Excel 및 R : 단순회귀분석

I. Excel 단순회귀분석

II. R 단순회귀분석

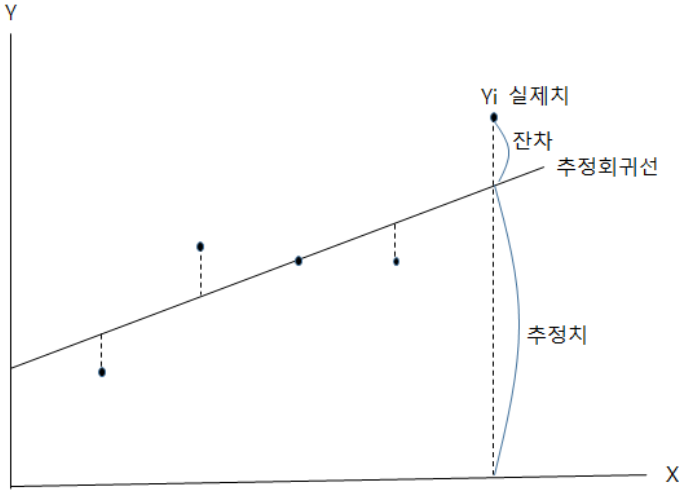
1. 회귀분석

- 회귀분석이란 한 변수와 다른 변수 사이의 관계를 분석하는 방법
- 회귀분석에서 변수는 영향을 주는 독립변수와 영향을 받는 종속변수로 구분
- 독립변수는 확정변수로 가정하고 종속변수는 확률변수로 가정
- 회귀분석의 목적은 독립변수의 주어진 값으로 종속변수의 평균값을 예측

2. 보통최소자승법

- 잔차(=실제치-추정치)의 합계가 최소가 되도록 하는 것이 바람직함
- 잔차의 합이 0이 되는 식은 유일하지가 않음
- 잔차의 제곱의 합이 최소가 되게 하는 회귀식을 구하게 되는데 이러한 추정방법을 보통최소자승법 (Ordinary Least Squares: OLS)이라고 함
- 아래 그림은 단순회귀모형, 추정 회귀선 및 잔차를 나타내 주고 있음

- 모형 : $Y_i = \beta_0 + \beta_1 X_i + u_i$
- 추정된 회귀선 : $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_i$
- 잔차 : $e_i = Y_i - \widehat{Y_i}$



- 단순회귀모형 회귀계수의 추정식은 각각 다음과 같음

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n} X_i Y_i - \overline{X}\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \overline{X}\sum_{i=1}^{n} X_i} = \frac{cov(X, Y)}{var(X)} \qquad \widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X}$$

(1) 추정식을 이용한 계산

- (예 1) X 와 Y에 대한 데이터가 아래 좌측과 같을 때 최소자승법으로 추정한 회귀계수를 각각 계산해 보라
  - · A2부터 A6셀에 X 데이터를 입력하고, B2부터 B6셀에 Y 데이터를 입력
  - · A7에 =sum(a2:a6)을 입력하여 $\sum_{i=1}^{5} X_i$를 구하고, A8에 =average(a2:a6)를 입력하여 $\bar{X}$를 구함
  - · 동일한 방법으로 B7에 $\sum_{i=1}^{5} Y_i$를 구하고, B8에 $\bar{X}$를 구함
  - · C2부터 C6셀에 X^2를 계산하고, D2부터 D6셀에 XY를 계산
  - · C7에 sum(c2:c6)을 입력하여 $\sum_{i=1}^{5} X_i^2$를 구하고, D7에 =sum(d2:d6)을 입력하여 $\sum_{i=1}^{5} X_i Y_i$를 구함
  - · $\hat{\beta}_1 = \frac{\sum_{i=1}^{5} X_i Y_i - \bar{X} \sum_{i=1}^{5} Y_i}{\sum_{i=1}^{5} X_i^2 - \bar{X} \sum_{i=1}^{5} X_i}$를 계산하기 위해 C10에 =(D7-A8*B7)/(C7-A8*A7)을 입력하면 1.4가 계산됨
  - · $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$를 계산하기 위해 C11에 =B8-C10*A8을 입력하면 0.4가 계산됨

X	Y
2	4
3	4
4	6
5	6
6	10

	A	B	C	D	E	F
1	X	Y	X^2	X*Y		
2	2	4	4	8		
3	3	4	9	12		
4	4	6	16	24		
5	5	6	25	30		
6	6	10	36	60		
7	20	30	90	134		
8	4	6				
9						
10		beta1=	1.4			
11		beta0=	0.4			

=sum(a2:a6) =average(a2:a6) =sum(c2:c6) =sum(d2:d6)

(2) 행렬추정식을 이용한 계산

- 행렬을 이용하면 회귀계수는 다음과 같이 추정할 수 있음

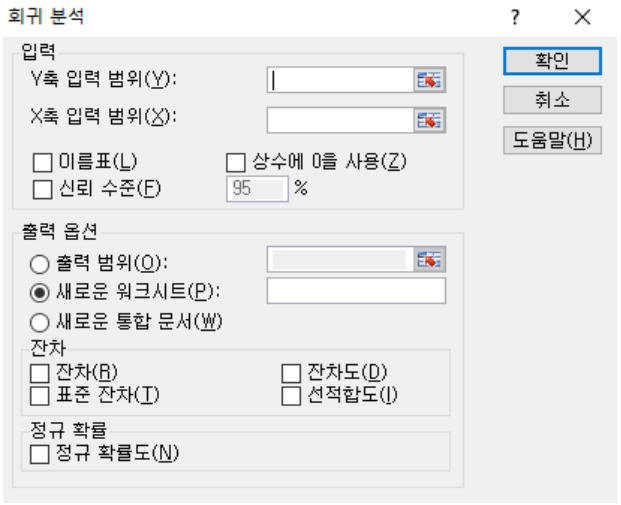
$$\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = (X'X)^{-1}X'Y$$

- A14부터 B18셀에 X 행렬을 만드는데, A14부터 A18까지는 상수항 $\hat{\beta}_0$추정에 필요한 1을 입력하고, B14부터 B18까지는 $\hat{\beta}_1$추정에 필요한 X 데이터를 A2부터 A6셀에 복사해 옴
- D14부터 H15셀에 X의 전치행렬을 만들고, A22부터 B23에 X'X , D22부터 E23에 $(X'X)^{-1}$, A26부터 A27에 X'Y를 만듦
- D26부터 D27을 선택하고 =MMULT(D22:E23,A26:A27) 입력하여 Ctrl + shift + Enter를 누르면 회귀계수를 계산해 줌

	A	B	C	D	E	F	G	H
1	X	Y	X^2	X*Y				
12								
13	X			X'				
14	1	2		1	1	1	1	1
15	1	3		2	3	4	5	6
16	1	4						
17	1	5						
18	1	6						
19								
20								
21	X'X			(X'X)^-1				
22	5	20		1.8	-0.4			
23	20	90		-0.4	0.1			
24								
25	X'Y							
26	30		beta0=	0.4				
27	134		beta1=	1.4				

(3) 데이터 분석을 이용한 계산

- 데이터-분석-데이터 분석을 실행하면 나타나는 통계 데이터분석 대화상자에서 회귀분석을 실행하면 아래 좌측 그림과 같은 회귀분석 대화상자가 나타남
- 먼저 종속변수의 데이터 범위를 나타내는 Y축 입력 범위에 $B$2:$B$6을 입력
- 독립변수의 데이터 범위를 나타내는 X축 입력 범위에 $A$2:$A$6을 입력
- 회귀 분석의 결과를 인쇄할 영역을 나타내는 출력 범위에 $B$29를 입력하고 확인을 누르면 아래 우측 그림과 같은 회귀 분석 결과를 출력



	A	B	C	D	E	F	G	H	I	J
1	X	Y	X^2	X*Y						
29		요약 출력								
30										
31		회귀분석 통계량								
32		다중 상관계	0.903696							
33		결정계수	0.816667							
34		조정된 결정	0.755556							
35		표준 오차	1.21106							
36		관측수	5							
37										
38		분산 분석								
39			자유도	제곱합	제곱 평균	F 비	유의한 F			
40		회귀	1	19.6	19.6	13.36364	0.035353			
41		잔차	3	4.4	1.466667					
42		계	4	24						
43										
44			계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
45		Y 절편	0.4	1.624808	0.246183	0.821423	-4.77086	5.570863	-4.77086	5.570863
46		X 1	1.4	0.382971	3.655631	0.035353	0.181216	2.618784	0.181216	2.618784
47										

- (예 2) 소비(단위 : 10억 원)와 소득(단위 : 10억 원)의 자료(consumption.xlsx)로 소비함수를 추정하고 예측하는 회귀분석을 해보라
  - http://kanggc.iptime.org/book/data/consumption.xlsx로 파일을 다운로드(하단 좌측)
  - 데이터-분석-데이터 분석을 실행하여 통계 데이터분석 대화상자에서 회귀분석을 선택하면 나타나는 회귀분석 대화상자에서 Y축 입력범위에 $C$2:$C$18을 선택하고, X축 입력 범위에 $B$2:$B$18을 선택하여 확인을 클릭하면 하단 우측 그림과 같은 소비함수 추정 결과가 나타남
  - 2017년 소비를 예측하기 위하여 B21셀에 식 =b17+b18*1700000을 입력하면 2017년 소비 예측치가 1099983인 것으로 계산됨

	A	B	C
1	연도	국민총소득	최종소비지출
2	2000	630,614.30	413,461.20
3	2001	683,447.10	460,668.30
4	2002	758,862.60	515,616.00
5	2003	807,778.00	535,967.40
6	2004	874,238.70	562,020.30
7	2005	912,608.60	602,345.40
8	2006	962,446.60	643,408.00
9	2007	1,040,091.80	691,740.40
10	2008	1,104,414.30	740,804.60
11	2009	1,148,981.80	769,588.60
12	2010	1,266,579.80	819,821.20
13	2011	1,340,529.80	873,522.60
14	2012	1,391,595.50	911,938.20
15	2013	1,439,644.40	942,267.20
16	2014	1,490,763.90	972,924.90
17	2015	1,568,383.10	1,006,005.60
18	2016	1,639,066.50	1,047,482.40

	A	B	C	D	E	F	G	H	I
1	요약 출력								
2									
3	회귀분석 통계량								
4	다중 상관	0.99871							
5	결정계수	0.997421							
6	조정된 결	0.99725							
7	표준 오차	10600.86							
8	관측수	17							
9									
10	분산 분석								
11		자유도	제곱합	제곱 평균	F 비	유의한 F			
12	회귀	1	6.52E+11	6.52E+11	5802.256	7.81E-21			
13	잔차	15	1.69E+09	1.12E+08					
14	계	16	6.54E+11						
15									
16		계수	표준 오차	t 통계량	P-값	하위 95%	상위 95%	하위 95.0%	상위 95.0%
17	Y 절편	30545.6	9609.735	3.17861	0.006231	10062.94	51028.27	10062.94	51028.27
18	X 1	0.629081	0.008259	76.17254	7.81E-21	0.611478	0.646684	0.611478	0.646684
19									
20		2017년 소비							
21		1099983							

II. R 단순회귀분석

b3-ch5-2.R

```
x<-c(2,3,4,5,6)
y<-c(4,4,6,6,10)
x;y
(n<-length(x))

(sumx<-sum(x));(sumy<-sum(y))
(mx=mean(x));(my=mean(y))
(xy<-x*y)
(sumxy<-sum(xy));(sumxsq<-sum(x^2));(sumysq<-sum(y^2))
beta1<-(sumxy-mx*sumy)/(sumxsq-mx*sumx)
beta0<-my-beta1*mx
beta0;beta1
(b1hat<-cov(x,y)/var(x))
(b0hat<-mean(y)-b1hat*mean(x))

X<-matrix(c(1,1,1,1,1,2,3,4,5,6), nrow=5, ncol=2)
Y<-matrix(c(4,4,6,6,10), nrow=5)
N<-length(Y)
X
(XPX<-t(X)%*%X)
(XPXINV<-solve(XPX))
(XPY<-t(X)%*%Y)
(beta<-XPXINV%*%XPY)

ols<-lm(y~x)
summary(ols)
```

```
> (sumx<-sum(x))
[1] 20

> (sumy<-sum(y))
[1] 30

> (mx=mean(x))
[1] 4

> (my=mean(y))
[1] 6
```

```
> (sumxy<-sum(xy))
[1] 134

> (sumxsq<-sum(x^2))
[1] 90

> (sumysq<-sum(y^2))
[1] 204
```

```
> beta0;beta1
[1] 0.4
[1] 1.4

> (b1hat<-cov(x,y)/var(x))
[1] 1.4

> (b0hat<-mean(y)-b1hat*mean(x))
[1] 0.4
```

```
> X
     [,1] [,2]
[1,]    1    2
[2,]    1    3
[3,]    1    4
[4,]    1    5
[5,]    1    6
```

```
> (XPX<-t(X)%*%X)
     [,1] [,2]
[1,]    5   20
[2,]   20   90
```

```
> (XPXINV<-solve(XPX))
     [,1] [,2]
[1,]  1.8 -0.4
[2,] -0.4  0.1
```

```
> (XPY<-t(X)%*%Y)
     [,1]
[1,]   30
[2,]  134
```

```
> (beta<-XPXINV%*%XPY)
     [,1]
[1,]  0.4
[2,]  1.4
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.400      1.625   0.246   0.8214
x              1.400      0.383   3.656   0.0354 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.211 on 3 degrees of freedom
Multiple R-squared:  0.8167,    Adjusted R-squared:  0.7556
F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535
```

b3-ch5-4.R

```
library(openxlsx)

df<-read.xlsx("http://kanggc.iptime.org/book/data/consumption-e.xlsx")

year<-df[,1]
gdp<-df[,2]
consumption<-df[,3]

y<-ts(gdp, start=c(2000), frequency=1)
c<-ts(consumption, start=c(2000), frequency=1)

n<-length(y)

ny<-y[2:n]
nc<-c[2:n]
lagc<-c[1:n-1]

lm_s<-lm(c~y)
ols_s<-summary(lm_s)
ols_s
y0<-1700000
(chat_s<-summary(lm_s)$coef[1]+summary(lm_s)$coef[2]*y0)
```

다음 장에 계속

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.055e+04  9.610e+03   3.179  0.00623 **
y           6.291e-01  8.259e-03  76.173  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10600 on 15 degrees of freedom
Multiple R-squared:  0.9974,    Adjusted R-squared:  0.9972
F-statistic:  5802 on 1 and 15 DF,  p-value: < 2.2e-16
```

```
> y0<-1700000

> (chat_s<-summary(lm_s)$coef[1]+summary(lm_s)$coef[2]*y0)
[1] 1099983
```