회귀분석 CH3

중어중문학과 2019131238 정예린

Exercise 3.3

Table 3.10 shows the scores in the final examination F and the scores in two preliminary examinations P_1 and P_2 for 22 students in a statistics course. The data can be found at the book's Web site.

(a) Fit each of the following models to the data :

Model 1 : $F=\beta_0+\beta_1 P_1+\quad\quad+\epsilon$

Model 2 : $F=\beta_0+\quad\quad+\beta_2 P_2+\epsilon$

Model 3 : $F=\beta_0+\beta_1 P_1+\beta_2 P_2+\epsilon$

(R 프로그래밍 코드)

데이터(P083.txt) 읽기

```
> data <- read.table("C:/data/P083.txt", header=T)
> Y <- cbind(data[,1])
> X1 <- cbind(data[,2])
> X2 <- cbind(data[,3])
```

Model 1 선형회귀 적합

```
> model_1 <- lm(Y ~ X1)
> summary(model_1)

Call:
lm(formula = Y ~ X1)

Residuals:
   Min     1Q Median     3Q    Max 
-8.844 -2.020 -0.587  4.043  7.938 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -22.3424    11.5640  -1.932   0.0676 .  
X1            1.2605     0.1399   9.008 1.78e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.081 on 20 degrees of freedom
Multiple R-squared:  0.8023,	Adjusted R-squared:  0.7924 
F-statistic: 81.14 on 1 and 20 DF,  p-value: 1.779e-08
```

→ Model 1에 대한 적합값 : $F=-22.34+1.26P_1$

Model 2 선형회귀 적합

```
> model_2 <- lm(Y ~ X2)
> summary(model_2)

Call:
lm(formula = Y ~ X2)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.4323  -1.5027   0.5421   2.2580   7.5165 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.85355    7.56181  -0.245    0.809    
X2           1.00427    0.09059  11.086 5.44e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.275 on 20 degrees of freedom
Multiple R-squared:   0.86,	Adjusted R-squared:  0.853 
F-statistic: 122.9 on 1 and 20 DF,  p-value: 5.442e-10
```

→ Model 2에 대한 적합값 : $F = -1.85 + 1.00P_2$

Model 3 선형회귀 적합

```
> model_3 <- lm(Y ~ X1 + X2)
> summary(model_3)

Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max 
-8.7328 -2.1703  0.3938  2.6443  6.3660 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -14.5005     9.2356  -1.570  0.13290   
X1            0.4883     0.2330   2.096  0.04971 * 
X2            0.6720     0.1793   3.748  0.00136 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.953 on 19 degrees of freedom
Multiple R-squared:  0.8863,	Adjusted R-squared:  0.8744 
F-statistic: 74.07 on 2 and 19 DF,  p-value: 1.069e-09
```

→ Model 3에 대한 적합값 : $F = -14.50 + 0.49P_1 + 0.67P_2$

∴ R 프로그램 구동 결과에 따라 각 모델에 대한 적합값은

Model 1 : $F = -22.34 + 1.26P_1$,

Model 2 : $F = -1.85 + 1.00P_2$,

Model 3 : $F = -14.50 + 0.49P_1 + 0.67P_2$이다.

(b) Test whether β_0 = 0 in each of the three models.

Model 1, 2, 3에 대해 F검정을 진행한다. (H_0 : β_0 = 0 vs $H_1 : \beta_0 \neq$ 0)

(R 프로그래밍 코드 : broom 패키지 → 각 모형의 통계량과 성능을 한번에 확인하기 위함)

\# Model 1

```
> glance(model_1)
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic    p.value    df logLik   AIC
      <dbl>         <dbl> <dbl>     <dbl>      <dbl> <dbl>  <dbl> <dbl>
1     0.802         0.792  5.08      81.1   1.78e-8      1  -65.9  138.
# ... with 4 more variables: BIC <dbl>, deviance <dbl>,
#   df.residual <int>, nobs <int>
```

→ 검정통계량 F = 81.1, P-value는 매우 작은 값(1.78e-08)으로 p값 < α이다.

∴ 따라서 H_0를 기각할 수 있다 ⇒ β_0은 0이 아니라고 할 수 있다.

\# Model 2

```
> glance(model_2)
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic    p.value    df logLik   AIC
      <dbl>         <dbl> <dbl>     <dbl>      <dbl> <dbl>  <dbl> <dbl>
1     0.860         0.853  4.27      123.  5.44e-10      1  -62.1  130.
# ... with 4 more variables: BIC <dbl>, deviance <dbl>,
#   df.residual <int>, nobs <int>
```

→ 검정통계량 F = 123.5, P-value는 매우 작은 값(5.44e-10)으로 p값 < α이다.

∴ 따라서 H_0를 기각할 수 있다 ⇒ β_0은 0이 아니라고 할 수 있다.

\# Model 3

```
> glance(model_3)
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic    p.value    df logLik   AIC
      <dbl>         <dbl> <dbl>     <dbl>      <dbl> <dbl>  <dbl> <dbl>
1     0.886         0.874  3.95      74.1   1.07e-9      2  -59.8  128.
# ... with 4 more variables: BIC <dbl>, deviance <dbl>,
#   df.residual <int>, nobs <int>
```

→ 검정통계량 F = 74.1, P-value는 매우 작은 값(1.07e-9)으로 p값 < α이다.

∴ 따라서 H_0를 기각할 수 있다 ⇒ β_0은 0이 아니라고 할 수 있다.

(c) Which variable individually, P_1 or P_2, is a better predictor of F?

```
> cor(Y, X1)
          [,1]
[1,] 0.8956842
```

```
> cor(Y, X2)
          [,1]
[1,] 0.9273811
```

$\Rightarrow Corr(F \sim P_2) > Corr(F \sim P_1)$

또 (b)의 통계량에서 P_2의 R-squred값(0.860)이 P_1의 R-squared값(0.802)보다 크므로, P_2가 P_1보다 F의 변동을 잘 설명할 수 있다고 말할 수 있다.

∴ P_2가 P_1보다 더 좋은 설명변수이다.

(d) Which of the three models would you use to predict the final examination scores for a student who scored 78 and 85 on the first and second preliminary examinations, respectively? What is your prediction in this case?
→ (b)의 R코드 결과값을 봤을 때 Model 3의 R-squred값(0.886)이 다른 두 모델 (Model 1 : 0.860, Model 2 : 0.802)보다 높기 때문에 Model 3을 이용하는 것이 가장 적절할 것이다.
Model 3 에서 $P_1 = 78$, $P_2 = 85$ ⇒ $F = -14.50 + 0.49 \times 78 + 0.67 \times 85 = 80.71$
∴ Model 3을 이용해 final examination score = 80.71이라고 predict할 수 있다.

Exercise 3.6
Table 3.11 shows the regression output, with some numbers erased, when a simple regression model relating a response variable Y to a predictor variable X_1 is fitted based on 20 observations. Complete the 13 missing numbers, then compute $Var(Y)$ and $Var(X_1)$.

Table 3.11 Regression Output When Y is Regressed on X_1 for 20 Observations

ANOVA Table				
Source	Sum of Squares	df	Mean Square	F-Test
Regression	1848.76	**(1)**	**(2)**	**(3)**
Residuals	**(4)**	**(5)**	**(6)**	
Coefficients Table				
Variable	Coefficient	s.e.	t-Test	p-value
Constant	−23.4325	12.74	**(7)**	0.0824
X_1	**(8)**	0.1528	8.32	< 0.0001
n = **(9)**	R^2 = **(10)**	R_a^2 = **(11)**	$\hat{\sigma}$ = **(12)**	df = **(13)**

(1) SSR의 자유도는 p(설명변수 개수) = 1이다. ∴ 1
(2) MSR = SSR/p = 1848.76/1 = 1848.76이다. ∴ 1848.76
(9) 문제에서 20개 관측치에 대한 적합값이라고 했으므로 n = 20이다. ∴ 20
(5) SSE의 자유도는 n-p-1 = 20-1-1 = 18이다. ∴ 18
(7) $\hat{\beta}_0 = -23.4325, SE(\hat{\beta}_0) = 12.74$가 주어졌으므로 $t_0 = -23.4325/12.74 = -1.839$ ∴ -1.839
(8) $SE(\hat{\beta}_1) = 0.1528, t_1 = 8.32$가 주어졌으므로 $\hat{\beta}_1 = 0.1528 \times 8.32 = 1.2713$ ∴ 1.2713
(3) $F = t^2$이므로 $F = 8.32^2 = 69.32$ ∴ 69.32
(6) MSE = MSR/F = 1848.76/69.32 = 26.67 ∴ 26.67
(4) SSE = MSE×(n-p-1) = 26.67×18 = 480.06
(10) R^2 = SSR/SST = SSR/(SSR+SSE) = 1848.76/(1848.76+480.06) = 0.7939 ∴ 0.7939

(11) R_a^2 = 1 – [SSE/(n-p-1)]/[SST/(n-1)] = 1 – 26.67/(2328.82/19) = 0.7824 ∴ 0.7824

(12) $\hat{\sigma}$ = $\sqrt{SSE/(n-p-1)}$ = $\sqrt{26.67}$ = 5.164 ∴ 5.164

(13) df = n-p-1 = 18이다. ∴ 18

∴ 위와 같은 계산에 따라 빈칸을 채운 표는 다음과 같다.

ANOVA Table				
Source	Sum of Squares	df	Mean Square	F-Test
Regression	1848.76	1	1848.76	69.32
Residuals	480.06	18	26.67	
Coefficients Table				
Variable	Coefficient	s.e.	t-Test	p-value
Constant	-23.4325	12.74	-1.839	0.0824
X_1	1.2713	0.1528	8.32	< 0.0001
n = 20	R^2 = 0.7939	R_a^2 = 0.7824	$\hat{\sigma}$ = 5.164	df = 18

이때 $SE(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\Sigma(x_i - \bar{x})^2} = \hat{\sigma}/\sqrt{(n-1)Var(X)}$ 이므로 $Var(X) = (\hat{\sigma}/SE(\hat{\beta}_1))^2/(n-1)$

→ $Var(X) = (5.164/0.1528)^2/19 = 33.80^2/19 = 60.13$ ∴ $Var(X) = 60.13$

$Var(Y) = SST/(n-1) = (1848.76 + 480.06)/19 = 2382.82/19 = 122.57$ ∴ $Var(Y) = 122.57$

Exercise 3.8

Construct the 95% confidence intervals for the individual parameters β_1 and β_2 using the regression output in Table 3.5.

Table 3.5 Regression Output for Supervisor Performance Data

Variable	Coefficient	s.e.	t-Test	p-value
Constant	10.787	11.5890	0.93	0.3616
X_1	0.613	0.1610	3.81	0.0009
X_2	−0.073	0.1357	−0.54	0.5956
X_3	0.320	0.1685	1.90	0.0699
X_4	0.081	0.2215	0.37	0.7155
X_5	0.038	0.1470	0.26	0.7963
X_6	−0.217	0.1782	−1.22	0.2356
$n = 30$	$R^2 = 0.73$	$R_a^2 = 0.66$	$\hat{\sigma} = 7.068$	df = 23

위 표에 의해 $\hat{\beta}_1 = 0.613,\ SE(\hat{\beta}_1) = 0.1610$, $\hat{\beta}_2 = -0.073,\ SE(\hat{\beta}_2) = 0.1357$

→ 95% CI for β_1 : $\hat{\beta}_1 \pm t_{(n-p-1,\,\alpha/2)} \times SE(\hat{\beta}_1)$, 이때 $n-p-1 = 30-6-1 = 23$, $\alpha = 0.05$

t table에 의해 $t_{(23,\,0.025)} = 2.069$이므로 $0.613 \pm 2.069 \times 0.1610 \approx (0.28, 0.95)$

→ 95% CI for β_2 : $\hat{\beta}_2 \pm t_{(n-p-1,\,\alpha/2)} \times SE(\hat{\beta}_2)$, 이때 $n-p-1 = 30-6-1 = 23$, $\alpha = 0.05$

t table에 의해 $t_{(23,\,0.025)} = 2.069$이므로 $-0.073 \pm 2.069 \times 0.1357 \approx (-0.35, 0.21)$

∴ 95% CI for β_1 : $(0.28, 0.95)$, 95% CI for β_2 : $(-0.35, 0.21)$

Exercise 3.10

Using the Supervisor Performance data, test the hypothesis $H_0 : \beta_1 = \beta_3 = 0.5$ in each of the following models:

(R 프로그래밍 코드)

데이터(P060.txt) 읽기 → a, b의 두 모델에서는 설명변수 $X_1, X_{2,} X_3$만 사용

```
> data <- read.table("C:/data/P060.txt", header=T)
> Y <- cbind(data[,1])
> X1 <- cbind(data[,2])
> X2 <- cbind(data[,3])
> X3 <- cbind(data[,4])
```

(a) $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \epsilon$

(R 프로그래밍 코드)

RM : $Y = \beta_0 + 0.5W + \epsilon$ ($W = X_1 + X_3$)에 대한 적합

```
> W <- X1+X3
> model_A <- lm(Y ~ W)
> summary(model_A)

Call:
lm(formula = Y ~ W)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.2052  -5.8973  -0.0372   5.4364  13.0172 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.98821    7.38841   1.352    0.187    
W            0.44439    0.05914   7.514 3.49e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.133 on 28 degrees of freedom
Multiple R-squared:  0.6685,    Adjusted R-squared:  0.6566 
F-statistic: 56.46 on 1 and 28 DF,  p-value: 3.487e-08
```

→ $\hat{Y} = 9.988 + 0.444(X_1 + X_3)$으로 추정된다. 이때 $\beta_1 = \beta_3 = 0.5$을 검정하기 위한 t검정 :

$|t| = \dfrac{0.44439 - 0.5}{0.05914} = 0.9403$, $0{,}9403 < t_{(27,\,0.025)} = 2.052$이므로 H_0를 기각할 수 없다.

(b) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

(R 프로그래밍 코드)

RM : $Y = \beta_0 + \beta_2 X_2 + 0.5W + \epsilon$ ($W = X_1 + X_3$)에 대한 적합

```
> W <- X1+X3
> model_B <- lm(Y ~ X2+W)
> summary(model_B)

Call:
lm(formula = Y ~ X2 + W)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.3879  -5.2613   0.1132   5.9849  13.6592 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.15323    7.68777   1.451    0.158    
X2          -0.08652    0.13558  -0.638    0.529    
W            0.47230    0.07407   6.376  7.9e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.21 on 27 degrees of freedom
Multiple R-squared:  0.6734,	Adjusted R-squared:  0.6492 
F-statistic: 27.83 on 2 and 27 DF,  p-value: 2.75e-07
```

→ $\hat{Y} = 11.15323 - 0.08652X_2 + 0.47230(X_1 + X_3)$으로 추정된다.

이때 $\beta_1 = \beta_3 = 0.5$을 검정하기 위한 t검정 : $|t| = \dfrac{0.47230 - 0.5}{0.07407} = 0.3740$,

이때 검정통계량 $0,3740 < t_{(27,\, 0.025)} = 2.052$이므로 H_0를 기각할 수 없다.

Exercise 3.15

Cigarette Consumption Data: A national insurance organization wanted to study the consumption pattern of cigarettes in all 50 states and the District of Columbia. The variables chosen for the study are given in Table 3.16. The data from 1970 are given in Table 3.17. The states are given in alphabetical order. The data can be found at the book's Website.

In (a)-(b) below, specify the null and alternative hypotheses, the test used, and your conclusion using a 5% level of significance.

변수 및 모형 설정

문제에서 주어진 내용에 따르면, Cigarette Consumption Data에 대한 연구에서

반응변수(Y) : Sales, 설명변수($X_1 \sim X_6$) : Age, HS, Income, Black, Female, Price이다.

→ 중회귀모형 : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon$

(a) Test the hypothesis that the variable Female is not needed in the regression equation relating Sales to the six predictor variables.

→ 귀무가설 $H_0 : \beta_5 = 0$ vs 대립가설 $H_1 : \beta_5 \neq 0$, t-Test를 사용해 검정한다.

(R 프로그래밍 코드)

데이터(P088.txt) 읽기 (table의 1열은 State)

```
> data <- read.table("C:/data/P088.txt", header=T)
> Y <- cbind(data[,8])
> X1 <- cbind(data[,2])
> X2 <- cbind(data[,3])
> X3 <- cbind(data[,4])
> X4 <- cbind(data[,5])
> X5 <- cbind(data[,6])
> X6 <- cbind(data[,7])
```

다중회귀분석 적합

```
> lm <- lm(Y ~ X1+X2+X3+X4+X5+X6)
> summary(lm)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.398 -12.388  -5.367   6.270 133.213 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 103.34485  245.60719   0.421  0.67597   
X1            4.52045    3.21977   1.404  0.16735   
X2           -0.06159    0.81468  -0.076  0.94008   
X3            0.01895    0.01022   1.855  0.07036 . 
X4            0.35754    0.48722   0.734  0.46695   
X5           -1.05286    5.56101  -0.189  0.85071   
X6           -3.25492    1.03141  -3.156  0.00289 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.17 on 44 degrees of freedom
Multiple R-squared:  0.3208,	Adjusted R-squared:  0.2282 
F-statistic: 3.464 on 6 and 44 DF,  p-value: 0.006857
```

→ β_5에 대한 $P-value=0.85071$로 $\alpha=0.05$보다 큰 값을 가지므로 H_0을 기각할 수 없다. 따라서 5% 유의수준에서 Female 변수가 회귀식에 필요하지 않다고 결론지을 수 있다.

(b) Test the hypothesis that the variables Female and HS are not needed in the above regression equation.

→ 귀무가설 $H_0:\beta_2=\beta_5=0$ vs $H_1:\beta_2\neq 0,\beta_5\neq 0$ or both, F-test를 이용해 검정한다.

(R 프로그래밍 코드)

ANOVA 분산분석

```
> lmB <- lm(Y ~ X2+X5)
> anova(lmB, lm)
Analysis of Variance Table

Model 1: Y ~ X2 + X5
Model 2: Y ~ X1 + X2 + X3 + X4 + X5 + X6
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)   
1     48 49310                                
2     44 34926  4     14384 4.5303 0.003758 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

→ $P-value = 0.00376$으로 $\alpha = 0.05$보다 작은 값을 가지므로 H_0를 기각할 수 있다.
따라서 Female 과 HS 변수가 회귀식에 필요하지 않다고 결론지을 수 없다.

(c) Compute the 95% confidence interval for the true regression coefficient of the variable Income.

→ (a)의 R코드 결과값에서 Income(X_3)의 회귀계수(β_3)에 대해 $\hat{\beta}_3 = 0.019, SE(\hat{\beta}_3) = 0.0102$

이때 95% CI : $\hat{\beta}_3 \pm t_{(n-p-1,\alpha/2)} \times SE(\hat{\beta}_3) = 0.019 \pm 2.01 \times 0.0102 = (-0.002, 0.04)$

∴ Income 변수의 회귀계수에 대한 95% 신뢰구간은 (-0.002, 0.04)이다.

(d) What percentage of the variation in Sales can be accounted for when Income is removed from the above regression equation? Explain.

→ Income(X_3) 변수가 제거되고 난 뒤 회귀식의 결정계수(R^2)를 구하면 다음과 같다.

(R 프로그래밍 코드)

```
> remX3 <- lm(Y ~ X1+X2+X4+X5+X6)
> glance(remX3)
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl>
1     0.268         0.186  28.9      3.29  0.0129     5  -241.  496.
# ... with 4 more variables: BIC <dbl>, deviance <dbl>,
#   df.residual <int>, nobs <int>
```

∴ $R^2 = 0.268$

따라서 Sales의 변동의 26.8%를 Income 변수가 제거되고 난 뒤의 회귀식으로 계산할 수 있다.

(e) What percentage of the variation in Sales can be accounted for by the three variables: Price, Age, and Income? Explain.

→ Price(X_6), Age(X_1), Income(X_3) 변수로 구성된 회귀식의 결정계수(R^2)는 다음과 같다.

(R 프로그래밍 코드)

```
> X136 <- lm(Y ~ X1+X3+X6)
> glance(X136)
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl>
1     0.303         0.259  27.6      6.82 0.000657     3  -240.  489.
# ... with 4 more variables: BIC <dbl>, deviance <dbl>,
#   df.residual <int>, nobs <int>
```

∴ $R^2 = 0.303$

따라서 Sales의 변동의 30.3%를 Price, Age와 Income 변수로 계산할 수 있다.

(f) What percentage of the variation in Sales that can be accounted for by the variable Income, when Sales is regressed on only Income? Explain.

→ Income(X_3) 변수만 포함하는 회귀식의 결정계수(R^2)는 다음과 같이 구할 수 있다.

(R 프로그래밍 코드)

```
> onlyX3 <- lm(Y ~ X3)
> glance(onlyX3)
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl>
1     0.106        0.0881  30.6      5.83  0.0195     1  -246.  498.
# ... with 4 more variables: BIC <dbl>, deviance <dbl>,
#   df.residual <int>, nobs <int>
```

∴ $R^2 = 0.106$

따라서 Sales의 변동의 10.6%를 Income 변수로 계산할 수 있다.