Part 2. R 통계분석 (데이터 분석 전문가 양성과정)

14

회귀분석의 유형

경북대학교 배준현 교수
(joonion@knu.ac.kr)

14. 회귀분석의 유형

- **선형회귀의 유형:**
  - **단순 선형회귀:** *simple(univariate) linear regression*
    - 한 개의 독립변수와 종속변수 간의 단순한(일차) 선형 관계
  - **다중 선형회귀:** *multiple(multivariate)* linear regression
    - 두 개 이상의 독립변수와 종속변수 간의 선형 관계
  - **다항 선형회귀:** *polynomial linear regression*
    - 종속변수와 한 개의 독립변수의 다항식으로 구성된 비선형 관계

14. 회귀분석의 유형

- lm() 함수와 *formula*:
  - `lm(formula, data)`:
    - `formula`: 종속변수와 독립변수 간의 관계를 설명하는 형식
    - `data`: *formula*를 적용할 데이터 객체

종속변수 ~ 독립변수1 + 독립변수2 + 독립변수3

14. 회귀분석의 유형

▪ 포뮬러 심볼: *formula symbols*

심볼	설명	사용법	해석
.	종속변수를 제외한 모든 변수	y ~ .	y ~ x1 + x2 + … + xn
:	독립변수 간의 상호작용	a:b	
*	독립변수 간의 모든 가능한 상호작용	a * b	a + b + a:b
^	지정한 차수까지의 상호작용	(a + b)^2	(a + b) * (a + b)
-	독립변수를 제외함	(a + b)^2 – a:b	a + b + a:c
I()	괄호 안의 연산자를 산술적으로 해석	y ~ x +I(w+z^2)	u <- w + x^2 y ~ x + u

14. 회귀분석의 유형

- Prestige 데이터셋
  - 캐나다의 인구조사 데이터(1971년): 변수 6개, 관측값 102개
    - education: 재직자의 평균 교육기간 (years)
    - income: 재직자의 평균 소득 (dollars)
    - women: 여성 재직자의 비율
    - prestige: 직업에 대한 명성 점수 (1960년대 중반에 실시된 사회 조사 결과)
    - census: 캐나다의 직업 코드
    - type: 직업 분류: bc: blue color, prof: professional, wc: white color

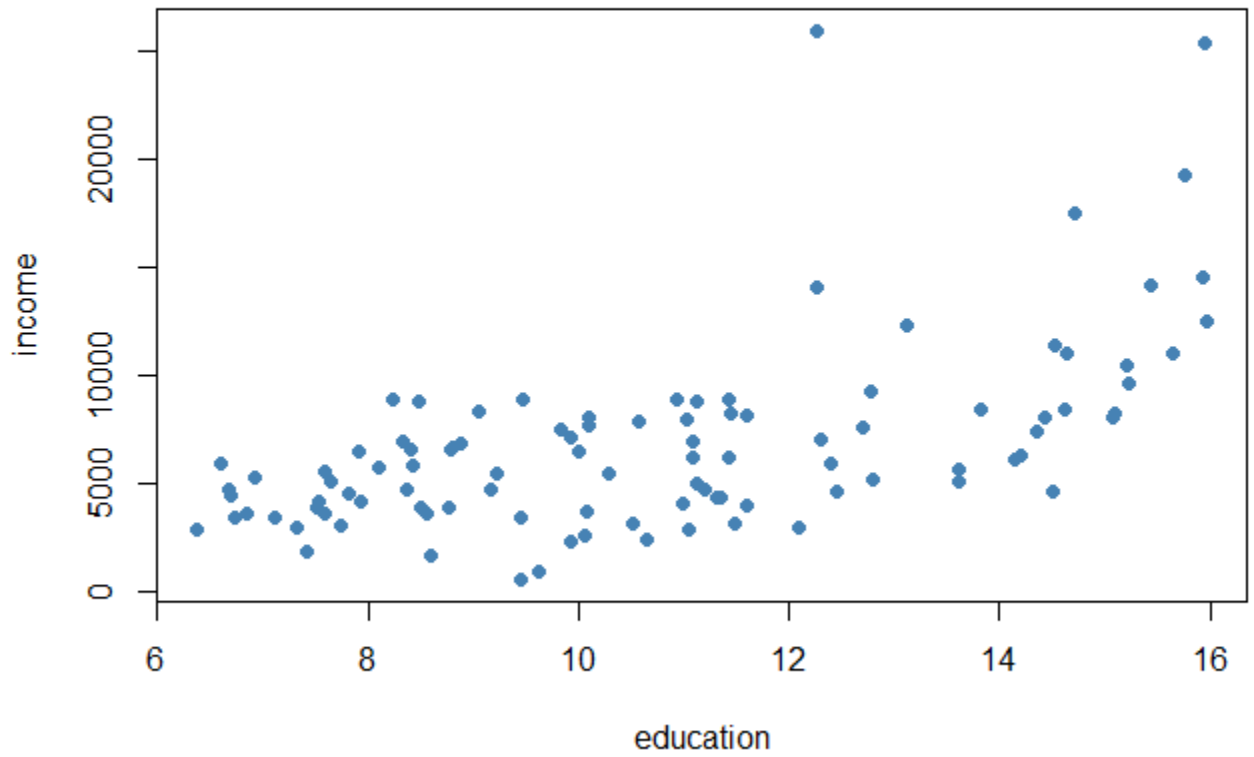
14. 회귀분석의 유형

- **단순 선형회귀**: *simple* linear regression
  - **교육기간과 평균소득 간에는 선형 관계가 있을까?**
    - **종속변수: 평균소득**(income)
    - **독립변수: 교육기간**(education)

```
> library(car)
> str(Prestige)
'data.frame':  102 obs. of  6 variables:
 $ education: num  13.1 12.3 12.8 11.4 14.6 ...
 $ income   : int  12351 25879 9271 8865 8403 11030 8258 14163 11377 11023 ...
 $ women    : num  11.16 4.02 15.7 9.11 11.68 ...
 $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...
 $ census   : int  1113 1130 1171 1175 2111 2113 2133 2141 2143 2153 ...
 $ type     : Factor w/ 3 levels "bc","prof","wc": 2 2 2 2 2 2 2 2 2 2 ...
```

14. 회귀분석의 유형

```
> df <- Prestige
> plot(income ~ education, data = df, pch = 19, col = "steelblue")
```

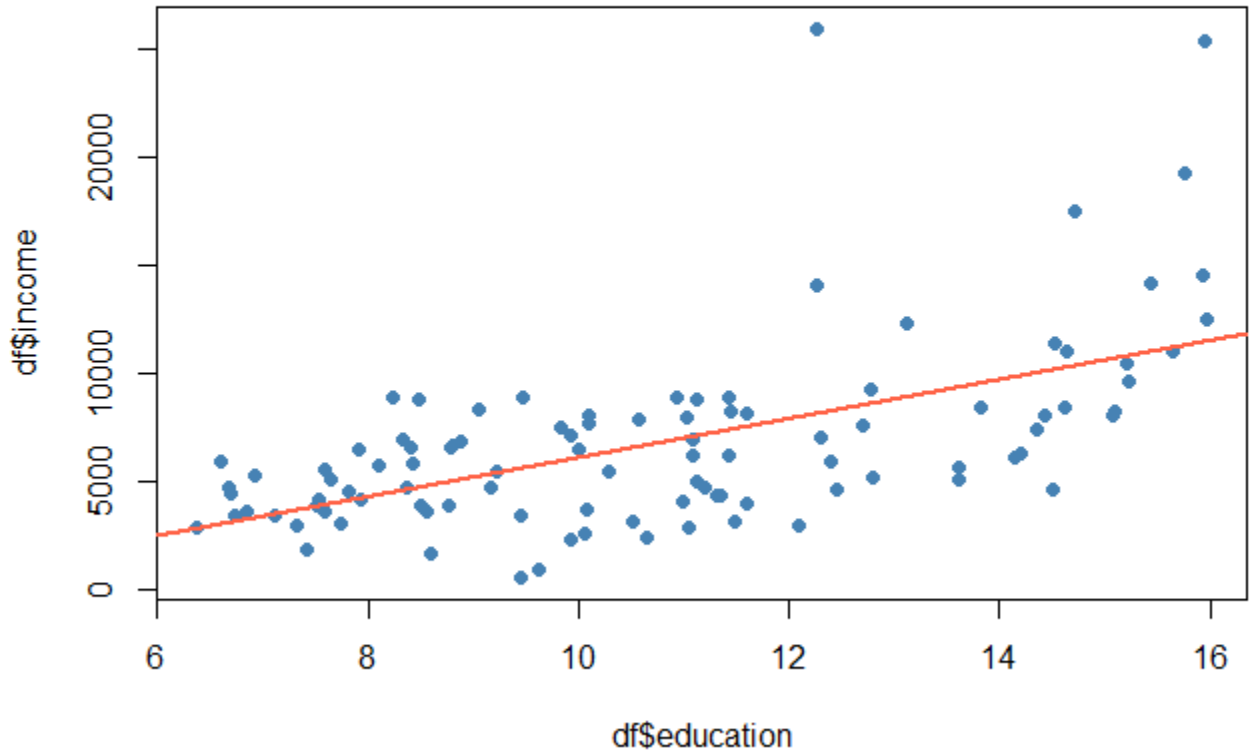


14. 회귀분석의 유형

```
> cor(df$education, df$income)
[1] 0.5775802

> formula <- income ~ education
> lm(formula = formula, data = Prestige)

Call:
lm(formula = formula, data = Prestige)

Coefficients:
(Intercept)    education
    -2853.6        898.8
```

14. 회귀분석의 유형

```
> model <- lm(formula = formula, data = Prestige)
> abline(model, lwd = 2, col = "tomato")
```



14. 회귀분석의 유형

```
> summary(model)

Call:
lm(formula = formula, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-5493.2 -2433.8   -41.9  1491.5 17713.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2853.6     1407.0  -2.028   0.0452 *
education      898.8      127.0   7.075 2.08e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3483 on 100 degrees of freedom
Multiple R-squared:  0.3336,     Adjusted R-squared:  0.3269
F-statistic: 50.06 on 1 and 100 DF,  p-value: 2.079e-10
```

14. 회귀분석의 유형

```
> summary(resid(model))
Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-5493.20 -2433.80   -41.92     0.00  1491.50 17713.14

> confint(model)
2.5 %     97.5 %
(Intercept) -5645.1114  -62.05979
education     646.7782 1150.84748

> anova(model)
Analysis of Variance Table

Response: income
           Df     Sum Sq   Mean Sq F value    Pr(>F)
education   1  607421386 607421386   50.06 2.079e-10 ***
Residuals 100 1213392025  12133920
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **다중 선형회귀**: *multiple* linear regression
  - **종속변수에 영향을 미치는 독립변수가 여러 개일 경우**
    - **다중 회귀식**: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$
  - **평균소득에 영향을 주는 요인은 무엇일까?**
    - **종속변수**: **평균소득**(income)
    - **독립변수**: **교육**(education), **성별**(women), **명성**(prestige)

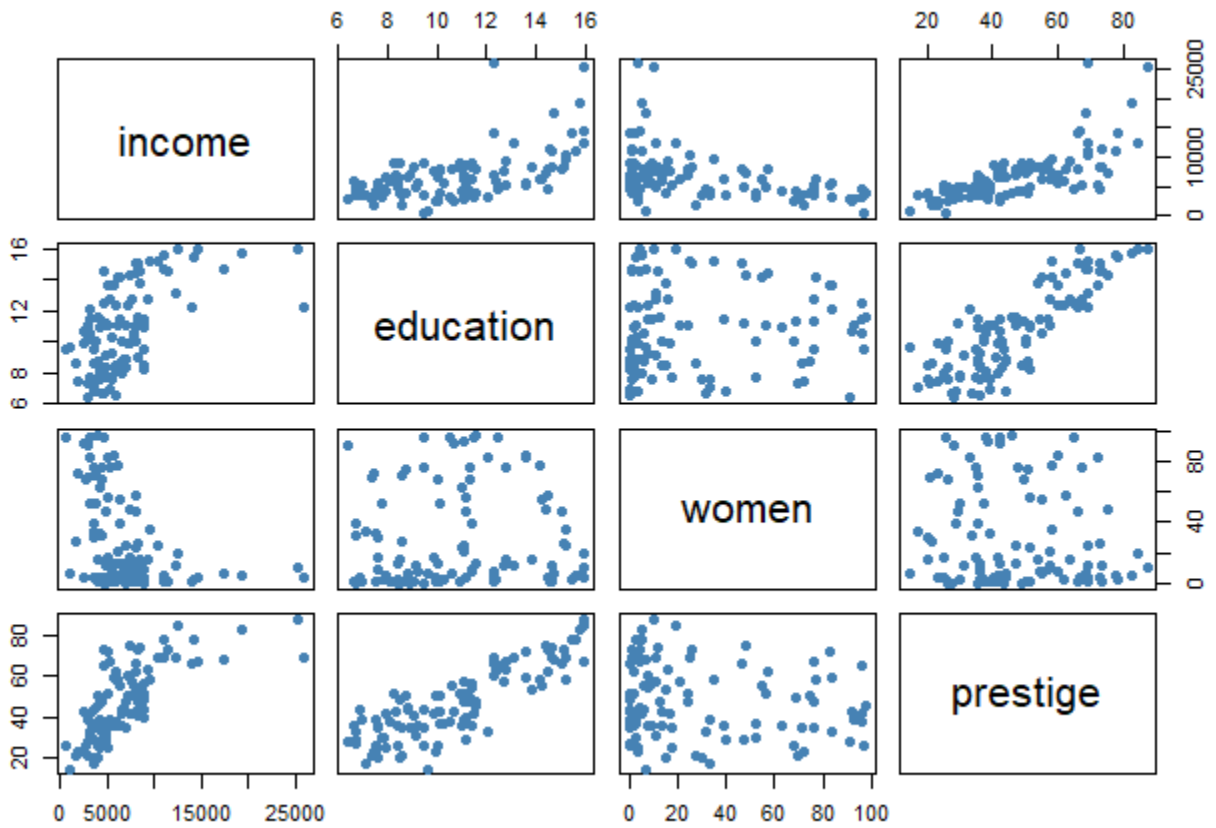
income ~ education + women + prestige

14. 회귀분석의 유형

```
> library(car)
> df <- subset(Prestige, select = c(2, 1, 3, 4))
> str(df)
'data.frame':  102 obs. of  4 variables:
 $ income   : int  12351 25879 9271 8865 8403 11030 8258 14163 11377 11023 ...
 $ education: num  13.1 12.3 12.8 11.4 14.6 ...
 $ women    : num  11.16 4.02 15.7 9.11 11.68 ...
 $ prestige : num  68.8 69.1 63.4 56.8 73.5 77.6 72.6 78.1 73.1 68.8 ...

> cor(df)
                income  education       women   prestige
income      1.0000000 0.57758023 -0.44105927  0.7149057
education   0.5775802 1.00000000  0.06185286  0.8501769
women      -0.4410593 0.06185286  1.00000000 -0.1183342
prestige    0.7149057 0.85017689 -0.11833419  1.0000000
```

14. 회귀분석의 유형

```
> plot(df, pch = 19, col = "steelblue")
```



```
> formula = income ~ education + women + prestige
> lm(formula, data = df)

Call:
lm(formula = formula, data = df)

Coefficients:
(Intercept)    education        women     prestige
     -253.8        177.2        -50.9        141.4
```

14. 회귀분석의 유형

```
> model <- lm(income ~ ., data = df)
> summary(model)

Call:
lm(formula = income ~ ., data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-7715.3  -929.7  -231.2   689.7 14391.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -253.850   1086.157  -0.234    0.816
education    177.199    187.632   0.944    0.347
women        -50.896      8.556  -5.948 4.19e-08 ***
prestige     141.435     29.910   4.729 7.58e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2575 on 98 degrees of freedom
Multiple R-squared:  0.6432,     Adjusted R-squared:  0.6323
F-statistic: 58.89 on 3 and 98 DF,  p-value: < 2.2e-16
```

14. 회귀분석의 유형

```
> library(stargazer)
> stargazer(model, type="text", no.space = TRUE)
================================================
                        Dependent variable:
                    ----------------------------
                               income
------------------------------------------------
education                      177.199
                              (187.632)
women                        -50.896***
                               (8.556)
prestige                     141.435***
                               (29.910)
Constant                       -253.850
                             (1,086.157)
------------------------------------------------
Observations                     102
R2                              0.643
Adjusted R2                     0.632
Residual Std. Error      2,574.709 (df = 98)
F Statistic           58.890*** (df = 3; 98)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

- **다항 선형회귀:** *polynomial* linear regression
  - 종속변수를 독립변수의 다항식이 더 잘 설명하는 경우
    - 다항 회귀식: $y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \cdots + \beta_n x^n$
  - 교육기간과 평균소득의 관계를 직선보다 더 잘 설명하는 곡선이 있을까?
    - 종속변수: 평균소득(income)
    - 독립변수: 교육기간(education)

14. 회귀분석의 유형

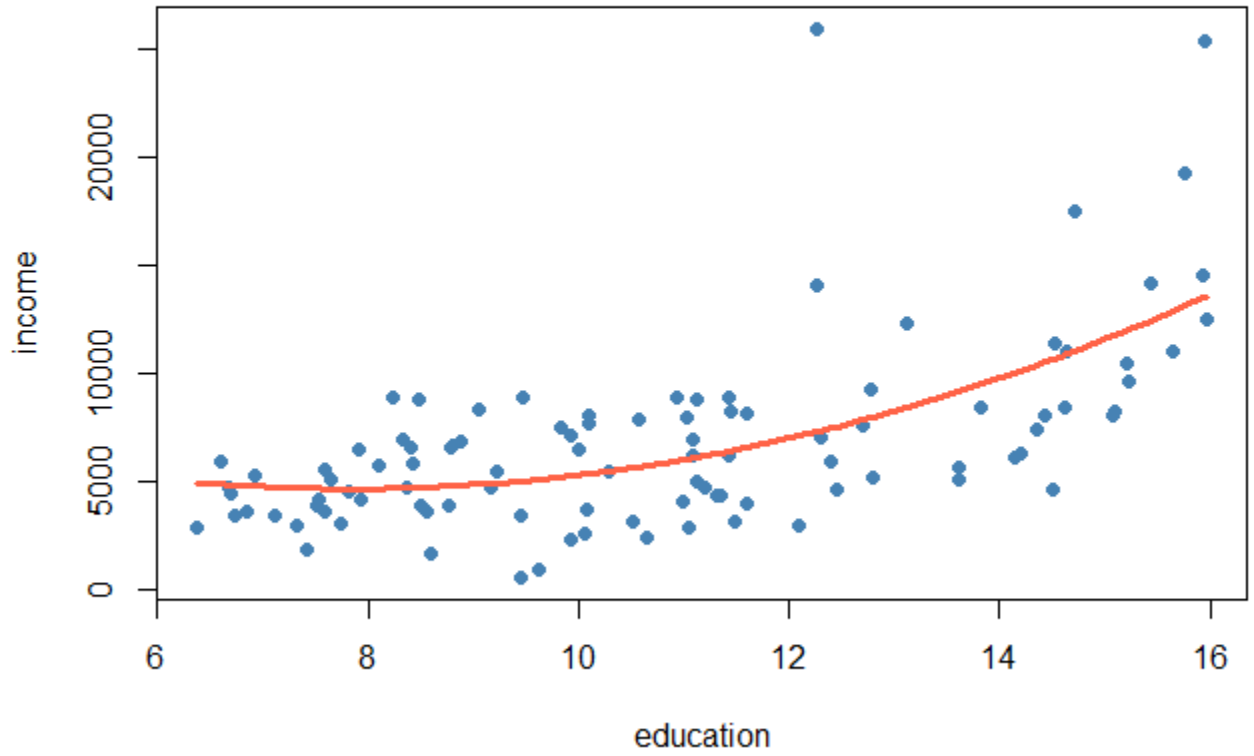
```
> library(car)
> formula <- income ~ education + I(education^2)
> lm(formula, data = Prestige)

Call:
lm(formula = formula, data = Prestige)

Coefficients:
   (Intercept)       education  I(education^2)
       12918.2         -2102.9           134.2
```

14. 회귀분석의 유형

```
> model <- lm(formula, data = Prestige)
> summary(model)

Call:
lm(formula = formula, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-5951.4 -2091.1  -358.2  1762.4 18574.2

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    12918.23    5762.27   2.242  0.02720 *
education      -2102.90    1072.73  -1.960  0.05277 .
I(education^2)   134.18      47.64   2.817  0.00586 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3369 on 99 degrees of freedom
Multiple R-squared:  0.383,      Adjusted R-squared:  0.3706
F-statistic: 30.73 on 2 and 99 DF,  p-value: 4.146e-11
```

14. 회귀분석의 유형

```
> plot(income ~ education, data = Prestige, pch = 19, col = "steelblue")
> library(dplyr)
> with(Prestige,
       lines(arrange(data.frame(education, fitted(model)), education),
             lty = 1, lwd = 3, col = "tomato"))
```



Any Questions?