자 갈턴(F. Galton)이 아버지와 아들의 키에 대한 유전 관계를 통계적으로 조사한 논문에서 regression 처음 사용
  - 연속형 두 변수 사이에 어떤 관계가 있는지 파악
  - 종속 변수가 독립변수에 의해 어떻게 설명 혹은 예측이 되는지를 알아보는 통계적 방법

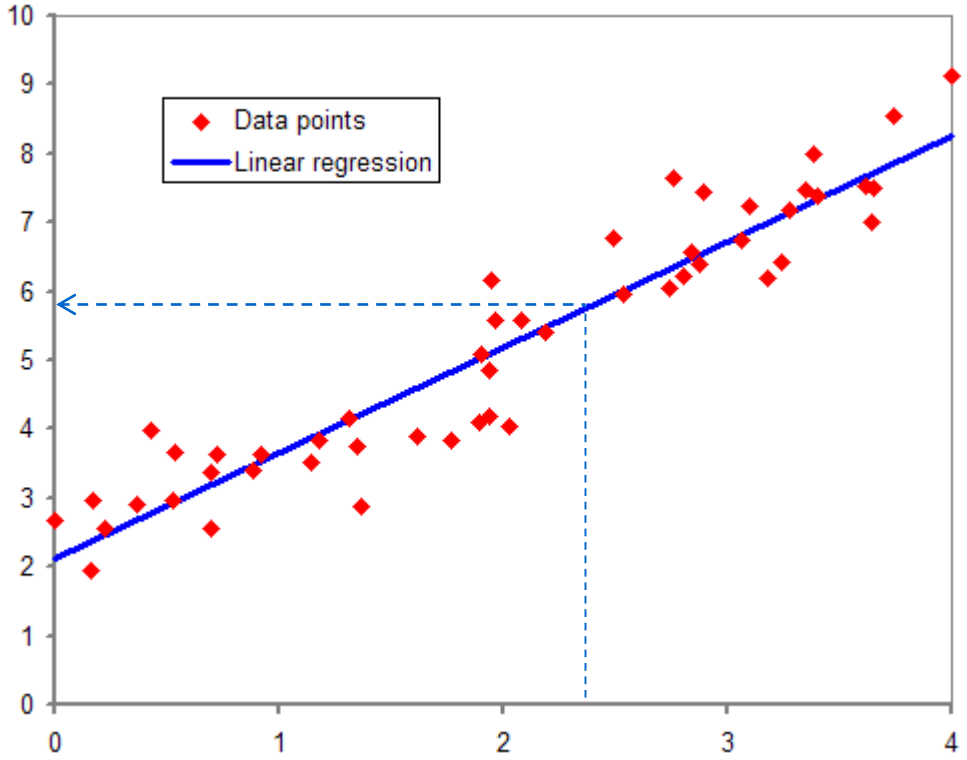
$$y = \alpha + \beta x$$

종속 변수 (dependent variable)
결과변수 (outcome variable)
반응변수 (response variable)

독립변수 (independent variable)
설명변수 (explanatory variable)
예측변수 (predictor variable)
위험인자 (risk factor)

회귀 분석 개요

- 회귀분석의 과정
  - (1) training data 로 부터 회귀식을 구한다
  - (2) 회귀식을 이용하여 y 값을 예측한다



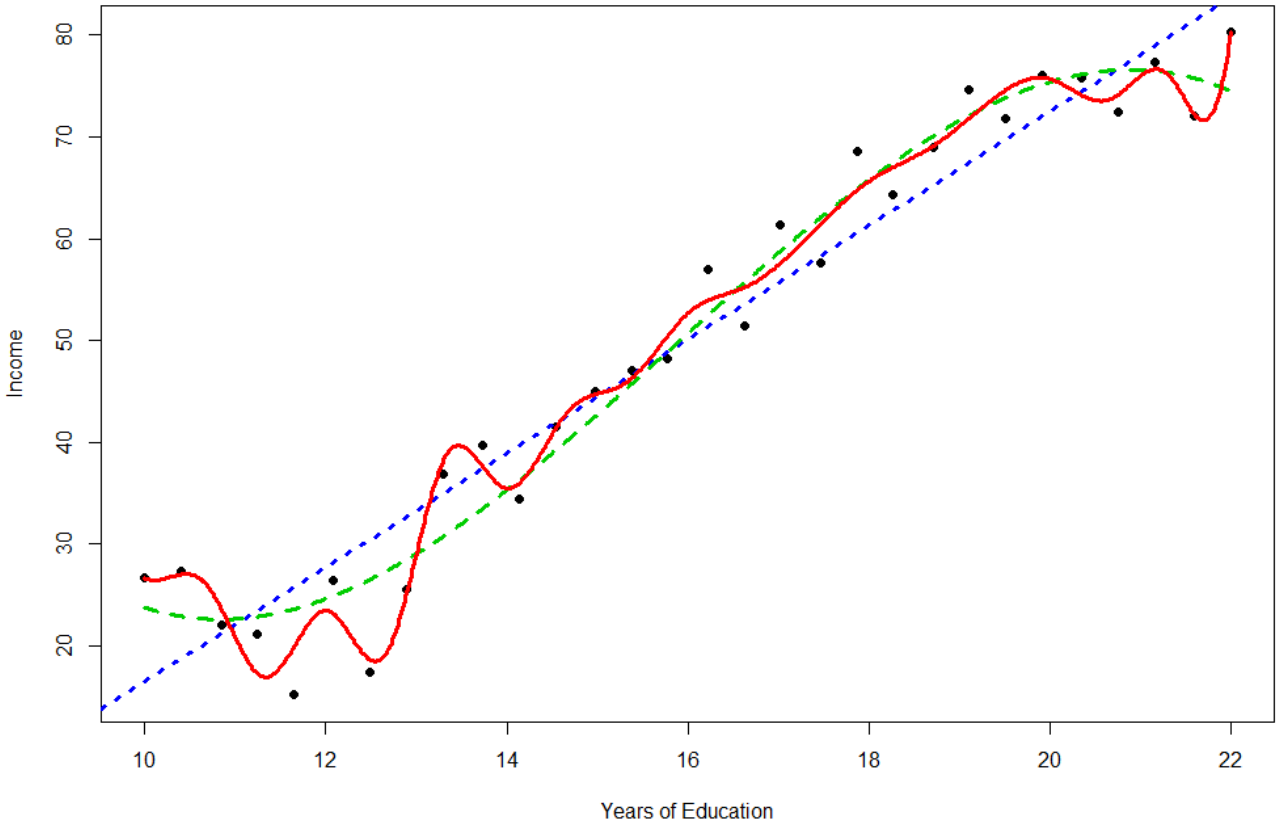
https://ko.wikipedia.org/wiki/회귀분석

회귀 분석 개요

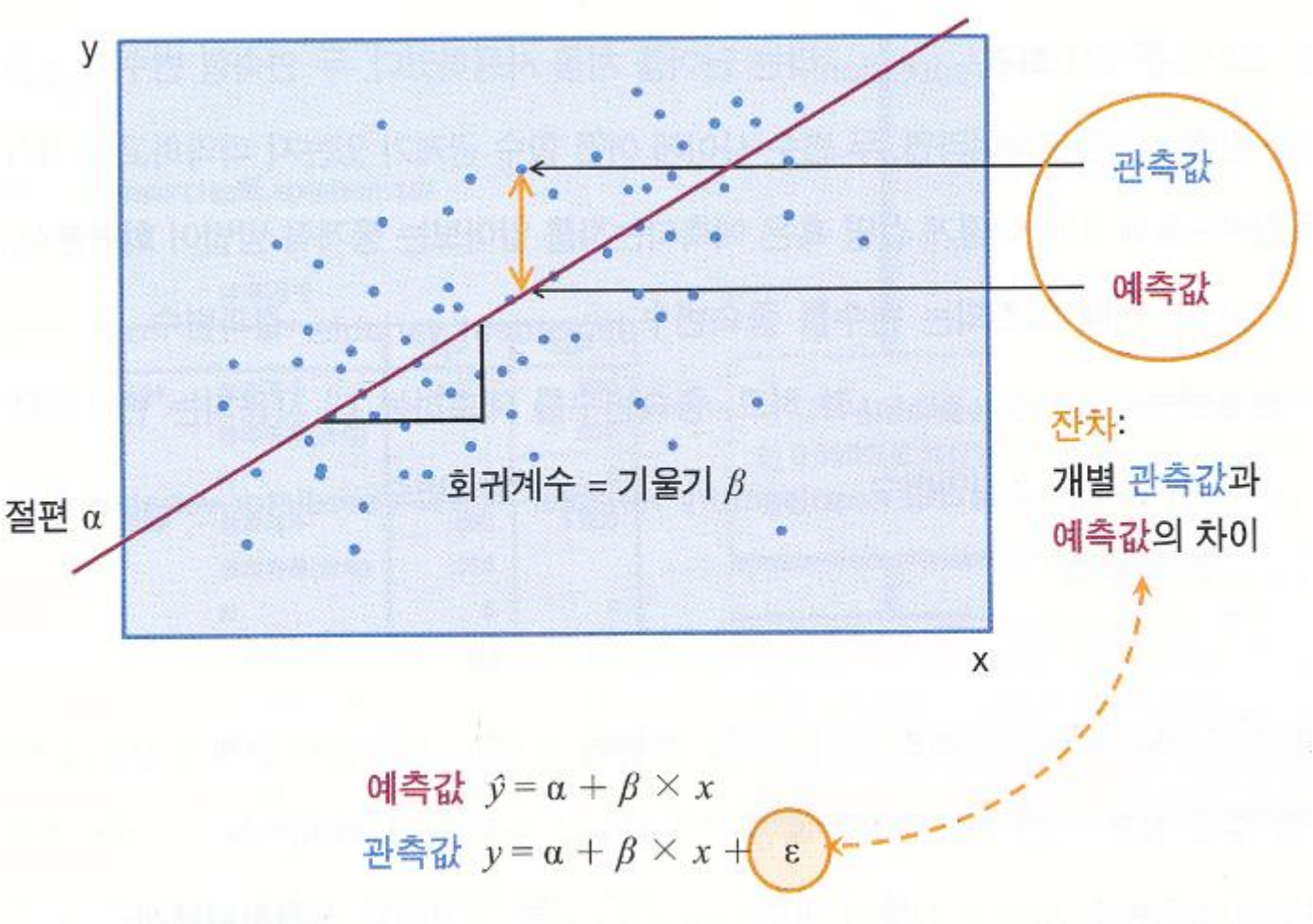
- 회귀 분석
  - 회귀분석은 함수 형태에 따라 1차, 2차, 3차 함수로 분류
  - 보건 의학 연구에서는 주로 1차 함수인 선형 회귀분석 (linear regression)이 주로 이용
  - 선형 회귀식
    - 단순 회귀분석 (simple linear regression) : 두 변수 사이의 선형 관계를 하나의 회귀 직선으로 표현
    - 다중 회귀분석 (multiple linear regression) : 여러 개의 독립변수들에 의해 결과 변수를 예측

회귀 분석 개요

- 예제 : 교육수준과 수입과의 관계
  - $\alpha + \beta X$
  - $\alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$
  - $\alpha + g(X)$



단순 회귀분석


y
x
관측값
예측값
잔차:
개별 관측값과
예측값의 차이
회귀계수 = 기울기 β
절편 α
예측값 $\hat{y} = \alpha + \beta \times x$
관측값 $y = \alpha + \beta \times x + \varepsilon$

단순 회귀분석

- 독립변수(x) 와 종속변수(y) 가 선형 관계에 있을 때 $y = \alpha + \beta x$ 의 선형 회귀식을 추정하는 것이 목적
- α, β : **회귀계수**(regression coefficient). β 만 회귀계수라고 부르기도 함
- α : **절편**(intercept), β :**기울기**(slope)
- 회귀분석에서는 기울기에 관심
- 회귀식은 현상을 설명하는 하나의 모형이라는 의미에서 **회귀 모형** (regression model) 이라고 부르기도 한다
- **잔차/오차항** (residual error): 표본집단/모집단에서 관측값과 회귀선에 의한 예측 값의 차이. (회귀식 또는 회귀 모형에 의해 설명이 되지 않는 부분) .
  - 회귀분석이 적절히 수행되었는지를 판단하는 근거

- 단순회귀 분석은 다중회귀분석, 로지스틱 회귀분석, Cox의 비례 위험 모형 등으로 확장됨

단순 회귀분석

- 회귀 모형에 대한 기본 가정
  - 독립변수(x)와 종속변수(y)가 선형 관계에 있어야 한다. (선형성)
  - 모든 독립변수(x)의 값에서 종속변수(y) 는 정규분포를 이룰 것 (오차항의 정규성)
  - 잔차들은 서로 영향을 받지 않고 독립적으로 분포 할 것 (오차항의 독립성)
  - 모든 독립변수(x)의 값에서 종속변수(y)의 분산은 같을 것 (오차항의 등분산성)

  - 회귀분석시 회귀선의 추정및 검정 뿐만 아니라 위의 네가지 기본 가정이 잘 충족되고 있는지 살펴 보아야 한다

단순 회귀분석

- 회귀 모형에 대한 기본 가정

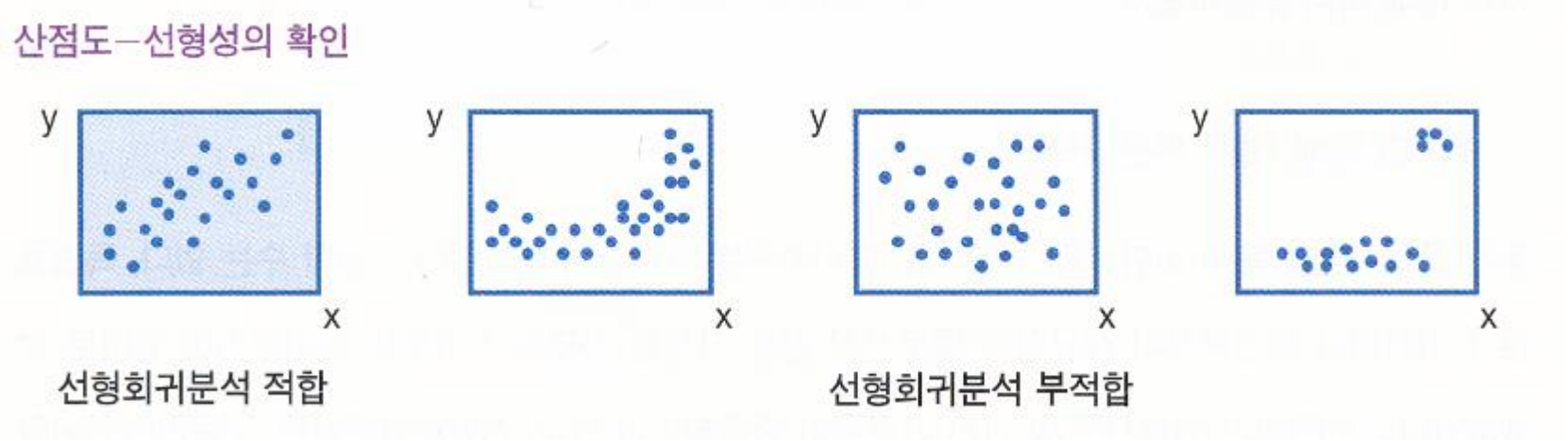
1	선형성	• 독립변수(x)와 종속변수(y)의 관계는 선형 관계에 있다. • 산점도로 확인
2	오차항의 정규성	• 모든 독립변수(x)의 값에서 종속변수(y)는 정규 분포를 이룬다. • 정규 P-P 곡선으로 확인
3	오차항의 독립성	• 개별 잔차들은 서로 독립이다. • 잔차산점도, Durbin-Watson 통계량으로 확인
4	오차항의 등분산성	• 모든 독립변수(x)의 값에서 종속변수(y)의 분산은 같다. • 잔차산점도로 확인

단순 회귀분석

- 회귀 모형에 대한 기본 가정

  1) 선형성

  - 통계 프로그램은 선형관계가 없는 두 변수 사이에서도 회귀식을 추정해 낸다.
  - 산점도를 통해 두 변수사이에 실제 선형 관계가 있는 지 눈으로 확인하는 것이 필요

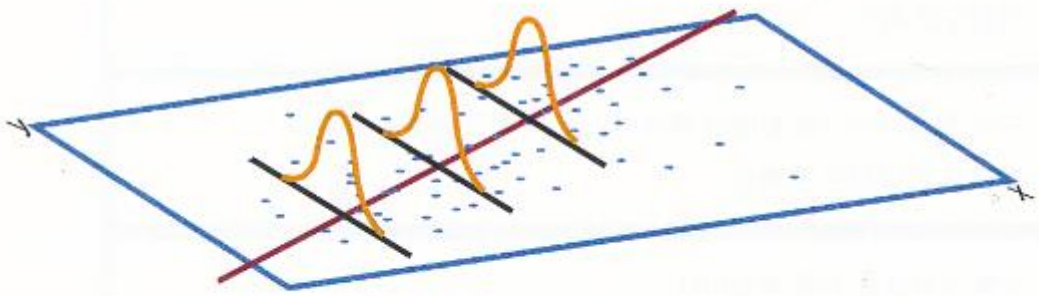


단순 회귀분석

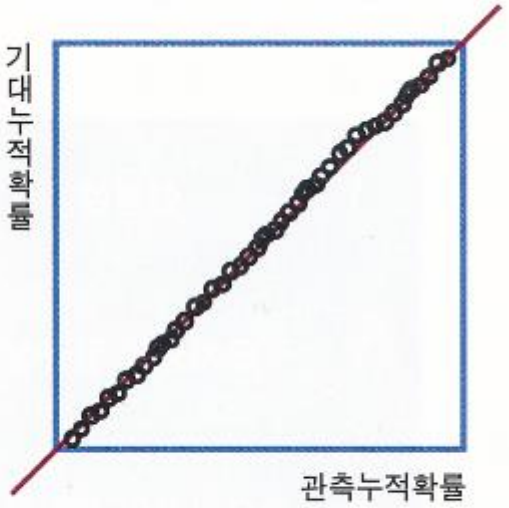
- 회귀 모형에 대한 기본 가정

  2) 오차항의 정규성

  - 회귀선을 그었을 때 모든 독립변수(x)의 값에서 종속변수(y) 의 관측 값은 정규분포를 이룰 것
  - **정규 p-p 곡선**을 그려 눈으로 확인 가능



오차항의 정규성에 대한 3차원 모식도



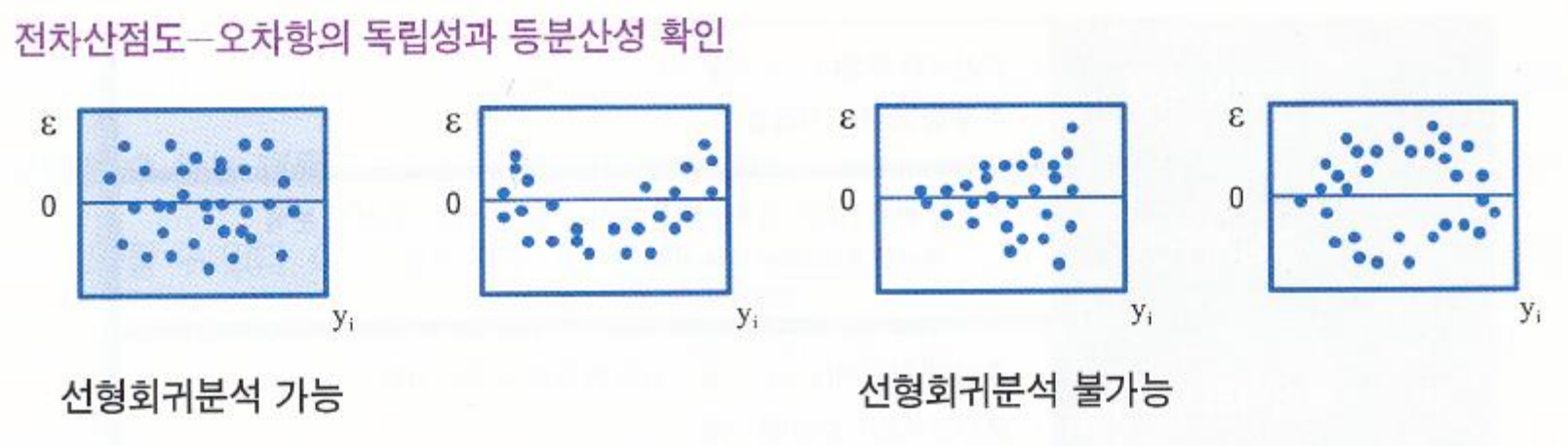
정규 p-p 도표

단순 회귀분석

- 회귀 모형에 대한 기본 가정

  3) 오차항의 독립성

  - 잔차들은 서로 연관성이 없이 독립적이어야
  - **잔차 산점도**를 그렸을 때 모든 예측값에 대해 잔차들이 불규칙성을 보여야 한다
  - 만일 어떤 규칙성이 관찰되면 이에 대한 설명이 필요

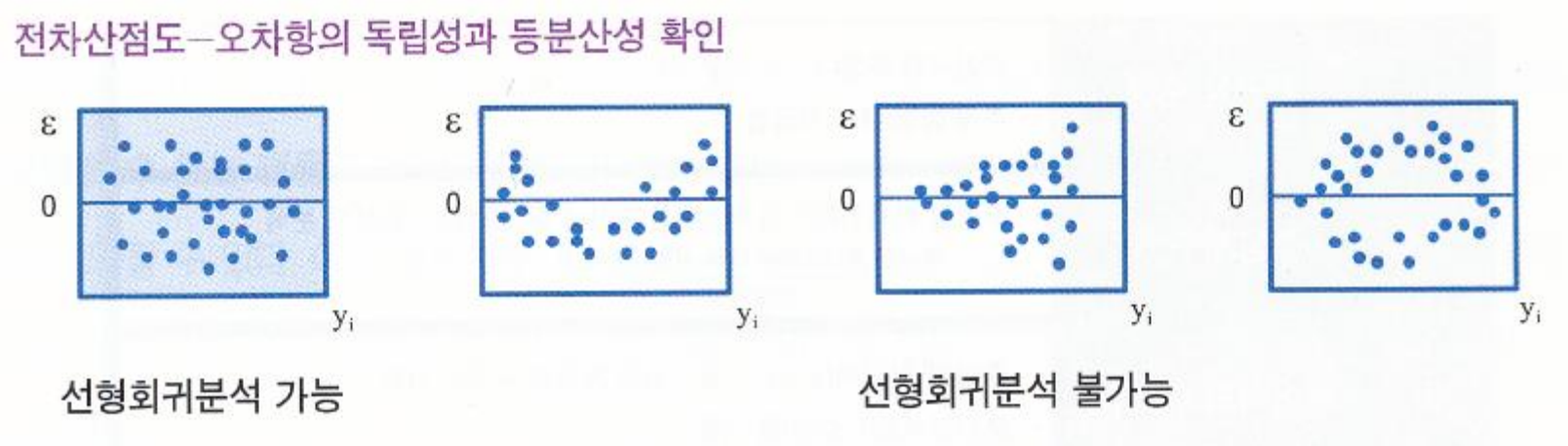


단순 회귀분석

- 회귀 모형에 대한 기본 가정

  4) 오차항의 등분산성

  - 모든 독립변수(x) 에 대해 잔차의 분산이 같다는 의미
  - 예를 들어 x 가 커질수록 분산이 커진다면 등분산성 위배
  - 산점도(잔차 산점도)를 통해 확인 가능

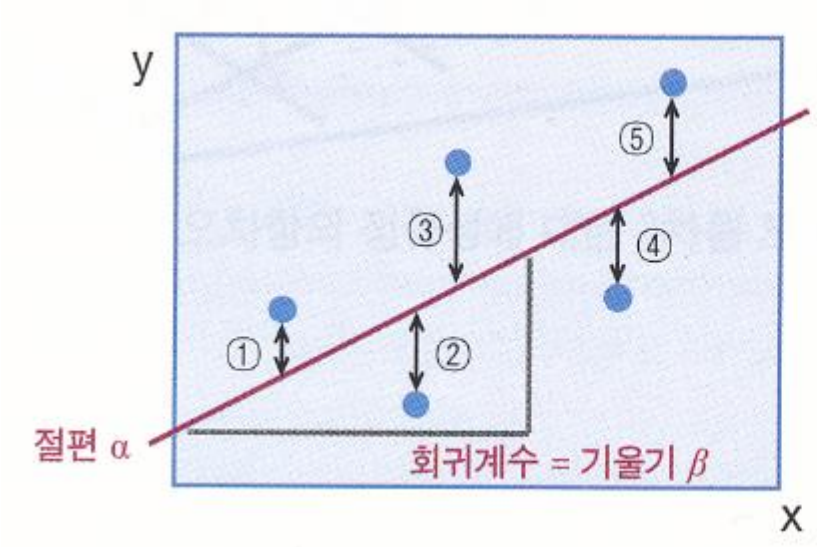


단순 회귀분석

- 단순 회귀 분석의 시행

  1) 회귀식($y = \alpha + \beta x$)의 추정

  - **최소제곱법**(least square method) : 모든 잔차의 제곱의 합을 최소화 하는 직선을 결정
  - 추정 직선에서 관측치들이 떨어진 정도가 작을 수록 회귀식의 자료에 대한 적합도가 더 높다



잔차의 제곱합을 최소화하는

$(①^2 + ②^2 + ③^2 + ④^2 + ⑤^2)$

회귀식을 추정

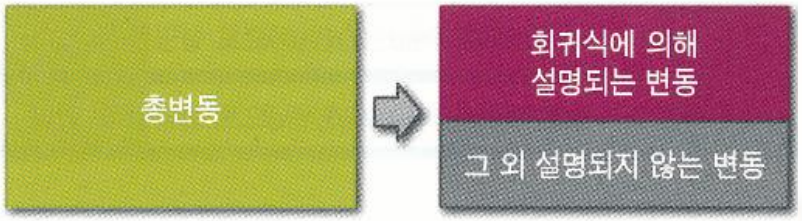
즉, 절편 α와 회귀계수 β를 추정

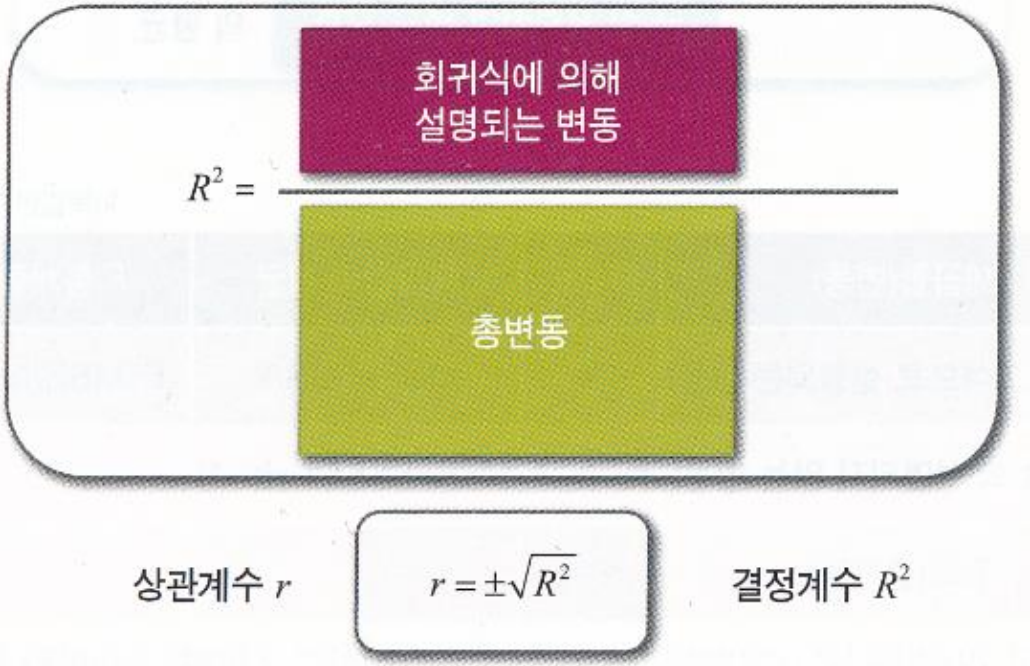
단순 회귀분석

- 단순 회귀 분석의 시행

  2) 결정계수 R^2

  - 회귀식에서 y 의 총변동은 이 회귀식으로 설명될수 있는 변동과 그 외 설명되지 않는 변동으로 설명
  - **결정계수**(coefficient of determination) : 총변동중 회귀식에 의해 설명되는 변동
  - $0 \le R^2 \le 1$. 두 변수 사이의 관계가 강할 수록 R^2 은 1 에 근접





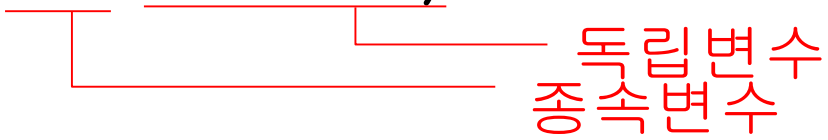
[실습] 단순회귀 분석

- 다음은 322명의 건강한 성인의 건강검진 자료이다.
- 허리둘레로 BMI 를 예측하는 회귀식을 구하라

환자번호	나이	키	몸무게	허리둘레	BMI	수축기 혈압
	age	**height**	**weight**	**waistline**	**BMI**	**SBP**

```
> # 자료 입력및 관찰
> ds = read.csv("6_correlation_and_regression.csv")
> attach(ds)
> head(ds)
  age height weight waistline  BMI SBP
1  57    164   62.0         85 23.1 147
2  25    172   54.0         65 18.3 116
3  57    157   59.0         83 23.9 122
4  43    170   87.8        104 30.4 130
5  52    155   50.0         83 20.8 120
6  27    163   76.0         83 28.6 128
>
```

```
> # 단순회귀분석 : BMI 와 waistline
> result = lm(BMI~waistline)
```



```
> summary(result)

Call:
lm(formula = BMI ~ waistline)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5197 -1.3087 -0.0444  1.2532 12.0683

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.97611    1.25481  -3.169  0.00168 **
waistline    0.33135    0.01386  23.906  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.186 on 320 degrees of freedom
Multiple R-squared:  0.6411,    Adjusted R-squared:  0.6399
F-statistic: 571.5 on 1 and 320 DF,  p-value: < 2.2e-16
```

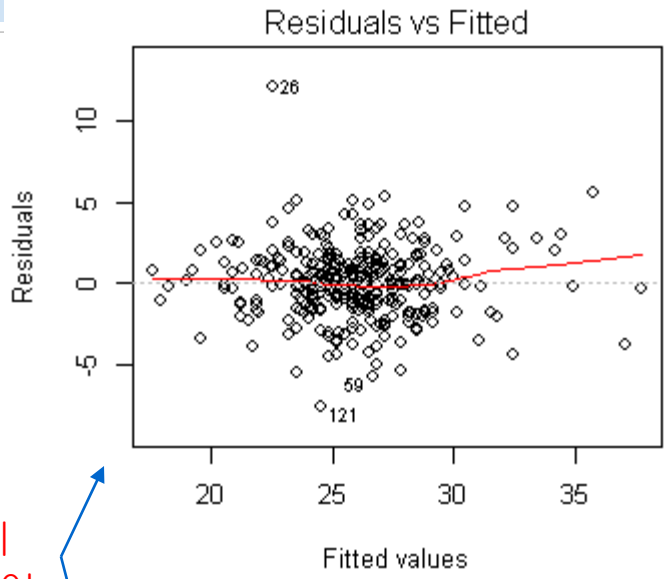
회귀식 : **BMI = -3.97611 + 0.33135*waistline**

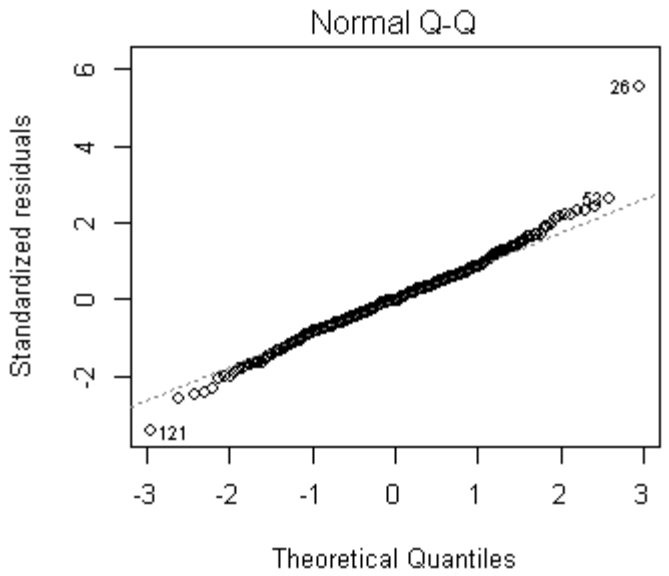
```
> par(mfrow=c(2,2))
> plot(result)

> detach(ds)
```

회귀식을 이용하여 허리둘레가 90 일 때 BMI 값을 예측해 보자

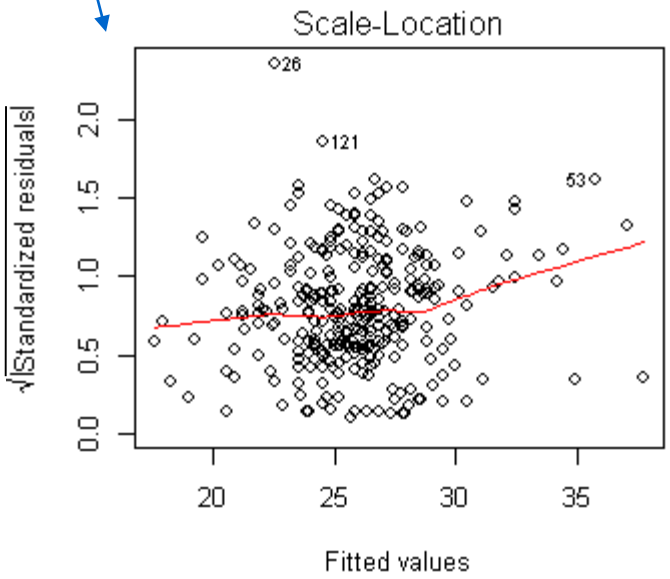
(잔차)

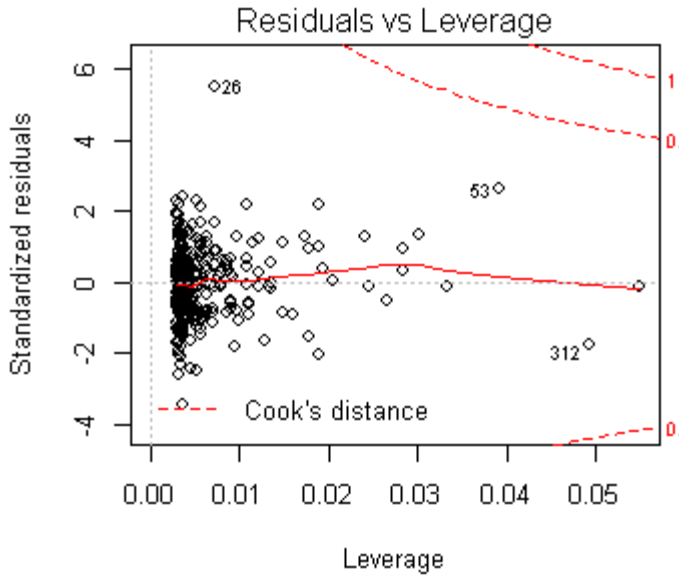




기준선에 놓여 있을 때 정규성 만족

Random 하게 분포하는 것이 바람직





점들이 한쪽에 몰려 있는 것이 바람직 (모델이 잘적합됨)

[과제 1]

- 실습의 데이터를 이용하여 키로 몸무게를 예측하는 회귀식을 작성하시오

- 이 회귀식을 이용하여 경우 키 180 인 경우 몸무게가 얼마인지 예측하시오

다중 선형회귀

- 대부분 실제 데이터 분석에서는 하나 이상의 변수가 관여
- 이 경우 다중 선형 회귀를 이용

- 단순 선형회귀
  - $Y = \alpha + \beta X$

- 다중 선형회귀
  - $Y = \alpha + \beta_1 X +$

age	height	weight	waistline	BMI	SBP
57	164	62	85	23.1	147
25	172	54	65	18.3	116
57	157	59	83	23.9	122
43	170	87.8	104	30.4	130

다중 선형회귀

- 장점
  - 수치 데이터를 모델화 하기 위한 가장 일반적인 접근법
  - 거의 모든 데이터를 모델화 할수 있다
  - 속성(변수)과 결과간의 견고성과 크기를 추정할 수 있다.

- 단점
  - 데이터에 대한 강한 가정을 만든다
  - 결측치가 포함된 경우 잘 작동되지 않는다
  - 수치 속성만 작동하고 범주형 데이터는 부가 처리가 필요하다
  - 모델을 이해하려면 통계적 지식이 필요하다

다중 선형회귀

```
> data(state.x77)
> head(state.x77)
```

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
California	21198	5114	1.1	71.71	10.3	62.6	20	156361
Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766

```
> states <-
as.data.frame(state.x77[,c("Murder","Population","Illite
racy","Income","Frost")])
```

Population(인구), Income(수입), Illiteracy(문맹률), Frost(추운날수) 로 부터 Murder(살인율)을 예측하는 모델을 만들어 보자

다중 선형회귀

```
# model fitting
> fit <- lm(Murder~.,data = states)
> fit
```

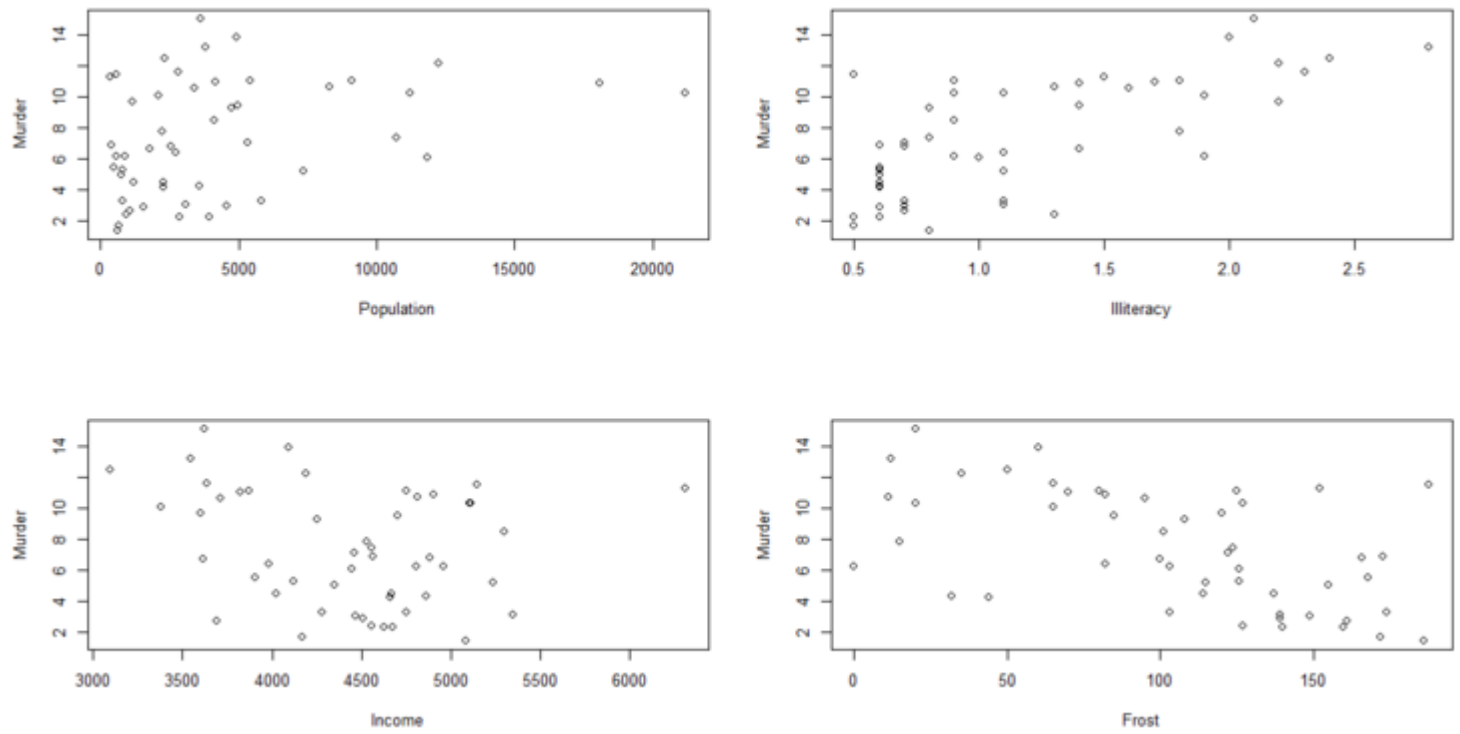
Murder~Population+Illiteracy+Income+Frost

```
Call:
lm(formula = Murder ~ ., data = states)

Coefficients:
(Intercept)   Population   Illiteracy       Income        Frost
  1.235e+00    2.237e-04    4.143e+00    6.442e-05    5.813e-04
```

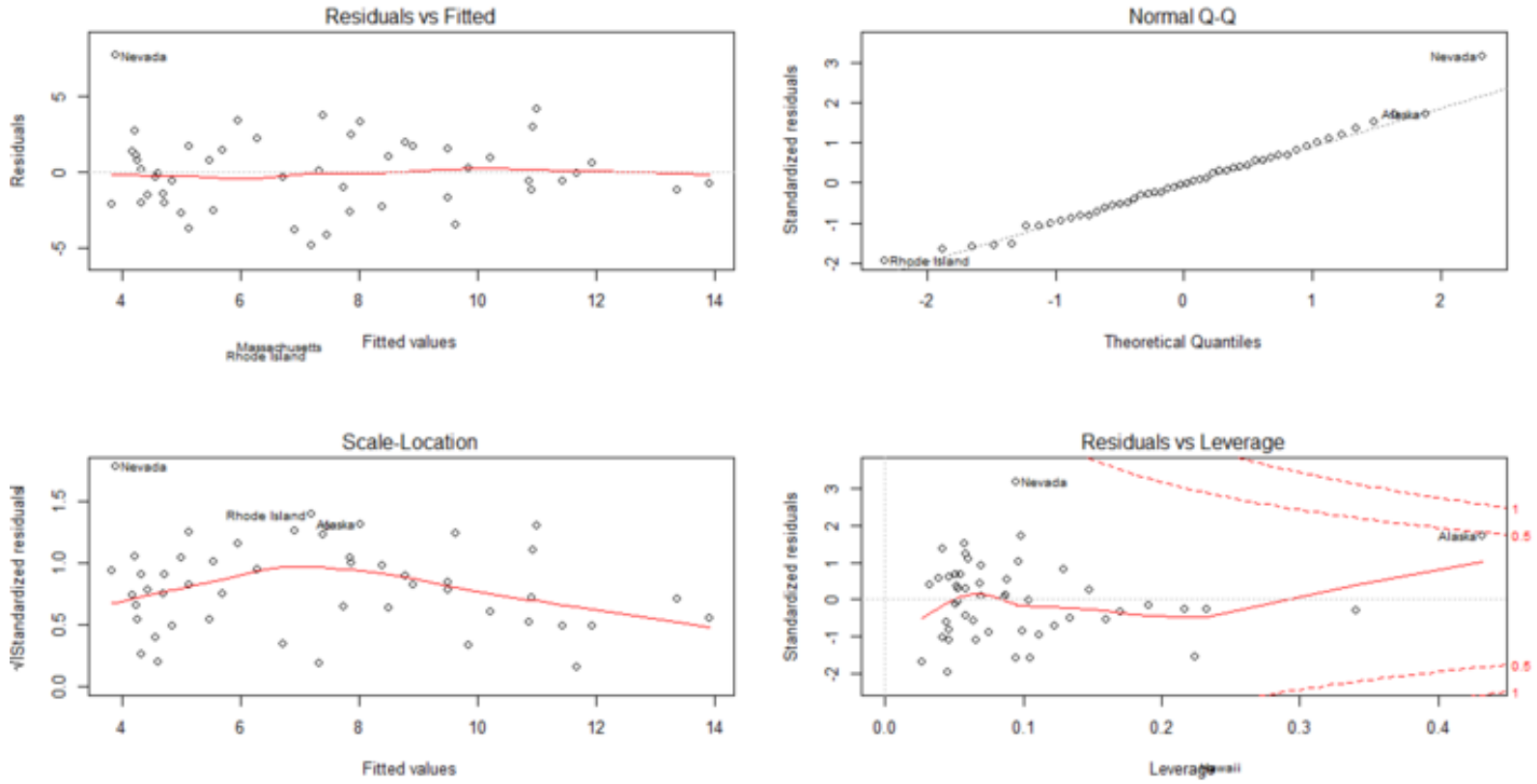
다중 선형회귀

```
# 데이터에 대한 산점도 그리기
> plot(Murder~.,data = states)
```



다중 선형회귀

```
# 회귀 모형의 적절성 확인
> plot(fit)
```



선형성과 정규성, 등분산성 모두 만족하는 것으로 보임

다중 선형회귀

```
# 네개의 독립변수에 대한 유의성 확인
> summary(fit)
Call:
lm(formula = Murder ~ ., data = states)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7960 -1.6495 -0.0811  1.4815  7.6210

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.235e+00  3.866e+00   0.319   0.7510
Population  2.237e-04  9.052e-05   2.471   0.0173 *
Illiteracy  4.143e+00  8.744e-01   4.738 2.19e-05 ***
Income      6.442e-05  6.837e-04   0.094   0.9253
Frost       5.813e-04  1.005e-02   0.058   0.9541
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.535 on 45 degrees of freedom
Multiple R-squared:  0.567,     Adjusted R-squared:  0.5285
F-statistic: 14.73 on 4 and 45 DF,  p-value: 9.133e-08
```

다중 선형회귀

```
# 유의한 두 변수만 가지고 회귀모형을 다시 구해 봄 (사실은 이렇게 하면 안됨)
> fit2 <- lm(Murder~Population + Illiteracy, data = states)
> summary(fit2)

Call:
lm(formula = Murder ~ Population + Illiteracy, data = states)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7652 -1.6561 -0.0898  1.4570  7.6758

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.652e+00  8.101e-01   2.039  0.04713 *
Population  2.242e-04  7.984e-05   2.808  0.00724 **
Illiteracy  4.081e+00  5.848e-01   6.978 8.83e-09 ***
```

다중 선형회귀

- 회귀식의 도출

```
Murder = 1.652 + 0.0002242*Population
                + 4.081*Illiteracy
```

- 예측값 과 실제 값의 비교

Population	Illiteracy	Murder(예측)	Murder(실제)	오차(실제-예측)
3615	2.1	11.03	15.1	4.07
365	1.5	7.86	11.3	3.44
2212	1.8	9.49	7.8	-1.69
2110	1.9	9.88	10.1	0.22
21198	1.1	10.89	10.3	-0.59
2541	0.7	5.08	6.8	1.72

다중 공선성(Multicollinearity)

- 다중 공선성 : 다중회귀분석에서 x변수(설명변수, 독립변수)들 끼리 상관관계가 존재할 경우 회귀 계수의 분산을 크게 하여, 회귀분석시 추정 회귀 계수를 믿을 수 없게 되는 문제.

- 다중 공선성을 고려하여 x변수를 선정하고 회귀식을 만들어야 함 (자동화된 tool 이 있음)

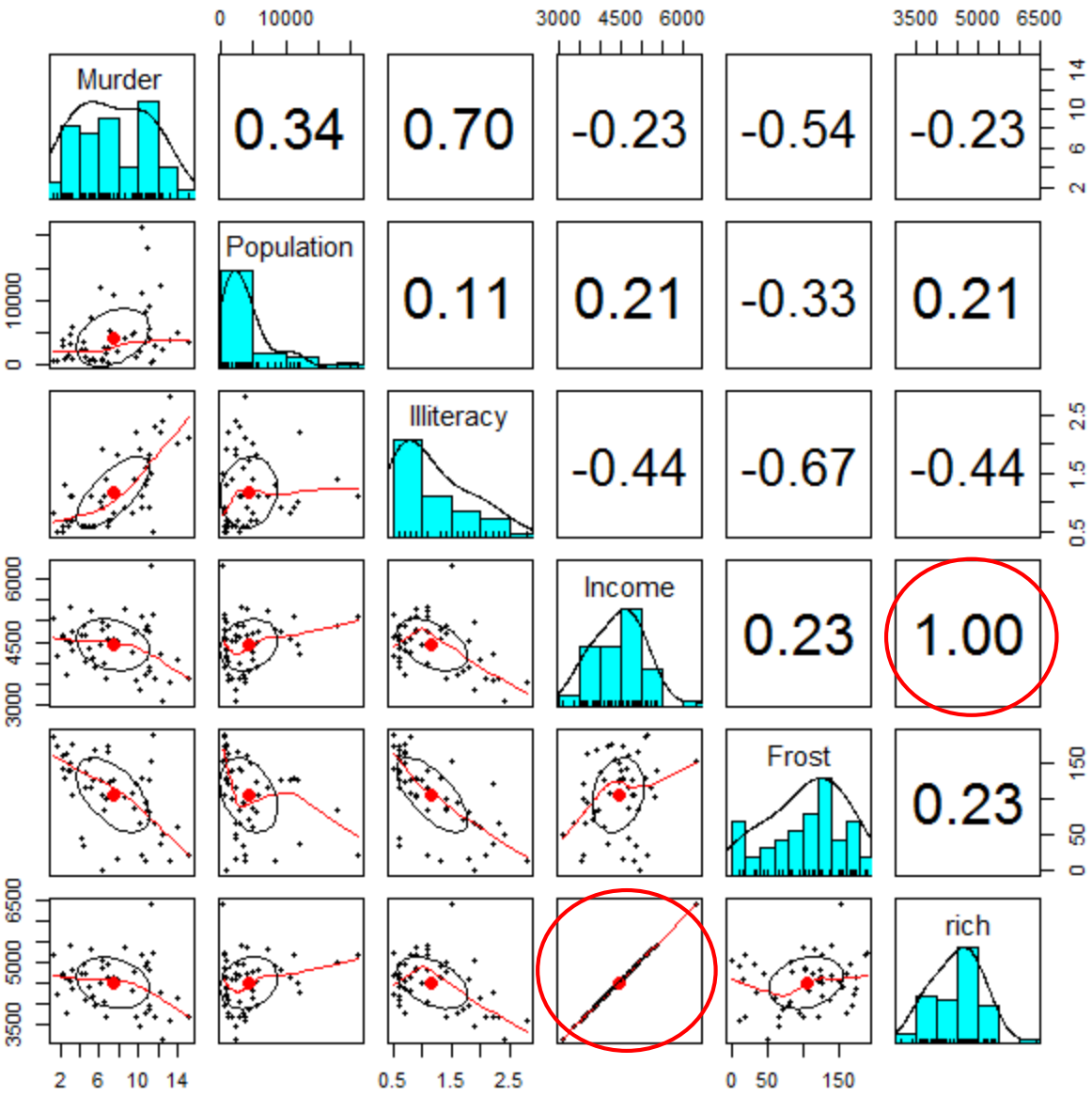
- 다중 공선성을 확인하기 위해 상관관계가 있는 변수를 하나 만들어 넣는다.

```
> rich <- sqrt(states$income) +  states$Income
> tmp <- cbind(states,rich)
<- as.factor(tmp$rich)
head(tmp)
```

다중 공선성(Multicollinearity)

```
# 산점도를 그려 다중 공선성 확인
> library(psych)
> library(car)
> pairs.panels(tmp[names(tmp)])
```

그래프에서 숫자는 상관관계를 나타내며, -1에 가까우면 음의 상관관계, 1에 가까우면 양의 상관관계.
그래프에서 1, 0.7과 같이 1에 가까운 수가 있으므로 다중공선성이 의심


Murder
Population
Illiteracy
Income
Frost
rich
0.34
0.70
-0.23
-0.54
-0.23
0.11
0.21
-0.33
0.21
-0.44
-0.67
-0.44
0.23
1.00
0.23
0 10000
3000 4500 6000
3500 5000 6500
2 6 10 14
0.5 1.5 2.5
0 50 150

다중 공선성(Multicollinearity)

```
# 수치로 다중 공선성 확인
> fit <- lm(Murder~.,data = tmp)
> car::vif(fit)
  Population   Illiteracy       Income        Frost          rich
1.317121e+00 3.422561e+00 1.221961e+07 2.278193e+00 1.222273e+07
> sqrt(car::vif(fit)) > 2
Population Illiteracy     Income      Frost       rich
     FALSE      FALSE       TRUE      FALSE       TRUE
```

car::vif() : car 패키지에 있는 vif() 함수를 사용.
동일한 이름의 함수가 여러 패키지에 있을 때 충돌을 피하기 위함
vif :분산팽창요인(Variance Inflation Factor)

sqrt(car::vif(fit)) 가 2보다 크면 다중공선성이 있는 것으로 판단
(느슨하게는 **car::vif(fit)** 가 10 보다 크면 다중공선성 판단)
위에서는 Income, rich 가 TRUE 이므로 다중공선성이 있다고 판단

다중 공선성(Multicollinearity)

- 다중 공선성의 해결 방법
  - 상관관계가 높은 독립변수중 하나 혹은 일부를 제거한다.
  - 변수를 변형시키거나 새로운 관측치를 이용한다.
  - 자료를 수집하는 현장의 상황을 보아 상관관계의 이유를 파악하여 해결한다.

다중 선형회귀에서 변수의 선택 (feature selection)

- 주어진 종속변수 y 에 대해 이를 잘 설명할 수 있는 독립변수들을 선별하는 작업이 필요

- 다음의 함수를 이용하여 이 작업을 자동적으로 실시할수 있다
  - 전진 선택(forward)
  - 후진 소거(backward)
  - 단계별 선택(stepwise) : 전진선택, 후진 소거를 동시에 적용

다중 선형회귀에서 변수의 선택 (feature selection)

```
## forward search ----------------------------
# 먼저 상수항만 포함된 회귀모형을 생성 (선택된 변수 없음)
> fit.con <- lm(Murder~1,data=tmp)
# 전진선택
> fit.forward <-
step(fit.con,scope=list(lower=fit.con,upper=fit),
                        direction = "forward")
> fit.forward
Call:
lm(formula = Murder ~ Illiteracy + Population, data = tmp)

Coefficients:
(Intercept)    Illiteracy    Population
  1.6515497     4.0807366     0.0002242
```

```
> fit.forward <- step(fit.con,scope=list(lower=fit.con,upper=fit),
+                     direction = "forward")
Start:  AIC=131.59
Murder ~ 1

             Df Sum of Sq    RSS     AIC
+ Illiteracy  1    329.98 337.76  99.516
+ Frost       1    193.91 473.84 116.442
+ Population  1     78.85 588.89 127.311
+ rich        1     35.39 632.35 130.871
+ Income      1     35.35 632.40 130.875
<none>                    667.75 131.594

Step:  AIC=99.52
Murder ~ Illiteracy

             Df Sum of Sq    RSS     AIC
+ Population  1    48.517 289.25  93.763
<none>                    337.76  99.516
+ Frost       1     5.387 332.38 100.712
+ Income      1     4.916 332.85 100.783
+ rich        1     4.914 332.85 100.783

Step:  AIC=93.76
Murder ~ Illiteracy + Population

         Df Sum of Sq    RSS    AIC
<none>                289.25 93.763
+ Income  1  0.057022 289.19 95.753
+ rich    1  0.056346 289.19 95.753
+ Frost   1  0.021447 289.22 95.759
```

AIC 값이 가장 작은 변수를 선택

다중 선형회귀에서 변수의 선택 (feature selection)

```
## backward elimination ----------------------------
> fit.backward <-
step(fit,scope=list(lower=fit.con,upper=fit),
                    direction = "backward")
> fit.backward
Call:
lm(formula = Murder ~ Population + Illiteracy + Income + rich,
    data = tmp)
Coefficients:
(Intercept)   Population   Illiteracy       Income         rich
  108.40286      0.00027      3.34423      3.17920     -3.15585
> summary(fit.backward)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.084e+02  6.447e+01   1.681 0.099613 .
Population   2.700e-04  8.755e-05   3.084 0.003486 **
Illiteracy   3.344e+00  8.036e-01   4.162 0.000141 ***
Income       3.179e+00  1.912e+00   1.663 0.103296
rich        -3.156e+00  1.898e+00  -1.663 0.103302
```

다중 선형회귀에서 변수의 선택 (feature selection)

```
> fit.backward <- step(fit,scope=list(lower=fit.con,upper=fit),
+                      direction = "backward")
Start:  AIC=96.55
Murder ~ Population + Illiteracy + Income + Frost + rich

             Df Sum of Sq    RSS     AIC
- Frost       1     1.212 272.45  94.771
<none>                    271.24  96.549
- rich        1    17.930 289.17  97.749
- Income      1    17.931 289.17  97.749
- Illiteracy  1    48.844 320.08 102.828
- Population  1    50.123 321.36 103.027

Step:  AIC=94.77
Murder ~ Population + Illiteracy + Income + rich

             Df Sum of Sq    RSS     AIC
<none>                    272.45  94.771
- rich        1    16.740 289.19  95.753
- Income      1    16.741 289.19  95.753
- Population  1    57.580 330.03 102.358
- Illiteracy  1   104.852 377.30 109.051
```

다중 선형회귀에서 변수의 선택 (feature selection)

```
## stepwise ----------------------------
> fit.both <-
step(fit.con,scope=list(lower=fit.con,upper=fit),
direction = "both")
> fit.both
Call:
lm(formula = Murder ~ Illiteracy + Population, data = tmp)

Coefficients:
(Intercept)    Illiteracy    Population
  1.6515497     4.0807366     0.0002242
> summary(fit.both)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.652e+00  8.101e-01   2.039  0.04713 *
Illiteracy  4.081e+00  5.848e-01   6.978 8.83e-09 ***
Population  2.242e-04  7.984e-05   2.808  0.00724 **
```

```
> fit.both <- step(fit.con,scope=list(lower=fit.con,upper=fit), direction = $
Start:  AIC=131.59
Murder ~ 1

             Df Sum of Sq    RSS     AIC
+ Illiteracy  1    329.98 337.76  99.516
+ Frost       1    193.91 473.84 116.442
+ Population  1     78.85 588.89 127.311
+ rich        1     35.39 632.35 130.871
+ Income      1     35.35 632.40 130.875
<none>                    667.75 131.594

Step:  AIC=99.52
Murder ~ Illiteracy

             Df Sum of Sq    RSS     AIC
+ Population  1     48.52 289.25  93.763
<none>                    337.76  99.516
+ Frost       1      5.39 332.38 100.712
+ Income      1      4.92 332.85 100.783
+ rich        1      4.91 332.85 100.783
- Illiteracy  1    329.98 667.75 131.594

Step:  AIC=93.76
Murder ~ Illiteracy + Population

             Df Sum of Sq    RSS     AIC
<none>                    289.25  93.763
+ Income      1     0.057 289.19  95.753
+ rich        1     0.056 289.19  95.753
+ Frost       1     0.021 289.22  95.759
- Population  1    48.517 337.76  99.516
- Illiteracy  1   299.646 588.89 127.311
```

회귀식에 의한 종속변수 예측

- fit.both 모델을 가지고 Murder 를 예측해 보자

```
## 점추정 ----------------------------
> pre_murder <- predict(fit.both, newdata = tmp)
> pre_murder <- as.data.frame(pre_murder)
> pre_murder
               pre_murder
Alabama         11.031529
Alaska           7.854483
Arizona          9.492775
Arkansas         9.877982
California      10.892657
Colorado         5.077722
Connecticut      6.835337
```

회귀식에 의한 종속변수 예측

```
## 구간 추정 ----------------------------
> pre_murder <- predict(fit.both, newdata = tmp,
interval = "confidence")
> pre_murder <- as.data.frame(pre_murder)
> pre_murder
                   fit       lwr       upr
Alabama       11.031529  9.716444 12.346615
Alaska         7.854483  6.810613  8.898352
Arizona        9.492775  8.394405 10.591146
Arkansas       9.877982  8.687628 11.068336
California    10.892657  8.070037 13.715277
Colorado       5.077722  4.157763  5.997681
Connecticut    6.835337  6.103519  7.567155
Delaware       5.454016  4.502294  6.405739
Florida        8.812096  7.853265  9.770927
```

회귀식에 의한 종속변수 예측

```
## 추정치와 실제값 비교 ----------------------------
> ttmp <- cbind(pre_murder,tmp$Murder)
> ttmp
                fit       lwr       upr tmp$Murder
Alabama    11.031529  9.716444 12.346615       15.1
Alaska      7.854483  6.810613  8.898352       11.3
Arizona     9.492775  8.394405 10.591146        7.8
Arkansas    9.877982  8.687628 11.068336       10.1
California 10.892657  8.070037 13.715277       10.3
Colorado    5.077722  4.157763  5.997681        6.8
Connecticut 6.835337  6.103519  7.567155        3.1
Delaware    5.454016  4.502294  6.405739        6.2
Florida     8.812096  7.853265  9.770927       10.7
```

[과제 2]

- 6_correlation_and_regression.csv 를 가지고 BMI 를 예측할 수 있는 회귀식을 만들어 보자

age	height	weight	waistline	BMI	SBP
57	164	62	85	23.1	147
25	172	54	65	18.3	116
57	157	59	83	23.9	122
43	170	87.8	104	30.4	130

  - BMI 예측에 도움이 되는 변수를 선택한다
  - 회귀 모델을 이용하여 BMI 를 예측해보자
  - 예측한 값과 실제 BMI 값의 차이를 비교해 보자