Example. 16.1

Good

EXAMPLE 16.1 A Bivariate Analysis of Variance

Several members of a species of sparrow were collected at the same location at four different times of the year. Two variables were measured for each bird: the fat content (in grams) and the fat-free dry weight (in grams). For the statement of the null hypothesis, μ_{ij} denotes the population mean for variable i (where i is 1 or 2) and month j (i.e., j is 1, 2, 3, or 4).

H_0: $\mu_{11} = \mu_{12} = \mu_{13} = \mu_{14}$ *and* $\mu_{21} = \mu_{22} = \mu_{23} = \mu_{24}$

H_A: Sparrows do not have the same weight of fat *and* the same weight of fat-free dry body tissue at these four times of the year.

$\alpha = 0.05$

	December		January		February		March	
	Fat weight	*Lean dry weight*	*Fat weight*	*Lean dry weight*	*Fat weight*	*Lean dry weight*	*Fat weight*	*Lean dry weight*
	2.41	4.57	4.35	5.30	3.98	5.05	1.98	4.19
	2.52	4.11	4.41	5.61	3.48	5.09	2.05	3.81
	2.61	4.79	4.38	5.83	3.36	4.95	2.17	4.33
	2.42	4.35	4.51	5.75	3.52	4.90	2.00	3.70
	2.51	4.36			3.41	5.38	2.02	4.06
$\overline{X}$:	2.49	4.44	4.41	5.62	3.55	5.07	2.04	4.02

Computer computation yields the following output:

Wilks' $\Lambda = 0.0178$, $F = 30.3$, DF = 6, 28, $P \ll 0.0001$.

Reject H_0.

Pillai's trace = 1.0223, $F = 5.23$, DF = 6, 30, $P = 0.0009$.

Reject H_0.

Lawley-Hotelling trace = 52.9847, $F = 115$, DF = 6, 26, $P \ll 0.0001$.

Reject H_0.

Roy's maximum root = 52.9421, $F = 265$, DF = 3, 15, $P \ll 0.0001$.

Reject H_0.

R

CODE

```
#ex16.1
month=c(rep('Dec',5),rep('Jan',4),rep('Feb',5),rep('Mar',5))
fat=c(2.41,2.52,2.61,2.42,2.51,4.35,4.41,4.38,4.51,3.98,3.48,3.36,3.52,3.41,1.98,2.05,2.17,2.00,2.02)
dry=c(4.57,4.11,4.79,4.35,4.36,5.30,5.61,5.83,5.75,5.05,5.09,4.95,4.90,5.38,4.17,3.81,4.33,3.
```

```
70,4.06)
ex16.1=data.frame(month,fat,dry)
model=manova(cbind(fat, dry)~factor(month))
summary(model,test = 'Wilks')
summary(model,test = 'Pillai')
summary(model,test = 'Hotelling-Lawley')
summary(model,test = 'Roy')
```

OUTPUT

```
> summary(model,test = 'Wilks')
              Df    Wilks approx F num Df den Df    Pr(>F)    
factor(month)  3 0.017737   30.374      6     28 5.082e-11 ***
Residuals     15                                              
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(model,test = 'Pillai')
              Df Pillai approx F num Df den Df    Pr(>F)    
factor(month)  3 1.0249   5.2558      6     30 0.0008399 ***
Residuals     15                                            
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(model,test = 'Hotelling-Lawley')
              Df Hotelling-Lawley approx F num Df den Df    Pr(>F)    
factor(month)  3           52.973   114.78      6     26 < 2.2e-16 ***
Residuals     15                                                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(model,test = 'Roy')
              Df    Roy approx F num Df den Df    Pr(>F)    
factor(month)  3 52.928   264.64      3     15 3.301e-13 ***
Residuals     15                                            
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

SAS

CODE

```
data ex16_1;
input month$ fat dry @@;
cards;
D 2.41 4.57 D 2.52 4.11 D 2.61 4.79 D 2.42 4.35 D 2.51 4.36
J 4.35 5.30 J 4.41 5.61 J 4.38 5.83 J 4.51 5.75
F 3.98 5.05 F 3.48 5.09 F 3.36 4.95 F 3.52 4.90 F 3.41 5.38
M 1.98 4.19 M 2.05 3.81 M 2.17 4.33 M 2.00 3.70 M 2.02 4.06
; run;
proc glm data=ex16_1;
class month;
model fat dry  = month;
manova h = _all_ ;
run;
```

OUTPUT

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall month Effect
H = Type III SSCP Matrix for month
E = Error SSCP Matrix

S=2 M=0 N=6

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.01778151	30.33	6	28	<.0001
Pillai's Trace	1.02229007	5.23	6	30	0.0009
Hotelling-Lawley Trace	52.98465188	120.10	6	17	<.0001
Roy's Greatest Root	52.94208549	264.71	3	15	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

결과해석

12월, 1월, 2월, 3월에 참새들의 몸무게의 차이를 확인해 보려고 한다. 따라서 월 별로 같은 지역에서 참새들을 모아 순수 지방의 무게와 지방을 포함하지 않은 무게를 각각 참새별로 측정하였다. 따라서 종속변수는 지방의 무게와, 무지방 무게 2개가 되고 독립변수가 월이 되어 다변량 분산분석을 실시하였다.

귀무가설: 월별 참새들의 지방 무게의 모평균은 모두 같고 월별 참새들의 지방을 제외한 무게의 모평균 역시 모두 같다.

대립가설: 적어도 한 달의 참새들의 지방 무게의 모평균은 다른 달들과 다르고 적어도 한 달의 참새들의 지방을 제외한 무게의 모평균 역시 다른 달들과 다르다.

다변량 분산분석을 하는데 사용한 통계량으로 Wilks의 Lamda, Pillai의 Trace, Hotelling-Lawley의 Trace, Roy의 Maximum Root를 사용했고 분석 결과 모두 p-value가 매우 유의하게 나와 귀무가설을 기각할 충분한 근거가 된다. 따라서 월별로 참새들의 지방의 무게와 지방을 제외한 무게들의 모평균들이 모두 동일한 것은 아니다.

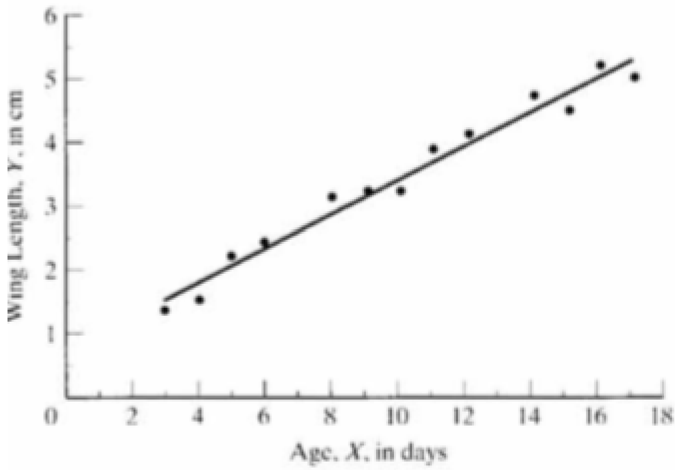
보통 이렇게 끝나지 않아요. 월별로 차이가 난다면 어느 달에 가장 높냐? 혹은 가장 낮냐? 차이가 있느냐? 등으로 질문을 받게 됩니다. 그런 부분을 설명하기 위한 통계량이 필요합니다.

Example. 17.1-17.4

EXAMPLE 17.1 Wing Lengths of 13 Sparrows of Various Ages. The Data Are Plotted in Figure 17.1.

Age (days) (X)	*Wing length* (cm) (Y)
3.0	1.4
4.0	1.5
5.0	2.2
6.0	2.4
8.0	3.1
9.0	3.2
10.0	3.2
11.0	3.9
12.0	4.1
14.0	4.7
15.0	4.5
16.0	5.2
17.0	5.0

$n = 13$



EXAMPLE 17.2 The Simple Linear Regression Equation Calculated (Using the "Machine Formula") by the Method of Least Squares, for the Data from the 13 Birds of Example 17.1

$n = 13$

$\sum X = 3.0 + 4.0 + \cdots + 17.0 = 130.0$

$\sum Y = 1.4 + 1.5 + \cdots + 5.0 = 44.4$

$\overline{X} = 130.0/13 = 10.0$

$\overline{Y} = 44.4/13 = 3.415$

$\sum X^2 = 3.0^2 + \cdots + 17.0^2 = 1562.00$

$\sum XY = (3.0)(1.4) + \cdots + (17.0)(5.0) = 514.80$

$$\sum x^2 = 1562.00 - \frac{(130.0)^2}{13} = 1562.00 - 1300.00 = 262.00$$

$$\sum xy = 514.80 - \frac{(130.0)(44.4)}{13} = 514.80 - 444.00 = 70.80$$

$$b = \frac{\sum xy}{\sum x^2} = \frac{70.80}{262.00} = 0.270 \text{ cm/day}$$

$$a = \overline{Y} - b\overline{X} = 3.415 \text{ cm} - (0.270 \text{ cm/day})(10.0 \text{ days})$$

$$= 3.415 \text{ cm} - 2.700 \text{ cm} = 0.715 \text{ cm}$$

So the simple linear regression equation is $\hat{Y} = 0.715 + 0.270X$.

EXAMPLE 17.3 Analysis of Variance Testing of H_0: $\beta = 0$ Against H_A: $\beta \neq 0$, Using the Data of Examples 17.1 and 17.2

$n = 13$

$\sum Y = 44.4$

$\sum Y^2 = 171.30$

$\sum xy = 70.80$ (from Example 17.2)

$\sum x^2 = 262.00$ (from Example 17.2)

$$\text{total SS} = \sum y^2 = 171.30 - \frac{(44.4)^2}{13} = 171.30 - 151.6431 = 19.656923$$

$$\text{regression SS} = \frac{\left(\sum xy\right)^2}{\sum x^2} = \frac{(70.80)^2}{262.00} = \frac{5012.64}{262.00} = 19.132214$$

total DF $= n - 1 = 12$

Source of variation	SS	DF	MS
Total	19.656923	12	
Linear regression	19.132214	1	19.132214
Residual	0.524709	11	0.047701

$$F = \frac{19.132214}{0.047701} = 401.1$$

$F_{0.05(1),1,11} = 4.84$

Therefore, reject H_0.

$$P \ll 0.0005 \quad [P = 0.00000000053]$$

$$r^2 = \frac{19.132214}{19.656923} = 0.97$$

$s_{Y \cdot X} = \sqrt{0.047701} = 0.218$ cm

EXAMPLE 17.4 Use of Student's t to Test H_0: $\beta = 0$ Against H_A: $\beta \neq 0$, Employing the Data of Examples 17.1 and 17.2

$n = 13$

$b = 0.270$ cm/day

$$s_b = \sqrt{\frac{s^2_{Y \cdot X}}{\sum x^2}} = \sqrt{\frac{0.047701}{262.00}} = \sqrt{0.00018206} = 0.0135 \text{ cm/day}$$

$$t = \frac{b - 0}{s_b} = \frac{0.270}{0.0135} = 20.000$$

$t_{0.05(2),11} = 2.201$

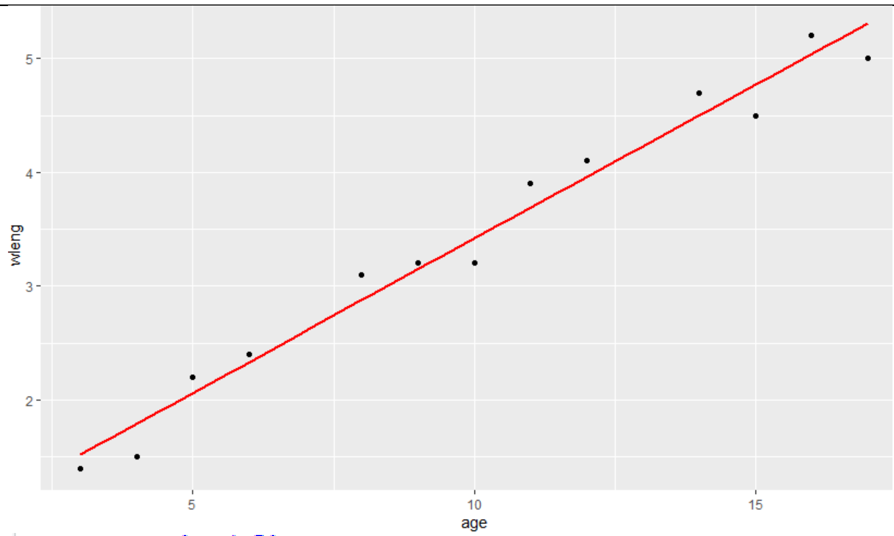
Therefore, reject H_0.

$$P \ll 0.01 \quad [P = 0.00000000027]$$

R	

CODE	#ex17.1-17.3
	age=c(3,4,5,6,8,9,10,11,12,14,15,16,17)
	wleng=c(1.4,1.5,2.2,2.4,3.1,3.2,3.2,3.9,4.1,4.7,4.5,5.2,5)
	ex17.1=data.frame(age, wleng)
	library(ggplot2)
	model=lm(wleng~age)
	ggplot(ex17.1,aes(x=age,y=wleng),
	xlim=c(0,18))+geom_point()+stat_smooth(method = 'lm', se=F, color='red')
	summary(model)
	par(mfrow=c(2,2))
	plot(model)

OUTPUT



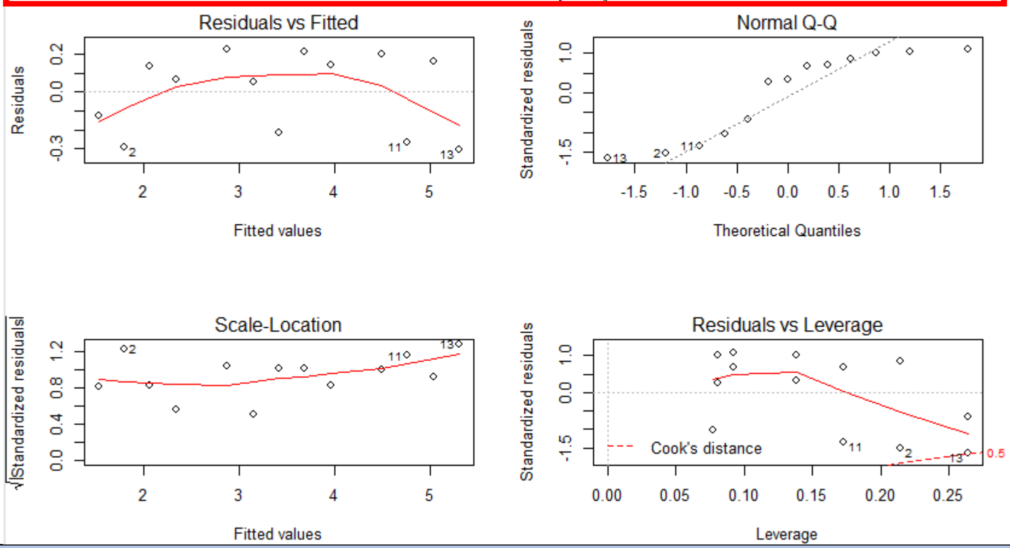

```
> summary(model)

Call:
lm(formula = wleng ~ age)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.30699 -0.21538  0.06553  0.16324  0.22507 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.71309    0.14790   4.821 0.000535 ***
age          0.27023    0.01349  20.027 5.27e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2184 on 11 degrees of freedom
Multiple R-squared:  0.9733,	Adjusted R-squared:  0.9709 
F-statistic: 401.1 on 1 and 11 DF,  p-value: 5.267e-10
```



SAS

CODE

```
data ex17_1;
input age length @@;
cards;
3 1.4 4 1.5 5 2.2 6 2.4 8 3.1 9 3.2
```

```
10 3.2 11 3.9 12 4.1 14 4.7 15 4.5 16 5.2 17 5
;run;
proc reg data=ex17_1;
model length=age;
run;
```

OUTPUT

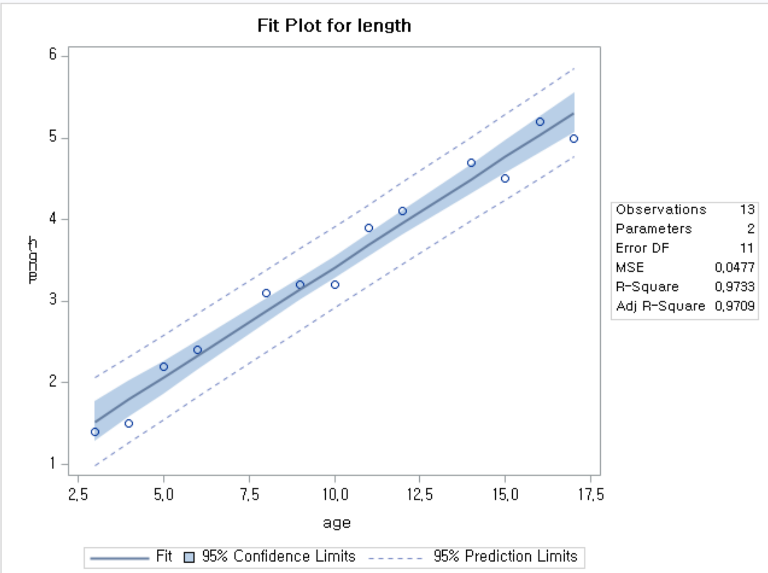
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	19.13221	19.13221	401.09	<.0001
Error	11	0.52471	0.04770		
Corrected Total	12	19.65692			

Root MSE	0.21841	R-Square	0.9733
Dependent Mean	3.41538	Adj R-Sq	0.9709
Coeff Var	6.39475		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.71309	0.14790	4.82	0.0005
age	1	0.27023	0.01349	20.03	<.0001



결과해석

13마리의 참새들의 나이와 날개 길이에 관계가 있는지 확인해보려고 한다. 먼저 13개의 자료를 이용하여 산점도를 그려봤더니 직선 형태를 나타내고 있는 것으로 보인다. 따라서 종속변수를 날개길이, 독립변수를 참새의 나이로 해서 회귀모형에 적합시켜보았다. 회귀식은 y=0.713+0.270x로 적합되며 절편항과 기울기 계수에 대한 검정은 매우 p-value가 매우 유의하게 나와 β=0이라는 귀무가설을 기각한다. 또한 결정계수는 0.97정도로 총변동중에서 적합된 회귀직선에 의해 설명되는 부분이 97%정도가 된다. 그리고 기울기 계수가 0.27로 age가 한 단위 증가할 때 날개길이는 0.27만큼 증가한다는 것을 알 수 있다.

Example. 17.5

EXAMPLE 17.5 Standard Errors of Predicted Values of Y

The regression equation derived in Example 17.2 is used for the following considerations. For this regression, $a = 0.72$ cm, $b = 0.270$ cm/day, $\overline{X} = 10.0$ days, $\sum x^2 = 262.00$ days^2, $n = 13$, $s^2_{Y \cdot X} = 0.047701$ cm^2, and $t_{0.05(2),11} = 2.201$.

a. Equation 17.26 is used when we wish to predict the mean value of $\hat{Y}_i$, given X_i, in the entire population. For example, we could ask, "What is the mean wing length of all 13.0-day-old birds in the population under study?"

$$\hat{Y}_i = a + bX_i$$
$$= 0.715 + (0.270)(13.0)$$
$$= 0.715 + 3.510$$
$$= 4.225 \text{ cm}$$

$$s_{\hat{Y}_i} = \sqrt{s^2_{Y \cdot X}\left[\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum x^2}\right]}$$
$$= \sqrt{0.047701\left[\frac{1}{13} + \frac{(13.0 - 10.0)^2}{262.00}\right]}$$
$$= \sqrt{(0.047701)(0.111274)}$$
$$= 0.073 \text{ cm}$$

$$95\% \text{ confidence interval} = \hat{Y}_i \pm t_{0.05(2),11}s_{\hat{Y}_i}$$
$$= 4.225 \pm (2.201)(0.073)$$
$$= 4.225 \pm 0.161 \text{ cm}$$
$$L_1 = 4.064 \text{ cm}$$
$$L_2 = 4.386 \text{ cm}$$

b. Equation 17.28 is used when we propose taking an additional sample of m individuals from the population and wish to predict the mean Y value, at a given X, for these m new data. For example, we might ask, "If ten 13.0-day-old birds were taken from the population, what would be their mean wing length?"

	$\hat{Y}_i = 0.715 + (0.270)(13.0) = 4.225$ cm $$(s_{\hat{Y}_i})_{10} = \sqrt{0.047701\left[\frac{1}{10} + \frac{1}{13} + \frac{(13.0 - 10.0)^2}{262.00}\right]}$$ $$= \sqrt{(0.047701)(0.211274)}$$ $$= 0.100 \text{ cm}$$ $$95\% \text{ prediction interval} = \hat{Y}_i \pm t_{0.05(2),11}(s_{\hat{Y}_i})_{10}$$ $$= 4.225 \pm (2.201)(0.100)$$ $$= 4.225 \pm 0.220 \text{ cm}$$ $$L_1 = 4.005 \text{ cm}$$ $$L_2 = 4.445 \text{ cm}$$ c. Equation 17.29 is used when we wish to predict the Y value of a single observation taken from the population as a specified X. For example, we could ask, "If one 13.0-day-old bird were taken from the population, what would be its wing length?" $\hat{Y}_i = 0.715 + (0.270)(13.0) = 4.225$ cm $$(s_{\hat{Y}_i})_1 = \sqrt{0.047701\left[1 + \frac{1}{13} + \frac{(13.0 - 10.0)^2}{262.00}\right]}$$ $$= \sqrt{(0.047701)(1.111274)}$$ $$= 0.230 \text{ cm}$$ $$95\% \text{ prediction interval} = \hat{Y}_i \pm t_{0.05(2),11}(s_{\hat{Y}_i})_1$$ $$= 4.225 \pm (2.201)(0.230)$$ $$= 4.225 \pm 0.506 \text{ cm}$$ $$L_1 = 3.719 \text{ cm}$$ $$L_2 = 4.731 \text{ cm}$$ Note from these three examples that the accuracy of prediction increases as does the number of data upon which the prediction is based. For example, predictions about a mean for the entire population will be more accurate than a prediction about a mean from 10 members of the population, which is more accurate than a prediction about a single member of the population.
R	

CODE	#ex17.5
	age=c(3,4,5,6,8,9,10,11,12,14,15,16,17)
	wleng=c(1.4,1.5,2.2,2.4,3.1,3.2,3.2,3.9,4.1,4.7,4.5,5.2,5)
	model=lm(wleng~age)
	ex17.5=data.frame(age, wleng)
	newdata=data.frame(age=13)
	predict(model, newdata, interval = "confidence")
	predict(model, newdata, interval = "predict", weights=10)
	predict(model, newdata, interval = "predict")

OUTPUT	> predict(model, newdata, interval = "confidence") fit lwr upr 1 4.226072 4.065719 4.386425 > predict(model, newdata, interval = "predict", weights=10) fit lwr upr 1 4.226072 4.005117 4.447026 > predict(model, newdata, interval = "predict") fit lwr upr 1 4.226072 3.719325 4.732818

SAS	
CODE	(see code below)

```
data ex17_5;
input age length @@;
cards;
3 1.4 4 1.5 5 2.2 6 2.4 8 3.1 9 3.2
10 3.2 11 3.9 12 4.1 14 4.7 15 4.5 16 5.2 17 5
;run;
data newdata;
input age;
cards;
13
;run;
data pred;
set ex17_5 newdata;
run;;
proc reg data=pred;
model length=age/p cli clm;
run;
```

OUTPUT

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	1.4	1.5238	0.1122	1.2768	1.7707	0.9833	2.0642	-0.1238
2	1.5	1.7940	0.1011	1.5715	2.0166	1.2643	2.3237	-0.2940
3	2.2	2.0642	0.0907	1.8647	2.2638	1.5438	2.5847	0.1358
4	2.4	2.3345	0.0811	2.1559	2.5130	1.8217	2.8473	0.0655
5	3.1	2.8749	0.0663	2.7290	3.0209	2.3726	3.3773	0.2251
6	3.2	3.1452	0.0621	3.0086	3.2817	2.6454	3.6449	0.0548
7	3.2	3.4154	0.0606	3.2821	3.5487	2.9165	3.9142	-0.2154
8	3.9	3.6856	0.0621	3.5490	3.8222	3.1859	4.1853	0.2144
9	4.1	3.9558	0.0663	3.8099	4.1018	3.4535	4.4582	0.1442
10	4.7	4.4963	0.0811	4.3177	4.6749	3.9835	5.0091	0.2037
11	4.5	4.7665	0.0907	4.5670	4.9661	4.2460	5.2870	-0.2665
12	5.2	5.0368	0.1011	4.8142	5.2593	4.5070	5.5665	0.1632
13	5.0	5.3070	0.1122	5.0600	5.5540	4.7666	5.8474	-0.3070
14	.	4.2261	0.0729	4.0657	4.3864	3.7193	4.7328	.

결과해석

a. 예제 17.1에서 적합된 회귀식에서 모집단에서 age=13인 새들의 평균 날개길이의 예측구간을 구했을 때 [4.064cm, 4.386cm]로 예측할 수 있다.

b. 예제 17.1에서 적합된 회귀식에서 모집단에서 추가로 age=13인 10마리의 참새들을 뽑았을 때 10마리의 평균 날개길이의 예측구간은 [4.005cm, 4.445cm]로 예측할 수 있다.

c. 예제 17.1에서 적합된 회귀식에서 모집단에서 추가로 age=13인 한 마리의 참새를 뽑았을 때 해당 참새의 평균 날개길이의 예측구간은 [3.719cm, 4.731cm]로 예측할 수 있다.

Example. 17.6

EXAMPLE 17.6 Hypothesis Testing with an Estimated Y Value

H_0: The mean population wing length of 13.0-day-old birds is not greater than 4 cm (i.e., H_0: $\mu_{\hat{Y}_{13.0}} \le 4$ cm).

H_A: The mean population wing length of 13.0-day-old birds is greater than 4 cm (i.e., H_A: $\mu_{\hat{Y}_{13.0}} > 4$ cm).

From Example 17.5b, $\hat{Y}_{13.0} = 4.225$ cm and $s_{\hat{Y}_{13.0}} = 0.073$ cm,

$$t = \frac{4.225 - 4}{0.073} = \frac{0.225}{0.073} = 3.082$$

$$t_{0.05(1),11} = 1.796$$

Therefore, reject H_0.

$$0.005 < P < 0.01 \quad [P = 0.0052]$$

R

CODE

```
#ex17.6
y=4.225
s=0.073
t=(y-4)/s
ct=qt(0.05,11, lower.tail = F)
if (t>ct){
   answer=("reject H0")
}
print(answer)
pt(t,11,lower.tail = F)
```

OUTPUT

```
> pt(t,11,lower.tail = F)
[1] 0.005215209
```

SAS

CODE

```
proc iml;
y=4.225;
s=0.073;
t=(y-4)/s;
ct=quantile("t", 0.95,11);
if abs(t) > ct then answer= "reject H0";
else answer= "cannot reject H0";
p=(1-probt(abs(t),11));
print answer p; run;
```

OUTPUT

answer	p
reject H0	0.0052152

결과해석

Age=13인 참새들의 날개길이의 모평균이 4보다 큰지 확인해보기 위해 가설검정을 실시하였다.

귀무가설: age=13인 참새들의 날개길이의 모평균은 4보다 작거나 같다.

대립가설: age=13인 참새들의 날개길이의 모평균은 4보다 크다.

	검정결과 p-value=0.005로 귀무가설을 기각하기에 충분한 증거가 된다. 따라서 age=13인 참새들의 날개길이의 모평균은 4보다 큰 것을 알 수 있다. 이는 예제 17.5의 a의 예측구간이 [4.064cm, 4.386cm]로 하한값도 4보다 크다는 것으로도 확인할 수 있다.

Example. 17.7

EXAMPLE 17.7 Inverse Prediction

We wish to estimate, with 95% confidence, the age of a bird with a wing length of 4.5 cm.

Predicted age:

$$\hat{X} = \frac{Y_i - a}{b} = \frac{4.5 - 0.715}{0.270} = 14.019 \text{ days}$$

To compute 95% confidence interval:

$t = t_{0.05(2),11} = 2.201$

$$K = b^2 - t^2 s_b^2 = 0.270^2 - (2.201)^2(0.0135)^2 = 0.0720$$

95% confidence interval:

$$\bar{X} + \frac{b(Y_i - \bar{Y})}{K} \pm \frac{t}{K}\sqrt{s_{Y \cdot X}^2\left[\frac{(Y_i - \bar{Y})^2}{\sum x^2} + K\left(1 + \frac{1}{n}\right)\right]}$$

$$= 10.0 + \frac{0.270(4.5 - 3.415)}{0.0720} \pm \frac{2.201}{0.0720}\sqrt{0.047701\left[\frac{(4.5 - 3.415)^2}{262.00} + 0.0720\left(1 + \frac{1}{13}\right)\right]}$$

$$= 10.0 + 4.069 \pm 30.569\sqrt{0.003913}$$

$$= 14.069 \pm 1.912 \text{ days}$$

$L_1 = 12.157$ days

$L_2 = 15.981$ days

R	

CODE	#ex17.7
	age=c(3,4,5,6,8,9,10,11,12,14,15,16,17)
	wleng=c(1.4,1.5,2.2,2.4,3.1,3.2,3.2,3.9,4.1,4.7,4.5,5.2,5)

	model=lm(wleng~age) library(chemCal) inverse.predict(model, 4.5)

OUTPUT

```
> inverse.predict(model, 4.5)
$Prediction
[1] 14.01369

$`Standard Error`
[1] 0.862344

$Confidence
[1] 1.898006

$`Confidence Limits`
[1] 12.11568 15.91170
```

SAS

CODE

```
proc iml;
xbar=10;
ybar=3.415;
sxx=262;
syx=0.047701;
a= 0.715;
s=0.270;
sb=0.0135;
t= 2.201;
X = (4.5-a) / s;
K = s**2 - (t**2 * sb**2);
H = (t / K) * sqrt (syx * (K * (1+(1/13)) + ((4.5- ybar)**2) /
sxx));
LL = xbar + ((s * (4.5 - ybar)) / K) - H;
UL = xbar + ((s * (4.5 - ybar)) / K) + H;
print K LL UL;
run; quit;
```

OUTPUT

K	LL	UL
0.0720171	12.155784	15.979783

결과해석

앞의 예제들과는 반대로 날개길이가 4.5일 때의 참새의 age, 즉 한 종속변수 값에 대한 독립변수의 예측 값에 대한 신뢰구간을 구해보았다. 구해본 결과 95%신뢰구간은 [12.157,15.981]로 날개길이가 4.5cm이면 참새들의 평균 age는 12일에서 16일 가까이 되는 것으로 예측할 수 있다.

Example. 17.8

EXAMPLE 17.8a Regression Data Where There Are Multiple Values of Y for Each Value of X

	Age (yr)	Systolic blood pressure (mm Hg)		
i	X_i	Y_{ij}	n_i	$\overline{Y}_i$
1	30	108, 110, 106	3	108.0
2	40	125, 120, 118, 119	4	120.5
3	50	132, 137, 134	3	134.3
4	60	148, 151, 146, 147, 144	5	147.2
5	70	162, 156, 164, 158, 159	5	159.8

$k = 5; \quad i = 1 \text{ to } 5; \quad j = 1 \text{ to } n_i; \quad N = 20$

$\sum\sum X_{ij} = 1050 \qquad \sum\sum Y_{ij} = 2744$

$\sum\sum X_{ij}^2 = 59{,}100 \qquad \sum\sum Y_{ij}^2 = 383{,}346 \qquad \sum\sum X_{ij}Y_{ij} = 149{,}240$

$\sum x^2 = 3975.00 \qquad \sum y^2 = 6869.20 \qquad \sum xy = 5180.00$

$\overline{X} = 52.5 \qquad \overline{Y} = 137.2$

$$b = \frac{\sum xy}{\sum x^2} = \frac{5180.00}{3975.00} = 1.303 \text{ mm Hg/yr}$$

$$a = \overline{Y} - b\overline{X} = 137.2 - (1.303)(52.5) = 68.79 \text{ mm Hg}$$

Therefore, the least-squares regression line is $\hat{Y}_{ij} = 68.79 + 1.303X_{ij}$.

EXAMPLE 17.8b Statistical Analysis of the Regression Data of Example 17.8a

H_0: The population regression is linear.

H_A: The population regression is not linear.

total SS $= \sum y^2 = 6869.20$ total DF $= N - 1 = 19$

$$\text{among-groups SS} = \sum_{i=1}^{k} \frac{\left(\sum_{j=1}^{n_i} Y_{ij}\right)^2}{n_i} - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}\right)^2}{N}$$

$$= 383{,}228.73 - 376{,}476.80 = 6751.93$$

among-groups DF $= k - 1 = 4$

within-groups SS $=$ total SS $-$ among-groups SS

$= 6869.20 - 6751.93 = 117.27$

within-groups DF $=$ total DF $-$ among-groups DF

$= 19 - 4 = 15$

deviations-from-linearity SS $=$ among-groups SS $-$ regression SS

$= 6751.93 - 6750.29 = 1.64$

deviations-from-linearity DF $=$ among-groups DF $-$ regression DF

$= 4 - 1 = 3$

Source of variation	SS	DF	MS
Total	6869.20	19	
Among groups	6751.93	4	
Linear regression	6750.29	1	
Deviations from linearity	1.64	3	0.55
Within groups	117.27	15	7.82

$$F = \frac{0.55}{7.82} = 0.070$$

Since $F < 1.00$, do not reject H_0.

$$P > 0.25 \quad [P = 0.975]$$

H_0: $\beta = 0$.

H_A: $\beta \neq 0$.

$$\text{regression SS} = \frac{\left(\sum xy\right)^2}{\sum x^2} = \frac{(5180.00)^2}{3975.00} = 6750.29$$

Source of variation	SS	DF	MS
Total	6869.20	19	
Linear regression	6750.29	1	6750.29
Residual	118.91	18	6.61

$$F = \frac{6750.29}{6.61} = 1021.2$$

$F_{0.05(1),1,18} = 4.41$

Therefore, reject H_0.

$$P \ll 0.0005 \quad [P < 0.00000000001]$$

$$r^2 = \frac{6750.29}{6869.20} = 0.98$$

$s_{Y \cdot X} = \sqrt{6.61} = 2.57$ mm Hg

R	

CODE	#ex17.8
	x=c(30,30,30,40,40,40,40,50,50,50,60,60,60,60,60,70,70,70,70,70)
	y=c(108,110,106,125,120,118,119,132,137,134,148,151,146,147,144,162,156,164,158,159)
	model=lm(y~x)
	summary(model)

OUTPUT

```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.0050 -1.9186 -0.4421  2.0264  4.0893 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 68.78491    2.21607   31.04   <2e-16 ***
x            1.30314    0.04077   31.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.57 on 18 degrees of freedom
Multiple R-squared:  0.9827,    Adjusted R-squared:  0.9817 
F-statistic:  1022 on 1 and 18 DF,  p-value: < 2.2e-16
```

https://stackoverflow.com/questions/45364277/using - r - for - lack - of - fit - f - test

SAS

CODE

```
data ex17_8;
input x y @@;
cards;
30 108 30 110 30 106 40 125
40 120 40 118 40 119 50 132
50 137 50 134 60 148 60 151
60 146 60 147 60 144 70 162
70 156 70 164 70 158 70 159
;run;
proc rsreg data=ex17_8;
model y=x/ lackfit;
run;
```

proc reg에서 lackfit 옵션을 주면 됩니다.

OUTPUT

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	67.329545	9.045263	7.44	<.0001	134.064685
x	1	1.365224	0.375737	3.63	0.0021	26.083625
x*x	1	-0.000610	0.003672	-0.17	0.8699	-0.244172

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F
x	2	6750.482343	3375.241171	483.32	<.0001

결과해석

사람들의 연령에 따른 수축기혈압에 대한 회귀식을 적합시켰을 때 적합된 직선회귀식이 자료를 적합시키는데 적절한지 확인해보기 위해 연령별로 수축기혈압을 반복측정을 하여 적합결여검정을 실시하였다.

귀무가설: 적합된 직선회귀식은 적절하다.

대립가설: 적합된 직선회귀식은 적절하지 않다.

검정결과 p-value는 매우 유의한 값이 나와 귀무가설을 기각한다. 따라서 적합된 직선회귀식은 적절하지 않다고 판단할 수 있다. 따라서 잔차의 산점도를 그려 어떤 형태의 회귀식이 적절한지 판단해보고, 적절한 변수변환을 통해 직선회귀식으로 적합시킬 수 있도록 하는 조치가 필요하다.

Example. 17.9

EXAMPLE 17.9 Regression Data Before and After Logarithmic Transformation of Y

Original data (as plotted in Figure 17.11a), indicating the variance of Y (namely, s_Y^2) at each X:

X	Y	s_Y^2
5	10.72, 11.22, 11.75, 12.31	0.4685
10	14.13, 14.79, 15.49, 16.22	0.8101
15	18.61, 19.50, 20.40, 21.37	1.4051
20	24.55, 25.70, 26.92, 28.18	2.4452
25	32.36, 33.88, 35.48, 37.15	4.2526

Transformed data (as plotted in Figure 17.11b), indicating the variance of log Y (namely, $s_{\log Y}^2$) at each X:

X	$\log Y$	$s_{\log Y}^2$
5	1.03019, 1.04999, 1.07004, 1.09026	0.000668
10	1.15014, 1.16997, 1.19005, 1.21005	0.000665
15	1.26975, 1.29003, 1.30963, 1.32980	0.000665
20	1.39005, 1.40993, 1.43008, 1.44994	0.000665
25	1.51001, 1.52994, 1.54998, 1.56996	0.000666

R

CODE

```
#ex17.9
x=c(5,5,5,5,10,10,10,10,15,15,15,15,20,20,20,20,25,25,25,25)
y=c(10.72,11.22,11.75,12.31,14.13,14.79,15.49,16.22,18.61,
     19.50,20.40,21.37,24.55,25.70,26.92,28.18,32.36,33.88,35.48,37.15)
model=lm(y~x)
summary(model)
logy=log10(y)
logmodel=lm(logy~x)
summary(logmodel)
```

OUTPUT	see below

```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9265 -1.2519 -0.2257  0.9631  4.0905 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.25200    0.98115   4.334    4e-04 ***
x            1.15230    0.05917  19.476 1.52e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.871 on 18 degrees of freedom
Multiple R-squared:  0.9547,	Adjusted R-squared:  0.9522 
F-statistic: 379.3 on 1 and 18 DF,  p-value: 1.523e-13
> summary(logmodel)

Call:
lm(formula = logy ~ x)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.030244 -0.015017 -0.000158  0.015043  0.030198 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.940095   0.012355   76.09   <2e-16 ***
x           0.023993   0.000745   32.20   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02356 on 18 degrees of freedom
Multiple R-squared:  0.9829,	Adjusted R-squared:  0.982 
F-statistic:  1037 on 1 and 18 DF,  p-value: < 2.2e-16
```

SAS

CODE

```
data ex17_9;
input x y @@;
cards;
5 10.72 5 11.22 5 11.75 5 12.31
10 14.13 10 14.79 10 15.49 10 16.22
15 18.61 15 19.5 15 20.4 15 21.37
20 24.55 20 25.7 20 26.92 20 28.18
25 32.36 25 33.88 25 35.48 25 37.15
; run;
proc rsreg data=ex17_9;
model y=x/ lackfit;
run;
data ex17_9_2;
set ex17_9;
logy= log10(y);
run;
proc rsreg data=ex17_9_2;
model logy=x/ lackfit;
run;
```

OUTPUT

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	9.752000	1.387010	7.03	<.0001	19.965071
x	1	0.209443	0.211399	0.99	0.3357	11.523000
x*x	1	0.031429	0.006913	4.55	0.0003	3.142857

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	0.940359	0.025997	36.17	<.0001	1.299915
x	1	0.023948	0.003962	6.04	<.0001	0.239930
x*x	1	0.000001508	0.000130	0.01	0.9909	0.000151

결과해석

회귀계수에 대한 추정이나 적합결여검정을 통해 유의하지 않은 항이나, 잘못 적합된 회귀식에 대해 로그변환과 같은 변수변환 방법을 사용한 후 회귀식을 적합시키면 유의한 결과를 얻을 수 있다.