3.3 표 3.10은 통계학 과목을 수강한 22명의 학생들에 대하여 기말시험 성적F와 두 번의 기초시험 성적 $P_1, P_2$를 보여준다. 데이터는 이책의 웹사이트에서도 찾을 수 있다.

(a) 데이터에 다음의 모형들 각각을 적합하여라.

$$\begin{aligned} &\text{모델 1: } F = \beta_0 + \beta_1 P_1 + \qquad + \epsilon \\ &\text{모델 2: } F = \beta_0 + \qquad + \beta_2 P_2 + \epsilon \\ &\text{모델 3: } F = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \epsilon \end{aligned}$$

(b) 세 모형 각각에 대하여 $\beta_0 = 0$를 검정하여라.

(c) 개별적으로 $P_1$과 $P_2$ 중 어느 것이 $F$에 대하여 더 좋은 예측변수인가?

(d) 첫 번째와 두 번째 기초시험 성적이 78, 85점인 학생에 대하여 기말시험 성적을 예측하기 위하여 어떤 모형을 사용하는 것이 좋은가? 이 경우 예측값은 얼마인가?

Solve)

(a) 모델 1, 2, 3에 데이터를 적합시키기 위한, R코드는 다음과 같다.

```
model1<-lm(F~P1, data=model)
model1
model2<-lm(F~P2, data=model)
model2
model3<-lm(F~P1+P2, data=model)
model3
```

여기서 model은 표3.10의 데이터이고, model1, 2, 3는 각각 모델 1, 2, 3를 나타낸다. 위의 코드를 통해 모델 1, 2, 3에 적합된 회귀식은 다음과 같다.

모델 1: $F = -22.342 + 1.261 P_1$

모델 2: $F = -1.854 + 1.004 P_2$

모델 3: $F = -14.5005 + 0.4883 P_1 + 0.6720 P_2$

(b) 모델 1, 2, 3의 $\beta_0 = 0$에 필요한 $\beta_0, s.e.(\beta_0)$는 summary함수를 통해 구할 수 있다.
모델 1의 유의수준 5% 하에서 $\beta_0 = 0$ 검정은 다음과 같다.((c)의 summary(model1)참고)

$t_0 = \dfrac{-22.342}{11.5640} = -1.932030$ , $\beta_0 = -22.342, s.e.(\beta_0) = 11.5640$

```
> qt(0.975,20)
[1] 2.085963
```

$1.392 = |t_0| < t_{0.025}(20) = 2.085963$이므로 귀무가설을 기각할 수 없다.

$\therefore \beta_0 = 0$

모델 2의 유의수준 5% 하에서 $\beta_0 = 0$ 검정은 다음과 같다.((c)의 summary(model2)참고)

$t_0 = \frac{-1.85355}{7.56181} = -0.245119$ , $\beta_0 = -1.85355, s.e.(\beta_0) = 7.56181$

```
> qt(0.975,20)
[1] 2.085963
```

$0.245119 = |t_0| < t_{0.025}(20) = 2.085963$이므로 귀무가설을 기각할 수 없다.

$\therefore \beta_0 = 0$

모델 3의 유의수준 5% 하에서 $\beta_0 = 0$ 검정은 다음과 같다.((c)의 summary(model3)참고)

$t_0 = \frac{-14.5005}{9.2356} = 1.570065$ , $\beta_0 = -14.5005, s.e.(\beta_0) = 9.2356$

```
> qt(0.975,20)
[1] 2.085963
```

$1.570065 = t_0 < t_{0.025}(20) = 2.085963$이므로 귀무가설을 기각할 수 없다.

$\therefore \beta_0 = 0$

(c) $P_1$과 $P_2$ 중 어느 것이 $F$에 대한 더 좋은 설명변수인지 모델 1, 2에 적합된 회귀식의 결정계수 $R^2$의 값을 통해 알 수 있다. 다음은 모델 1, 2의 summary함수 결과이다.

```
> summary(model1)

Call:
lm(formula = F ~ P1, data = model)

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 .  
P1            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
```

```
> summary(model2)

Call:
lm(formula = F ~ P2, data = model)

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    
P2           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
```

위의 코드를 통해, 모델 1의 $R^2$이 0.8023, 모델 2의 $R^2$이 0.86임을 알 수 있다. 따라서 모델 2의 $R^2$이 더 크므로, $P_2$가 $F$에 대한 더 좋은 결정계수라고 할 수 있다.

(d) 모델 1, 2, 3의 결정계수$R^2$을 통해, 어떤 모형이 $F$를 가장 잘 설명하는지 알 수 있다. 다만, 모델 1, 2와 모델 3은 모형 안에 있는 예측변수의 수가 다르다는 것을 조정하여, 수정결정계수$R_a^2$를 이용하여 비교하겠다. 다음은 모델3의 summary함수 결과이다.

```
> summary(model3)

Call:
lm(formula = F ~ P1 + P2, data = model)

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   
P1            0.4883     0.2330   2.096  0.04971 * 
P2            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

(c)에서 구한 모델 1, 2의 summary함수 결과에서 $R_a^2$는 각각 0.7924, 0.853임을 쉽게 알 수 있다. 또한 위의 모델3의 summary함수 결과에서는 $R_a^2$이 0.8744임을 구하였다. 이를 통해, 모델 3의 $R_a^2$이 가장 크므로, $F$를 가장 잘 설명한다고 할 수 있다.
그러므로 모델 3의 회귀식 $F = -14.5005 + 0.4883P_1 + 0.6720P_2$을 사용하는 것이 좋다.
식에 $P_1 = 78, P_2 = 85$를 대입하여, 계산하면 다음과 같다.

$$\begin{aligned} F &= -14.5005 + 0.4883P_1 + 0.6720P_2 \\ &= -14.5005 + 0.4883*78 + 0.6720*85 \\ &= 80.7069 \end{aligned}$$

따라서 기초시험이 78, 85점인 학생의 기말고사 성적은 80.7069점으로 예측할 수 있다.

3.5 다음 회귀식들을 비교함으로써 단순과 다중회귀계수 사이의 관계를 살펴볼 수 있다.

$$
\begin{aligned}
\hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 \\
\hat{Y} &= \hat{\beta}_0' + \hat{\beta}_1' X_1 \\
\hat{Y} &= \hat{\beta}_0'' + \qquad + \hat{\beta}_2' X_2 \\
\hat{Y} &= \hat{\alpha}_0 + \qquad + \hat{\alpha}_2 X_2 \\
\hat{Y} &= \hat{\alpha}_0' + \hat{\alpha}_1 X_1
\end{aligned}
$$

표 3.10의 시험성적 데이터를 이용하여 다음을 보여라. 단, $Y = F$, $X_1 = P_1$, $X_2 = P_2$.

(a) $\hat{\beta}_1' = \hat{\beta}_1 + \hat{\beta}_2 \hat{\alpha}_1$, 즉 $X_1$에 대한 $Y$의 단순회귀계수는 $X_2$의 다중회귀계수에 $X_1$에 대한 $X_2$의 회귀계수를 곱한 후 $X_1$의 다중회귀계수를 더한 것과 같다.

(b) $\hat{\beta}_2' = \hat{\beta}_2 + \hat{\beta}_1 \hat{\alpha}_2$, 즉 $X_2$에 대한 $Y$의 단순회귀계수는 $X_1$의 다중회귀계수에 $X_2$에 대한 $X_1$의 회귀계수를 곱한 후 $X_2$의 다중회귀계수를 더한 것과 같다.

Solve)

(a) 위의 회귀식을 차례대로 적합한 R코드는 다음과 같다.(model은 시험성적 데이터이다.)

```
lm.satis1<-lm(F~P1+P2, data=model)
lm.satis2<-lm(F~P1, data=model)
lm.satis3<-lm(F~P2, data=model)
lm.satis4<-lm(P1~P2, data=model)
lm.satis5<-lm(P2~P1, data=model)
```

$X_1$에 대한 $Y$의 단순회귀계수가 $X_2$의 다중회귀계수에 $X_1$에 대한 $X_2$의 회귀계수를 곱한 후 $X_1$의 다중회귀계수를 더한 것과 같다는 것을 보이는 과정은 다음과 같다.

1. $X_1$에 대한 $Y$의 단순회귀계수

```
> lm.satis2$coefficients[2]
      P1
1.260516
```

2. $X_2$의 다중회귀계수

```
> lm.satis1$coefficients[3]
       P2
0.6720356
```

3. $X_1$에 대한 $X_2$의 회귀계수

```
> lm.satis5$coefficients[2]
      P1
1.149014
```

4. $X_1$의 다중회귀계수

```
> lm.satis1$coefficients[2]
       P1
0.4883376
```

5. $X_2$의 다중회귀계수*$X_1$에 대한 $X_2$의 회귀계수+$X_1$의 다중회귀계수 (2번*3번+4번)

```
> lm.satis1$coefficients[3]*lm.satis5$coefficients[2]+lm.satis1$coefficients[2]
      P2
1.260516
```

위의 과정을 통해 1번과 5번이 같음을 확인했으므로, (a)를 보였다고 할 수 있다.

(b) 마찬가지로 (a)에서 회귀식을 적합한 R코드를 이용해서 풀어보자.
$X_2$에 대한 $Y$의 단순회귀계수가 $X_1$의 다중회귀계수에 $X_2$에 대한 $X_1$의 회귀계수를 곱한 후 $X_2$의 다중회귀계수를 더한 것과 같다는 것을 보이는 과정은 다음과 같다.

1. $X_2$에 대한 $Y$의 단순회귀계수

```
> lm.satis3$coefficients[2]
      P2
1.004267
```

2. $X_1$의 다중회귀계수

```
> lm.satis1$coefficients[2]
      P1
0.4883376
```

3. $X_2$에 대한 $X_1$의 회귀계수

```
> lm.satis4$coefficients[2]
       P2
0.6803307
```

4. $X_2$의 다중회귀계수

```
> lm.satis1$coefficients[3]
       P2
0.6720356
```

5. $X_1$의 다중회귀계수*$X_2$에 대한 $X_1$의 회귀계수+$X_2$의 다중회귀계수 (2번*3번+4번)

```
> lm.satis1$coefficients[2]* lm.satis4$coefficients[2]+lm.satis1$coefficients[3]
       P1
1.004267
```

위의 과정을 통해 1번과 5번이 같음을 확인했으므로, (b)를 보였다고 할 수 있다.

3.7 표 3.12는 예측변수 $X_1$에 대응변수 $Y$를 관계시키는 단순회귀모형을 18개의 관측값에 적합시켰을 때의 회귀분석 결과를 보여준다(일부 수치는 삭제되었음). 13개의 삭제된 수치들을 채워 넣어라. 그리고 $Var(Y)$와 $Var(X_1)$을 계산하여라.

| 분산분석표 | | | | |
|---|---|---|---|---|
| 요인 | 제곱합 | 자유도 | 평균제곱 | $F$-검정 |
| 회귀 | (1)2174.419956 | (3) 1 | (5)2174.419956 | (7)40.338028 |
| 잔차 | (2)862.479424 | (4)16 | (6)53.904964 | |
| 회귀계수표 | | | | |
| 변수 | 계수 | 표준오차 | $t$-검정 | $p$-값 |
| 상수 | 3.43179 | (8)12.950150 | 0.265 | 0.7941 |
| $X_1$ | (9) 0.902428 | 0.1421 | (10)6.350655 | $< 0.0001$ |
| (11)$n = 18$ | $R^2 = 0.716$ | (12)$R_a^2 = 0.69825$ | $\hat{\sigma} = 7.342$ | (13)$d.f. = 16$ |

Solve)

(1): $SSR$로 위의 표에 $R^2$으로부터 구할 수 있다.

$R^2 = 1 - \dfrac{SSE}{SST} = 0.716$이므로, (2)의 $SSE = 862.479424$를 대입하면, $SST = 3036.89938$을 구할 수 있다. 여기서 $SST = SSR + SSE$이므로, $SSR = 2174.419956$를 구할 수 있다.

(2): $MSE = \dfrac{SSE}{n-2} = \hat{\sigma}^2 = 53.904964$임을 알 수 있다,

이 식에 $n = 18$을 대입하면, (2) $SSE = 862.479424$를 구할 수 있다.

(3): 예측변수가 1개이므로, $p = 1$이다.

(4): $n = 18, p = 1$이므로, $n - p - 1 = 16$이다.

(5): 평균회귀제곱$(MSR)$으로, $MSR = \dfrac{SSR}{p}$이다.

따라서 $SSR = 2174.41995, p = 1$이므로. $MSR = 2174.419956$이다.

(6): 평균오차제곱$(MSE)$으로, $MSE = \dfrac{SSE}{n-2} = \hat{\sigma}^2$이다.

따라서 $MSE = \hat{\sigma}^2 = (7.342)^2 = 53.904964$이다.

(7): $F = \dfrac{MSR}{MSE}$이므로, $F = \dfrac{2174.41995}{53.904964} = 40.338028$이다.

(8): $s.e.(\widehat{\beta_0})$로, 위의 상수$(\widehat{\beta_0})$의 $t$값을 통해 구할 수 있다.

$t_0 = \dfrac{\widehat{\beta_0}}{s.e(\widehat{\beta_0})} = 0.265$이므로, $\widehat{\beta_0} = 3.43179$를 대입하면, $s.e.(\widehat{\beta_0}) = 12.950150$를 구할 수 있다.

(9); $\widehat{\beta_1}$로 $\dfrac{\widehat{\beta_1}}{s.e.(\widehat{\beta_1})} = t_1$임을 이용하여 구할 수 있다. (10)의 $t_1$을 대입하면 다음과 같다.

$\dfrac{\widehat{\beta_1}}{0.1421} = 6.350655 \Leftrightarrow \widehat{\beta_1} = 0.902428$으로, 따라서 (9)는 $\widehat{\beta_1} = 0.902428$를 구할 수 있다.

(10): 단순선형회귀식에서는 $F = t_1^2$ 이므로, $\sqrt{40.330828} = 6.350655$이다.

(11): 18개의 관측값이므로, $n = 18$이다.

(12): 수정결정계수($R_a^2$)로 $R_a^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$ 이다.

$SSE, SST, n, p$를 각각 대입하면 $R_a^2 = 1 - \frac{862.479424/16}{3036.89938/17} = 1 - 0.30175 = 0.69825$이다.

(13): (4)와 같은 16이다.

다음은 $Var(Y)$와 $Var(X_1)$을 계산하는 과정이다.

$$1)\ Var(Y) = \frac{\sum_{n=1}^{18}(y_i - \overline{y})^2}{n-1} = \frac{SST}{n-1} = \frac{3036.89938}{17} = 178.64114$$

$$2)\ Var(X_1) = \frac{\sum_{i=1}^{18}(x_i - \overline{x})^2}{n-1}$$

$$= \frac{2669.565643}{17} \quad ,(\ (s.e.(\widehat{\beta_1}) = \frac{\widehat{\sigma^2}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \frac{7.342}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = 0.1421$$

$$\Leftrightarrow \sum_{i=1}^{n}(x_i - \overline{x})^2 = 2669.565643\ )$$

$$= 157.033273$$

따라서 $Var(Y) = 178.64114$, $Var(X_1) = 157.033273$이다.

3.10 감독자 직무능력평가 데이터를 이용하여, 다음의 각 모형에 대하여 $H_0 : \beta_1 = \beta_3 = 0.5$를 검정하여라.

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

Solve)
(a) RM과 FM은 다음과 같다.

$$(\mathrm{RM}) : \mathrm{Y} = \beta_0 + \beta_1' (\mathrm{X}_1 + \mathrm{X}_3) + \epsilon$$

$$\Leftrightarrow Y = \beta_0 + 0.5(X_1 + X_3) + \epsilon \ , \ \beta_1' = 0.5$$

$$\Leftrightarrow Y - 0.5(X_1 + X_3) = \beta_0 + \epsilon$$

$$\Leftrightarrow Y' = \beta_0 + \epsilon \qquad , Y' = Y - 0.5(X_1 + X_3)$$

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

위의 모형을 적합시킨 R코드는 다음과 같다.(P060은 감독자 직무능력평가 데이터이다.)

```
Yd<-P060$Y-0.5*P060$X1-0.5*P060$X3
lm.a.RM<-lm(Yd~1,data = P060)
lm.a.RM
lm.a.FM<-lm(Y~X1+X3,data = P060)
lm.a.FM
```

다음은 $H_0 : \beta_1 = \beta_3 = 0.5$에 대한 유의수준 5%에서 검정이다.

```
> summary(lm.a.FM)$r.squared
[1] 0.7080152
> summary(lm.a.RM)

Call:
lm(formula = Yd ~ 1, data = P060)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.650  -6.025  -0.650   4.975  14.350 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)     3.15       1.30   2.424   0.0218 *
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.119 on 29 degrees of freedom
> yhat<-3.15+0.5*P060$X1+0.5*P060$X3
> SST <- sum((P060$Y-mean(P060$Y))^2)
> SSE <- sum((P060$Y-yhat)^2)
> RM_r.squared<-1-(SSE/SST)
> SST
[1] 4296.967
> SSE
[1] 1469.575
> RM_r.squared<-1-(SSE/SST)
> RM_r.squared
[1] 0.6579971
>Fvals<-((summary(lm.a.FM)$r.squared-RM_r.squared)/2)/
((1-summary(lm.a.FM)$r.squared)/27)
> Fvals
[1] 2.3126
```

RM, FM의 결정계수를 이용하여 구한 $F = 2.3126$이다.

$2.3126 = F < F_{0.05}(2,27) = 3.3354131$이므로, 귀무가설을 기각할 수 없다.

$\therefore \beta_1 = \beta_3 = 0.5$

(b) RM과 FM은 다음과 같다.

$$(RM):\ Y = \beta_0 + \beta_1^{'}(X_1 + X_3) + \beta_2 X_2 + \epsilon$$

$$\Leftrightarrow Y = \beta_0 + 0.5(X_1 + X_3) + \beta_2 X_2 + \epsilon\ ,\ \beta_1^{'} = 0.5$$

$$\Leftrightarrow Y - 0.5(X_1 + X_3) = \beta_0 + \beta_2 X_2 + \epsilon$$

$$\Leftrightarrow Y^{'} = \beta_0 + \beta_2 X_2 + \epsilon,\ Y^{'} = Y - 0.5(X_1 + X_3)$$

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

위의 모형을 적합시키기 위한 R코드는 다음과 같다.

```
Yd<-P060$Y-0.5*P060$X1-0.5*P060$X3
lm.b.RM<-lm(Yd~X2,data = P060)
lm.b.RM
lm.b.FM<-lm(Y~X1+X2+X3,data = P060)
lm.b.FM
```

다음은 $H_0 : \beta_1 = \beta_3 = 0.5$에 대한 유의수준 5%에서 검정이다.

```
> summary(lm.b.FM)$r.squared
[1] 0.7150044
> summary(lm.b.RM)

Call:
lm(formula = Yd ~ X2, data = P060)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9466  -5.4598  -0.2339   5.7817  14.3003

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.3378     5.8687   1.591    0.123
X2           -0.1165     0.1077  -1.081    0.289

Residual standard error: 7.098 on 28 degrees of freedom
Multiple R-squared:  0.04007,	Adjusted R-squared:  0.005784
F-statistic: 1.169 on 1 and 28 DF,  p-value: 0.2889
> yhat<-9.3378+0.5*P060$X1-0.1165*P060$X2+0.5*P060$X3
> SST <- sum((P060$Y-mean(P060$Y))^2)
> SSE <- sum((P060$Y-yhat)^2)
> SST
[1] 4296.967
> SSE
[1] 1410.694
> RM_r.squared<-1-(SSE/SST)
> RM_r.squared
[1] 0.6717001
> Fvals<-((summary(lm.b.FM)$r.squared-RM_r.squared)/2)/
((1-summary(lm.b.FM)$r.squared)/26)
> Fvals
[1] 1.975318
```

RM, FM의 결정계수를 이용하여 구한 $F = 1.975318$이다.

$1.975318 = F < F_{0.05}(2,26) = 3.369016$이므로 귀무가설을 기각할 수 없다.

$\therefore \beta_1 = \beta_3 = 0.5$

3.11 연습문제 2.10과 표 2.11의 데이터를 참고하여 다음에 답하여라. 데이터는 교재의 웹 사이트에도 있다.

(a) 연습문제 2.10의 (f)에서 선택된 반응변수를 이용하여, 절편항과 기울기가 모두 0이라는 귀무가설을 검정하여라.
(b) 비슷한 키의 사람들끼리 결혼하는 경향이 있는지에 대한 검정으로 연습문제 2.10(g), 2.10(h) 3.11(a)의 가설검정 중에서 어느 것을 선택하겠는가? 결론은 무엇인가?
(c) 만일 비슷한 키의 사람들끼리 결혼하는 경향이 있는지에 대한 검정으로 위의 검정들이 적절하지 않다면 어떠한 다른 가설 검정을 할 것인가? 그리고 그 가설 검정을 기반으로 한 결론은 무엇인가?

Solve)

(a) 다음은 각각 반응변수를 남편, 아내로 했을 때의 summary함수의 결과이다.

```
sexheight<-lm(Husband~Wife,data = height)
sexheight1<-lm(Wife~Husband,data = height)
> summary(sexheight)

Call:
lm(formula = Husband ~ Wife, data = height)

Residuals:
    Min      1Q  Median      3Q     Max
-16.7438 -4.2838 -0.1615  4.2562 17.7500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.81005   11.93231   3.169  0.00207 **
Wife         0.83292    0.07269  11.458  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.468 on 94 degrees of freedom
Multiple R-squared:  0.5828,	Adjusted R-squared:  0.5783
F-statistic: 131.3 on 1 and 94 DF,  p-value: < 2.2e-16

> summary(sexheight1)

Call:
lm(formula = Wife ~ Husband, data = height)

Residuals:
```

```
      Min        1Q    Median        3Q       Max
-19.4685  -3.9208    0.8301    3.9538   11.1287

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) 41.93015    10.66162    3.933 0.000161 ***
Husband       0.69965     0.06106   11.458  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.928 on 94 degrees of freedom
Multiple R-squared:  0.5828,     Adjusted R-squared:  0.5783
F-statistic: 131.3 on 1 and 94 DF,  p-value: < 2.2e-16
```

위의 결과를 통해, 각각의 결정계수가 0.5828로 같다는 것을 알 수 있다. 즉 남편이 반응변수일 때, 아내가 반응변수일 때, 각각의 설명변수가 반응변수를 설명하는지에 대한 정도가 같다는 것을 의미한다. 따라서 남편과 아내 모두 반응변수가 될 수 있다고 볼 수 있다. 또한 남편과 아내가 각각 반응변수일 때, 적합된 모형에서 설명변수의 계수와 상수항의 p값이 서로 거의 차이가 없고, 모두 0.025($\alpha = 0.05$, 양측검정)보다 작으므로, 절편항과 기울기가 모두 0이라는 귀무가설을 기각하게 된다.

(b) 2.10(g)는 기울기가 0인가를 가설 검정하는 것으로, 귀무가설의 기각 여부가 비슷한 키의 사람들끼리 결혼하는 경향이 있는지에 대해 잘 설명하고 있지 못한다고 생각한다.
다음으로 2.10(h)의 절편항이 0인가에 대한 가설 검정이 단독으로 이루어진다면, 2.10(g)와 같이 적절하지 않은 검정이 될 것이라고 생각한다.
하지만 기울기가 1인가에 대한 가설 검정이 이루어진 후에, 2.10(h)의 검정이 진행한다면, 비슷한 키의 사람들끼리 결혼하는 경향에 대해 의미 있는 결과를 도출할 수 있을것이라 생각한다.
하지만 3.11(a)는 절편항과 기울기가 모두 0인지를 결정하는 것이기 때문에, 가설검정에서 귀무가설의 기각여부가 비슷한 키의 사람들이 결혼하는 경향이 있는지에 대해 잘 설명하지 못하고 있다고 생각한다.
그러므로 2.10(g). 2.10(h), 3.11(a)의 가설검정 모두 적절지 않다고 판단된다.

(c) (b)에서 미리 언급했듯이, 기울기가 1인가에 대한 검정이 가장 필요하다고 생각한다.
비슷한 키의 사람들끼리 결혼을 한다면, $\hat{Y} = \hat{\beta_0} + X$의 모형에 적합될 것이기 때문이다.
따라서 남편이 반응변수일 때, $H_0 : \beta_1 = 1 \; vs \; H_1 : \beta_1 \neq 1$에 대한 검정은 다음과 같다.

$$t_1 = \frac{\hat{\beta_1} - 1}{s.e.(\hat{\beta}_1)} = \frac{0.83292 - 1}{0.07269} = -2.298527, \quad \hat{\beta_1} = 0.83292, \; s.e.(\hat{\beta_1}) = 0.0.7269$$

```
> qt(0.975,94)
[1] 1.985523
```

위의 결과는 (a)의 풀이에서 summary(sexheight)의 결과를 이용하였다.

$2.298527 = |t_1| > t_{(0.025)}(94) = 1.985523$이므로, 귀무가설을 기각할 수 있다.

따라서 기울기가 1이라고 할 수 없으므로, 비슷한 키의 사람들끼리 결혼하는 경향 있다고 할 근거가 충분하지 않다는 결론을 얻을 수 있다.

3.13 표 3.14는 주어진 회사에서 근로자의 급여(salary)에 관한 다중회귀분석의 결과를 보여 준다. 여기에서 예측변수들은 다음과 같다.

| | |
|---|---|
| 성별(Sex) | 지시변수(1 = 남자, 0 = 여자) |
| 교육수준(Education) | 고용 당시의 교육 년수 |
| 경력(Experience) | 이전의 근무경력 월수 |
| 근무기간(Months) | 현 직장에서의 근무 월수 |

아래의 (a)-(b)에서 귀무가설과 대립가설, 사용된 검정법, 결론을 유의수준 5% 하에서 서술하여라.

(a) 회귀의 전반적 적합에 대한 $F$검정을 구축하여라.

(b) 변수 성별, 교육수준, 근무기간의 효과를 조정한 후, 급여와 경력 사이에 양(positive)의 선형관계가 존재하는가?

(c) 남자, 교육수준 12년, 경력 10개월, 근무기간 15개월인 어떤 사람의 급여는 얼마로 예측되는가?

(d) 남자, 교육수준 12년, 경력 10개월, 근무기간 15개월인 사람들의 평균적인 급여는 얼마로 예측되는가?

(e) 여자, 교육수준 12년, 경력 10개월, 근무기간 15개월인 사람들의 평균적인 급여는 얼마로 예측되는가?

[표 3.14] 네 개의 예측변수들에 대한 급여의 회귀로부터 얻은 결과

| 분산분석표 | | | | |
|---|---|---|---|---|
| 요인 | 제곱합 | 자유도 | 평균제곱 | $F$-검정 |
| 회귀 | 23665352 | 4 | 2916338 | 22.98 |
| 잔차 | 22657938 | 88 | 257477 | |
| 회귀계수표 | | | | |
| 변수 | 계수 | 표준오차 | t-검정 | $p$ -값 |
| 상수 | 3526.4 | 327.7 | 10.76 | 0.000 |
| 성별 | 722.5 | 117.8 | 6.13 | 0.000 |
| 교육수준 | 90.02 | 24.69 | 3.65 | 0.000 |
| 경력 | 1.2690 | 0.5877 | 2.16 | 0.034 |
| 근무기간 | 23.406 | 5.201 | 4.50 | 0.000 |
| $n = 93$ | $R^2 = 0.515$ | $R_a^2 = 0.489$ | $\hat{\sigma} = 507.4$ | $d.f. = 88$ |

Solve)

(a) 귀무가설과 대립가설은 다음과 같다.

$$(RM): H_0 : Y = \beta_0 + \epsilon$$

$$(FM): H_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

위 가설의 검정은 표3.14에 분산분석표의 $F$통계량을 이용해 $F$검정으로 구할 수 있다. $22.98 = F > F_{(0.05)}(4, 88) = 2.475277$이므로 귀무가설을 기각, $FM$을 사용하는 것이 적절하다고 할 수 있다. 그러므로 모든 회귀계수는 0이 아니다.

(b) $H_0 : \beta_0 = 0 \; vs \; H_1 : \beta_0 > 0$

$t_1 = \dfrac{1.269}{0.5877} = 2.16$, $t_{(0.05)}(88) = 1.66, t_1 > t_{(0.05)}(88)$이므로, $\alpha = 0.05$하에서, 귀무가설을 기각한다. 따라서, 급여와 경력 사이에 양의 선형관계가 존재한다고 할 수 있다.

(c) 남자, 교육수준 12년, 경력 10개월, 근무기간 15개월은 $X_1 = 1, X_2 = 12, X_3 = 10, X_4 = 15$로 나타낼 수 있다, 따라서 관측에 대한 예측값은 다음과 같다.

$$\hat{Y} = 3526.4 + 722.5X_1 + 90.02X_2 + 1.2690X_3 + 23.406X_4$$

$$= 3526.4 + 722.5 \cdot 1 + 90.02 \cdot 12 + 1.2690 \cdot 10 + 23.406 \cdot 15$$

$$= 5692.92$$

따라서 남자, 교육수준 12년, 경력 10개월, 근무기간 15개월인 어떤 사람의 급여는 5692.92로 예측된다.

(d) 남자, 교육수준 12년, 경력 10개월, 근무기간 15개월은 $X_1 = 1, X_2 = 12, X_3 = 10, X_4 = 15$로 나타낼 수 있다, 따라서 관측에 대한 평균값은 다음과 같다.

$$\hat{\mu} = 3526.4 + 722.5X_1 + 90.02X_2 + 1.2690X_3 + 23.406X_4$$

$$= 3526.4 + 722.5 \cdot 1 + 90.02 \cdot 12 + 1.2690 \cdot 10 + 23.406 \cdot 15$$

$$= 5692.92$$

따라서 남자, 교육수준 12년, 경력 10개월, 근무기간 15개월인 어떤 사람의 평균적인 급여는 5692.92로 예측된다.

(e) 여자, 교육수준 12년, 경력 10개월, 근무기간 15개월은 $X_1 = 0, X_2 = 12, X_3 = 10, X_4 = 15$로 나타낼 수 있다, 따라서 관측에 대한 평균값은 다음과 같다.

$$\hat{\mu} = 3526.4 + 722.5X_1 + 90.02X_2 + 1.2690X_3 + 23.406X_4$$

$$= 3526.4 + 722.5 \cdot 0 + 90.02 \cdot 12 + 1.2690 \cdot 10 + 23.406 \cdot 15$$

$$= 4970.42$$

따라서 여자, 교육수준 12년, 경력 10개월, 근무기간 15개월인 어떤 사람의 평균적인 급여는 4970.42로 예측된다.