



총정리

정보통계학과 김성수교수

단순회귀모형

(예제) 표본상점의 광고료와 총판매액 자료에 대하여 회귀직선을 구하고, 산점도 위에 회귀직선을 그려보아라.

```
> market.lm = lm(Y ~ X, data=market)
```

```
> summary(market.lm)
```

Call:

```
lm(formula = Y ~ X, data = market)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2.02908	-1.35349	-0.05685	0.98903	2.51517

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.3282	1.4302	0.229	0.822
X	2.1497	0.1548	13.889	3.55e-09 ***

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

Residual standard error: 1.587 on 13 degrees of freedom

Multiple R-squared: 0.9369, Adjusted R-squared: 0.932

F-statistic: 192.9 on 1 and 13 DF, p-value: 3.554e-09

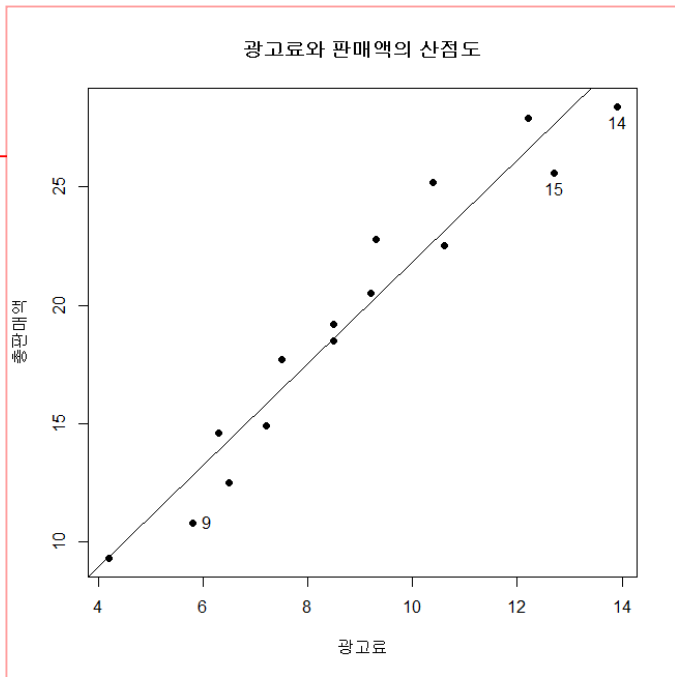
추정된 회귀식

$$\hat{Y} = 0.3282 + 2.1497 X$$

단순회귀모형

(예제) 표본상점의 광고료와 총판매액 자료에 대하여 회귀직선을 구하고, 산점도 위에 회귀직선을 그려보아라.

```
> plot(market$X, market$Y, xlab="광고료", ylab="총판매액", pch=19)  
> title("광고료와 판매액의 산점도")  
> abline(market.lm)  
> identify(market$X, market$Y)  
[1] 9 14 15
```



단순회귀모형 : 분산분석표

```
> market.lm = lm(Y ~ X, data=market)
```

```
> anova(market.lm)
```

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	1	485.57	485.57	192.9	3.554e-09 ***
Residuals	13	32.72	2.52		

분산분석 결과 해석 : $p\text{-값}=3.554 \times 10^{-9}$ 로 매우 작은 값이므로 $H_0 : \beta_1 = 0$ 을 기각.

단순회귀모형 : 결정계수, 추정값 표준오차

```
> market.lm = lm(Y ~ X, data=market)
```

```
> anova(market.lm)
```

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
X	1	485.57	485.57	192.9	3.554e-09 ***
Residuals	13	32.72	2.52		

$$\Rightarrow R^2 = 485.57 / (485.57 + 32.72) = 0.9369$$

이는 총변동 중에서 회귀직선에 의하여 설명되는 부분이 94%라는 의미로서, 추정된 회귀선의 정도가 높다는 것을 알 수 있음.

추정값 표준오차 $\Rightarrow S_{Y \cdot X} = \sqrt{MSE} = \sqrt{2.52} = 1.587$

단순회귀모형 : β_1 , β_0 신뢰구간 구하기

```
> market.lm = lm(Y ~ X, data=market)
```

```
> summary(market.lm)
```

```
...
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.3282	1.4302	0.229	0.822
X	2.1497	0.1548	13.889	3.55e-09 ***

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

Residual standard error: 1.587 on 13 degrees of freedom

Multiple R-squared: 0.9369, Adjusted R-squared: 0.932

F-statistic: 192.9 on 1 and 13 DF, p-value: 3.554e-09

β_1 의 95% 신뢰구간

```
> q.val = qt(0.975,13)
> 2.1497 - q.val*0.1548
[1] 1.815275
> 2.1497 + q.val*0.1548
[1] 2.484125
```

β_0 의 95% 신뢰구간

```
> q.val = qt(0.975,13)
> 0.3282 - q.val*1.4302
[1] -2.761559
> 0.3282 + q.val*1.4302
[1] 3.417959
```

단순회귀모형 : β_1 검정

```
> market.lm = lm(Y ~ X, data=market)
> summary(market.lm)
```

```
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.3282     1.4302    0.229   0.822
X              2.1497     0.1548   13.889 3.55e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.587 on 13 degrees of freedom
Multiple R-squared:  0.9369,    Adjusted R-squared:  0.932
F-statistic: 192.9 on 1 and 13 DF,  p-value: 3.554e-09
```

기각역 및 p-값 구하기

```
> # 유의수준 0.05 기각역
> qt(0.975, 13)
[1] 2.160369
> # 유의확률 p-값
> 2*(1-pt(13.889, 13))
[1] 3.553531e-09
```

⇒ 이 결과에서 기울기 β_1 의 추정값 $b_1 = 2.1497$ 이고, t-값

$$t_0 = \frac{2.1497}{0.1548} = 13.889$$

단순회귀모형 : 신뢰대 그리기

```
> pred.frame = data.frame(X=seq(3.5, 14.5, 0.2))  
> pc = predict(market.lm, int="c", newdata=pred.frame)    #기댓값 신뢰구간  
> pp = predict(market.lm, int="p", newdata=pred.frame)    #새로운 값 신뢰구간  
> head(pc, 3)
```

	fit	lwr	upr
1	7.852079	5.855247	9.848911
2	8.282014	6.344903	10.219125
3	8.711949	6.834076	10.589821

```
> head(pp, 3)
```

	fit	lwr	upr
1	7.852079	3.885278	11.81888
2	8.282014	4.344937	12.21909
3	8.711949	4.803678	12.62022

단순회귀모형 : 신뢰대 그리기

```
> pred.X = pred.frame$X
```

```
> pred.X
```

```
[1] 3.5 3.7 3.9 4.1 4.3 4.5 4.7 4.9 5.1 5.3 5.5 5.7 5.9 6.1
```

```
[15] 6.3 6.5 6.7 6.9 7.1 7.3 7.5 7.7 7.9 8.1 8.3 8.5 8.7 8.9
```

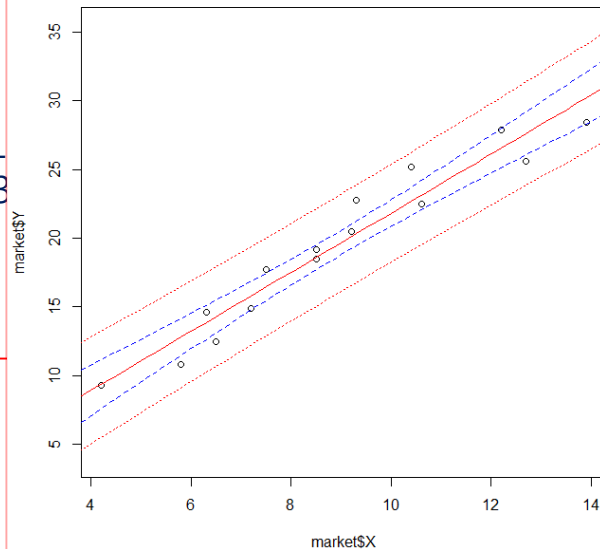
```
[29] 9.1 9.3 9.5 9.7 9.9 10.1 10.3 10.5 10.7 10.9 11.1 11.3 11.5 11.7
```

```
[43] 11.9 12.1 12.3 12.5 12.7 12.9 13.1 13.3 13.5 13.7 13.9 14.1 14.3
```

```
> plot(market$X, market$Y, ylim=range(market$Y, pp))
```

```
> matlines(pred.X, pc, lty=c(1,2,2), col="BLUE")
```

```
> matlines(pred.X, pp, lty=c(1,3,3), col="RED")
```



중회귀모형 : 모형적합

```
> market2 = read.table("c:/data/reg/market-2.txt", header=T)
```

```
> head(market2, 2)
```

	ID	X1	X2	Y
1	1	4.2	4.5	9.3
2	2	8.5	12.0	18.5

```
> market2.lm = lm(Y ~ X1+X2, data=market2)
```

```
> summary(market2.lm)
```

...

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.85041	0.84624	1.005	0.334770
X1	1.55811	0.14793	10.532	2.04e-07 ***
X2	0.42736	0.08431	5.069	0.000276 ***

Residual standard error: 0.9318 on 12 degrees of freedom
Multiple R-squared: 0.9799, Adjusted R-squared: 0.9765
F-statistic: 292.5 on 2 and 12 DF, p-value: 6.597e-11

적합된 회귀식 : $\hat{Y} = 0.85041 + 1.55811X_1 + 0.4273X_2$

결정계수 : 0.9799

F-값=292.5 이고, 유의확률 p-값= 6.597×10^{-11} 로서
적합된 중회귀모형이 이 데이터를 설명하는데 유의함.

(이는 귀무가설 $H_0: \beta_1 = \beta_2 = 0$ 이 기각되므로

β_1 과 β_2 가 동시에 영이 되지는 않을 것이라는 의미임)

중회귀모형 : 분산분석표

```
> anova(market2.lm)
```

Analysis of Variance Table

Response: Y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X1	1	485.57	485.57	559.283	1.955e-11	***
X2	1	22.30	22.30	25.691	0.0002758	***
Residuals	12	10.42	0.87			

중회귀모형의 분산분석 결과 해석

$$SS(X1) = 485.57$$

$$SS(X2|X1) = 22.30$$

, 여기서 $SS(X2|X1)$ 는 변수 $X1$ 이 적합된 후,
변수 $X2$ 가 추가되었을 때의 추가제곱합을 의미.
회귀제곱합 $SS(X1, X2) = SS(X1) + SS(X2|X1)$

〈분산분석표〉

요인	자유도	제곱합	평균제곱	F_0	Pr(>F)
회귀	2	507.87	253.94	292	6.597e-11
잔차	12	10.42	0.87		
계	14	518.29			

표준화 회귀모형

```
> install.packages("lm.beta")
> library(lm.beta)
> market2.lm = lm(Y ~ X1+X2, data=market2)
> market2.beta = lm.beta(market2.lm)
> print(market2.beta)
```

```
---
Standardized Coefficients::
(Intercept)          X1          X2
  0.0000000  0.7015566  0.3376137
```

```
> summary(market2.beta)
```

```
---
Coefficients:
              Estimate Standardized Std. Error t value Pr(>|t|)
(Intercept)  0.85041      0.00000    0.84624   1.005 0.334770
X1           1.55811      0.70156    0.14793  10.532 2.04e-07 ***
X2           0.42736      0.33761    0.08431   5.069 0.000276 ***
```

```
---
Residual standard error: 0.9318 on 12 degrees of freedom
Multiple R-squared:  0.9799,    Adjusted R-squared:  0.9765
F-statistic: 292.5 on 2 and 12 DF,  p-value: 6.597e-11
```

적합된 표준화 회귀모형

$$\hat{Y}^* = 0.7016Z_1 + 0.3376Z_2$$

※ 여기서 X1의 표준화계수가 X2의 표준화계수보다 크므로 상대적으로 X1의 영향이 더 큼을 알 수 있음.

신뢰구간

마켓데이터에 대하여 $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ 을 적합시켰을 때

(1) $x_1 = 10, x_2 = 10$ 에서 $E(y)$ 를 95% 신뢰구간으로 추정하고,

(2) $H_0 : \beta_1 = 0, H_0 : \beta_2 = 0$ 에 대하여 유의수준 $\alpha = 0.05$ 로

가설검정 하여보자.

```
> # 1 : 95% ( 99% 신뢰구간 추가 )
> pred.x = data.frame(X1=10, X2=10)
> pc = predict(market2.lm, int="c", newdata=pred.x)
> pc
      fit      lwr      upr
1 20.70503 19.95796 21.45209
> pc99 = predict(market2.lm, int="c", level=0.99, newdata=pred.x)
> pc99
      fit      lwr      upr
1 20.70503 19.65769 21.75236
```

다항회귀 적합

```
> maraton.lm = lm(m1990 ~ sect+I(sect^2)+I(sect^3), data=maraton)
> summary(maraton.lm)
```

...

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	917.592857	8.083355	113.516	3.61e-08	***
sect	13.785281	1.462847	9.424	0.000707	***
I(sect^2)	-0.683225	0.073387	-9.310	0.000741	***
I(sect^3)	0.012248	0.001077	11.375	0.000341	***

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

Residual standard error: 3.281 on 4 degrees of freedom

Multiple R-squared: 0.9983, Adjusted R-squared: 0.9969

F-statistic: 761.4 on 3 and 4 DF, p-value: 5.726e-06

적합된 3차 다항회귀모형식 :

$$\widehat{m1990} = 917.593 + 13.785 \times \text{sect} - 0.683 \times \text{sect}^2 + 0.012 \times \text{sect}^3$$

가변수를 이용한 회귀모형

✓ 독립변수에 이산형 변수가 포함되어 있는 경우에 사용

(자료 예)

<비누생산공정에서 비누부스러기 부산물의 양과 공정속도>

1번 생산공정 ($D=1$)			2번 생산공정 ($D=0$)		
부산물의 양		공정속도	부산물의 양		공정속도
Y	X	D	Y	X	D
218	100	1	140	105	0
248	125	1	277	215	0
360	220	1	384	270	0
351	205	1	341	255	0
470	300	1	215	175	0
394	255	1	180	135	0
332	225	1	260	200	0
321	175	1	361	275	0
410	270	1	252	155	0
260	170	1	422	320	0
241	155	1	273	190	0
331	190	1	410	295	0
275	140	1			
425	290	1			
367	265	1			

교호작용을 고려한 모형 :

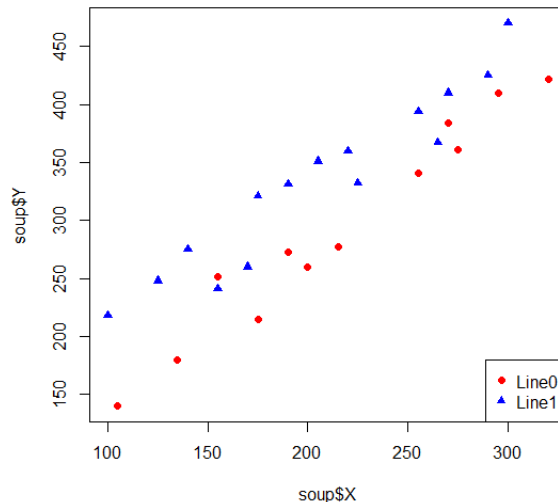
$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 XD + \epsilon$$

교호작용이 없는 모형 :

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \epsilon$$

두 그룹별 산점도 그리기

```
> soup = read.table("c:/data/reg/soup.txt", header=T)
> soup[c(1,15,16,27),]
      Y   X D
1  218 100 1
15 367 265 1
16  140 105 0
27 410 295 0
> soup$D = factor(soup$D, levels=c(0,1), label=c("Line0", "Line1"))
> plot(soup$X, soup$Y, type="n")
> points(soup$X[soup$D=="Line1"], soup$Y[soup$D=="Line1"],
        pch=17, col="BLUE")
> points(soup$X[soup$D=="Line0"], soup$Y[soup$D=="Line0"],
        pch=19, col="RED")
> legend("bottomright", legend=levels(soup$D),
        pch=c(19,17), col=c("RED", "BLUE"))
```



가변수를 이용한 회귀모형 : 교호작용을 고려한 경우

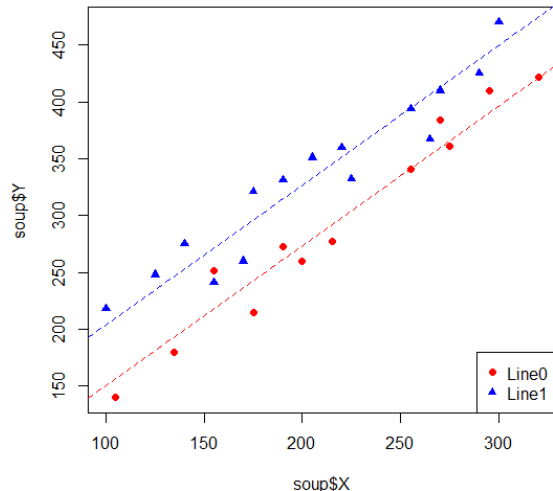
```
> soup2.lm = lm(Y ~ X+D+X:D, data=soup)
> summary(soup2.lm)
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.57446    20.86970   0.363  0.71996
X             1.32205     0.09262  14.273 6.45e-13 ***
DLine1      90.39086    28.34573   3.189  0.00409 **
X:DLine1     -0.17666     0.12884  -1.371  0.18355
---
Residual standard error: 20.75 on 23 degrees of freedom
Multiple R-squared:  0.9447,    Adjusted R-squared:  0.9375
F-statistic: 130.9 on 3 and 23 DF,  p-value: 1.341e-14
```

- ⇒ 교호작용항의 경우, 회귀계수의 추정값은 -0.1767 이고, t_0 값에
대한 유의확률은 0.18355 로서 유의수준 0.05 보다 크므로 유의하지
않음을 알 수 있음.
- ⇒ 따라서 이 자료의 경우에는 교호작용을 고려하지 않은 모형으로 적합
하는 것이 좋음.

가변수를 이용한 회귀모형 : 교호작용을 고려하지 않는 경우

```
> soup.lm = lm(Y ~ X+D, data=soup)
> summary(soup.lm)
...
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.28179    15.40701   1.771   0.0893 .
X             1.23074     0.06555  18.775 7.48e-16 ***
DLine1       53.12920     8.21003   6.471 1.08e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
                  ' ' 1

Residual standard error: 21.13 on 24 degrees of freedom
Multiple R-squared:  0.9402,    Adjusted R-squared:  0.9352
F-statistic: 188.6 on 2 and 24 DF,  p-value: 2.104e-15
> abline(27.28179, 1.23074, lty=2, col="RED")
> abline(27.28179+53.1292, 1.23074, lty=2, col="BLUE")
```



적합된 회귀모형 : $\hat{Y} = 27.282 + 1.231X + 54.129D$

⇒ 따라서 기울기가 동일하다고 가정하는 경우의 회귀모형 적합에서
두 생산라인의 차이는 54.129 가 됨을 알 수 있음.

특이값 검정

```
> forbes.res = ls.diag(forbes.lm)
> names(forbes.res)
[1] "std.dev"      "hat"          "std.res"      "stud.res"
[5] "cooks"        "dfits"        "correlation"  "std.err"
[9] "cov.scaled"   "cov.unscaled"
> resid.result = cbind(forbes.res$std.res, forbes.res$stud.res, forbes.res$hat)
> colnames(resid.result) = c("standardized resid", "studentized resid", "Hat")
> resid.result = round(resid.result,3)
> print(resid.result)
```

	standardized resid	studentized resid	Hat
[1,]	-0.728	-0.716	0.193
[2,]	-0.203	-0.197	0.200
[3,]	-0.150	-0.145	0.107
[4,]	0.052	0.050	0.098
[5,]	0.091	0.088	0.083
[10,]	-0.230	-0.223	0.064
[11,]	-0.400	-0.388	0.060
[12,]	3.707	12.374	0.064
[13,]	0.004	0.004	0.140
[16,]	-0.235	-0.227	0.210
[17,]	-0.260	-0.252	0.220

```
> rstudent(forbes.lm) #스튜던트화 잔차
```

1	2	3	4	5
-0.716454916	-0.196531386	-0.145084092	0.050361279	0.088080643

```
> Bonferroni 유의수준 0.01에서 기각치
```

```
> qt(0.01/(2*17), 14)
```

```
[1] -4.414447
```

```
> Bonferroni p-value for obs.12
```

```
> 2*17*(1-pt(12.374,14))
```

```
[1] 1.071262e-07
```

특이값 검정

```
> library(car)
> outlierTest(forbes.lm)
      rstudent unadjusted p-value Bonferonni p
12 12.37386      6.3025e-09    1.0714e-07
```

잔차분석 및 cook 통계량

```
> soil.diag = ls.diag(soil.lm)
> names(soil.diag)
[1] "std.dev"      "hat"          "std.res"      "stud.res"
[5] "cooks"        "dfits"        "correlation"  "std.err"
[9] "cov.scaled"   "cov.unscaled"
> diag.st = cbind(soil.diag$hat, soil.diag$std.res, soil.diag$stud.res, soil.diag$cooks)
> colnames(diag.st) = c("Hii", "ri", "ti", "Di")
> round(diag.st, 3)
```

	Hii	ri	ti	Di
[1,]	0.464	0.736	0.709	0.117
[2,]	0.248	0.599	0.569	0.029
[3,]	0.363	-0.014	-0.013	0.000
[4,]	0.299	-0.008	-0.008	0.000
[5,]	0.332	-0.131	-0.121	0.002
[6,]	0.118	-0.091	-0.084	0.000
[7,]	0.533	2.075	3.098	1.227
[8,]	0.530	-0.382	-0.358	0.041
[9,]	0.629	0.636	0.607	0.171
[10,]	0.188	-2.232	-3.851	0.289
[11,]	0.298	-0.454	-0.427	0.022

```
> Di = cooks.distance(soil.lm)
> round(Di, 3)
```

	1	2	3	4	5	6	7	8	9	10	11
	0.117	0.029	0.000	0.000	0.002	0.000	1.227	0.041	0.171	0.289	0.022

```
> library(car)
> outlierTest(soil.lm)
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
      rstudent unadjusted p-value Bonferonni p
10 -3.850967          0.0084505      0.092955
```

7번 : 영향력있는 관측값

오차의 등분산 - 스코어 검정

```
> library(car)
```

```
> ncvTest(goose.lm)
```

Non-constant Variance Score Test

Variance formula: ~ fitted.values

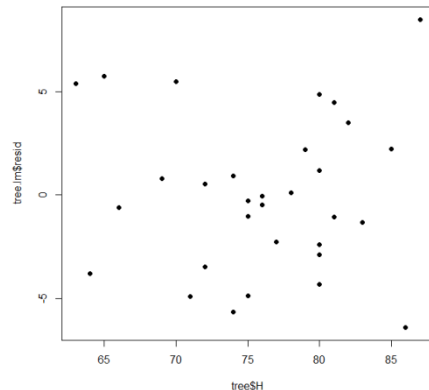
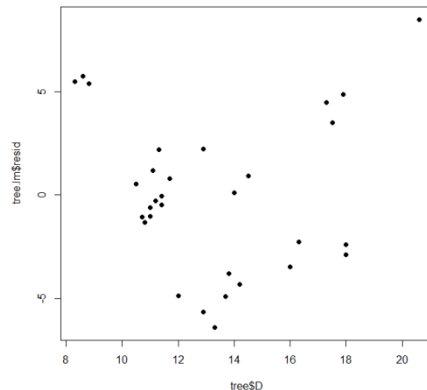
Chisquare = 81.41318 Df = 1 p = 1.831324e-19

스코어 검정의 $\chi^2=81.41$ 이고, 유의확률 p-값이 매우 작으므로
등분산 가정을 기각

선형성 진단

〈나무자료〉

순서	지름	높이	부피	순서	지름	높이	부피	순서	지름	높이	부피
1	8.3	70	10.3	12	11.4	76	21.0	23	14.5	74	36.3
2	8.6	65	10.3	13	11.4	76	21.4	24	16.0	72	38.3
3	8.8	63	10.2	14	11.7	69	21.3	25	16.3	77	42.6
4	10.5	72	16.4	15	12.0	75	19.1	26	17.3	81	55.4
5	10.7	81	18.8	16	12.9	74	22.2	27	17.5	82	55.7
6	10.8	83	19.7	17	12.9	85	33.8	28	17.9	80	58.3
7	11.0	66	15.6	18	13.3	86	27.4	29	18.0	80	51.5
8	11.0	75	18.2	19	13.7	71	25.7	30	18.0	80	51.0
9	11.1	80	22.6	20	13.8	64	24.9	31	20.6	87	77.0
10	11.2	75	19.9	21	14.0	78	34.5				
11	11.3	79	24.2	22	14.2	80	31.7				



```
> tree = read.table("c:/data/reg/tree.txt", header=T)
> head(tree, 3)
  num  D  H  V
1   1 8.3 70 10.3
2   2 8.6 65 10.3
3   3 8.8 63 10.2
> tree.lm = lm(V ~ D+H, data=tree)
> plot(tree$D, tree.lm$resid, pch=19)
> plot(tree$H, tree.lm$resid, pch=19)
```

변수 D의 잔차산점도의 경우
2차 함수 형태의 비선형성이 나타남.

오차의 정규성

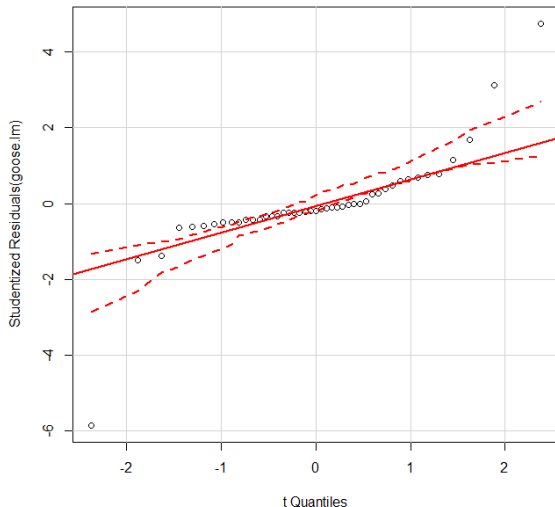
```
> goose.lm = lm(photo ~ obsA, data=goose)
> qqPlot(goose.lm)
> # 정규성 검정
> library(mvnormtest)
> goose.rstudent = rstudent(goose.lm)
> shapiro.test(goose.rstudent)
```

Shapiro-Wilk normality test

data: goose.rstudent

W = 0.7192, p-value = 5.971e-08

W 통계량의 값은 0.7192 이고, 유의확률 p-값이 매우 작으므로 정규성 가정을 기각함.



잔차가 직선의 형태를 벗어나 곡선의 형태로 직선에서 벗어나고 있음을 보이므로 정규성 가정에 위배되는 것으로 판단.

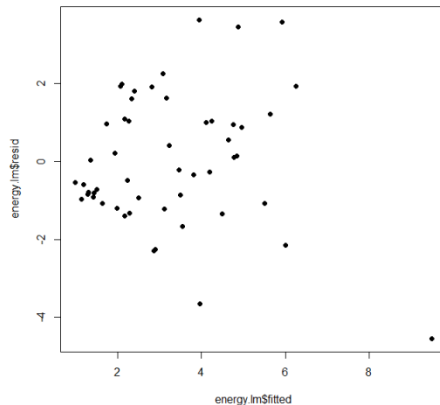
Box-Cox 변환

<53명의 주거지역 고객에 대한 수요(Y)와 에너지사용량(X)>

고객	X(KWH)	Y(KW)	고객	X(KWH)	Y(KW)	고객	X(KWH)	Y(KW)
1	679	0.79	19	745	0.77	37	770	1.74
2	292	0.44	20	435	1.39	38	724	4.10
3	1012	0.56	21	540	0.56	39	808	3.94
4	493	0.79	22	874	1.56	40	790	0.96
5	582	2.70	23	1543	5.28	41	783	3.29
6	1156	3.64	24	1029	0.64	42	406	0.44
7	997	4.73	25	710	4.00	43	1242	3.24
8	2189	9.50	26	1434	0.31	44	658	2.14
9	1097	5.34	27	837	4.20	45	1746	5.71
10	2078	6.85	28	1748	4.88	46	468	0.64
11	1818	5.84	29	1381	3.48	47	1114	1.90
12	1700	5.21	30	1428	7.58	48	413	0.51
13	747	3.25	31	1255	2.63	49	1787	8.33
14	2030	4.43	32	1777	4.99	50	3560	4.94
15	1643	3.16	33	370	0.59	51	1495	5.11
16	414	0.50	34	2316	8.19	52	2221	3.85
17	354	0.17	35	1130	4.79	53	1526	3.93
18	1276	1.88	36	463	0.51			

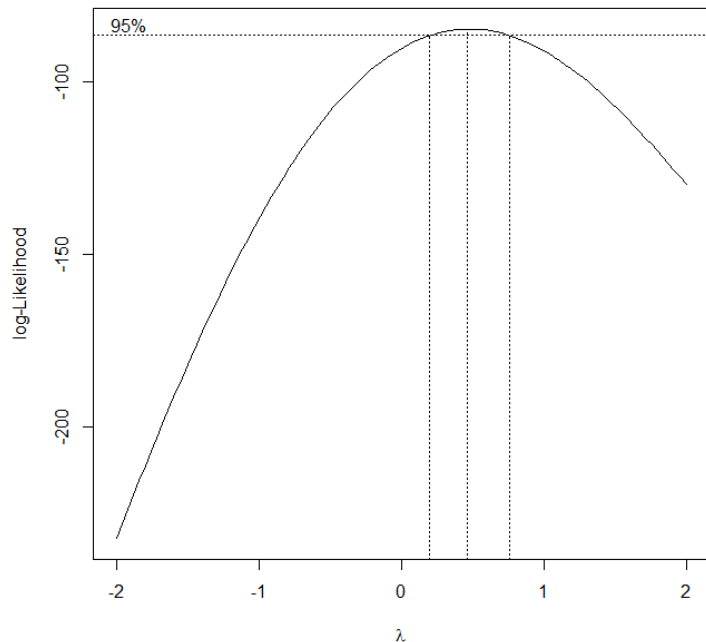
```
> energy = read.table("c:/data/reg/energy.txt", header=T)
> head(energy,3)
  customer      X      Y
1         1  679 0.79
2         2  292 0.44
3         3 1012 0.56
> energy.lm = lm(Y ~ X, data=energy)
> plot(energy.lm$fitted, energy.lm$resid, pch=19)
```

잔차산점도는 X가 증가함에 따라 잔차의
흩어짐이 많아짐 => 이분산성이 의심됨



Box-Cox 변환

```
> library(MASS)
> boxcox(Y~X, data=energy, lambda=seq(-2,2, 1/2), plotit=TRUE)
```



Box-Cox 변환그림에서는 log-likelihood 값이 최대가 되는 λ 값을 찾으면 됨. 그림에서 λ 는 0.5 가 됨. 이는 $\sqrt{\cdot}$ 변환에 해당.

로지스틱 회귀모형 : 이항자료

〈날다람쥐 Sugar Glider의 출현자료〉

p_no	occurr	con_metric	p_size_km
1	1	0.650	130.9
2	0	0.610	104.1
3	0	0.744	132.3
4	1	0.213	225.6
5	1	0.723	83.0
6	0	0.678	48.8
7	0	0.733	61.0
8	1	0.522	39.6
9	1	0.552	193.1
10	0	0.245	155.6

```
> glider <-  
  read.csv('c:/data/reg/sugar_glider_binomial.csv')  
> head(glider, 3)  
  p_no occur con_metric p_size_km  
1    1     1     0.650    130.9  
2    2     0     0.610    104.1  
3    3     0     0.744    132.3  
> logit_m2 <- glm(occurr~p_size_km,  
                  family=binomial(link=logit),  
                  data=glider)
```

반응변수 $y = \text{occur}$, 1=yes, 0=no 이므로 이항분포를 가정

로지스틱 회귀모형 : $\eta = \log it(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1$

$$\pi = E(Y | x) = \Pr(y = 1 | x)$$

로지스틱 회귀모형 : 이항자료

```
> summary(logit_m2)
```

```
...
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5541	-0.8980	-0.5157	0.8075	2.0394

Coefficients:


	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.528298	0.820251	-3.082	0.00205 **
p_size_km	0.021727	0.006893	3.152	0.00162 **

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 68.994 on 49 degrees of freedom
Residual deviance: 55.716 on 48 degrees of freedom
AIC: 59.716

Number of Fisher Scoring iterations: 3


$$\log\left(\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right) = -2.528 + 0.022 \times x$$

Deviance(이탈도)

: 선형회귀모형의 잔차제곱합을 일반화한 개념.
정규분포를 따르는 가정이 맞는 경우,
 χ^2 -분포를 따름.
“Residual deviance/df < 2”이면,
모형의 적합도에 큰 문제가 없다고 판단함.

정리된 자료의 로지스틱 회귀모형 적합

<구획 크기의 계급구간에서 구획 수, Sugar Glider 출현 구획 수, 표본비율>

p_size_km	구간의 중앙값	출현 구획 수	구획 수	표본비율
≤ 50.0	35.3	3	10	0.30
50.0 ~ 100.0	79.55	3	14	0.21
100.0 ~ 150.0	123.6	6	14	0.43
150.0 ~ 200.0	177.65	9	10	0.90
200.0 <	214.55	2	2	1.00

```
> glider_g <- read.csv('c:/data/reg/sugar_glider_binomial_g.csv')
> head(glider_g)
  p_size_med count cases
1      35.30    10     3
2      79.55    14     3
3     123.60    14     6
4     177.65    10     9
5     214.55     2     2
> y <- cbind(glider_g$cases, glider_g$count-glider_g$cases)
> logit_mg <- glm(y~glider_g$p_size_med, family=binomial(link=logit))
```

프로빗 모형

```
> probit_m <- glm(occurr~p_size_km, family=binomial(link=probit))
```

```
> summary(probit_m)
```

...

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5586	-0.9211	-0.5186	0.8041	2.0341

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.492825	0.460250	-3.244	0.001181	**
p_size_km	0.013023	0.003866	3.368	0.000757	***

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 68.994 on 49 degrees of freedom

Residual deviance: 55.797 on 48 degrees of freedom

AIC: 59.797

Number of Fisher Scoring iterations: 5

$$\Phi^{-1}(\hat{\pi}(x)) = \hat{\beta}_0 + \hat{\beta}_1 x = -1.493 + 0.013x$$

$$\hat{\pi}(x) = \Phi(\hat{\beta}_0 + \hat{\beta}_1 x) = \Phi(-1.493 + 0.013x)$$

모형이 적합

로그선형모형 : 개수형자료분석

< 고속도로 속도제한여부와 교통사고 건수 >

year	day	limit	y	year	day	limit	y
1961	1	no	9	1962	1	no	9
1961	2	no	11	1962	2	no	20
1961	3	no	9	1962	3	no	15
1961	4	no	20	1962	4	no	14
1961	5	no	31	1962	5	no	30
1961	6	no	26	1962	6	no	23
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

> library(MASS)

> data(Traffic)

> head(Traffic, 3)

	year	day	limit	y
1	1961	1	no	9
2	1961	2	no	11
3	1961	3	no	9

주요관심 내용 : 고속도로의 속도제한이 평균 사고건수에 어떤 영향을 주는가

분석모형 : $\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{92} x_{92} + \beta_{93} x_{93}$

$$x_1 = \begin{cases} 0 & \text{no} \\ 1 & \text{yes} \end{cases}, x_i = \begin{cases} 1 & \text{day} = i \\ 0 & \text{나머지} \end{cases}, i = 2, 3, \dots, 92, x_{93} = \begin{cases} 0 & \text{year} = 1961 \\ 1 & \text{year} = 1962 \end{cases}$$

로그선형모형 : 개수형자료분석

```
> Traffic$day <- as.factor(Traffic$day)
> Traffic$year <- as.factor(Traffic$year)
> log_m <- glm(y~limit+day+year, family=poisson(link=log), data=Traffic)
> summary(log_m)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	2.20984	0.23632	9.351	< 2e-16	***
limityes	-0.28424	0.04300	-6.610	3.86e-11	***
day2	0.54362	0.29633	1.834	0.066584	.
day3	0.28768	0.31180	0.923	0.356197	
day91	0.37539	0.31528	1.191	0.233800	
day92	0.64109	0.29876	2.146	0.031888	*
year1962	-0.02539	0.03458	-0.734	0.462927	

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 625.25 on 183 degrees of freedom

Residual deviance: 107.11 on 90 degrees of freedom

AIC: 1185.1

Number of Fisher Scoring iterations: 4

유의하지 않음

107.11/90=1.19 로 모형 적합

로그선형모형 : 개수형자료분석

```
> log_m1 <- glm(y~limit+day, family=poisson(link=log), data=Traffic)
> summary(log_m1)
```

Coefficients: $\log(\hat{\mu}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_{92} x_{92} = 2.20 - 0.30x_1 + 0.54x_2 + \dots + 0.65x_{92}$

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	2.19722	0.23570	9.322	< 2e-16 ***
limityes	-0.29627	0.03978	-7.448	9.46e-14 ***
day2	0.54362	0.29633	1.834	0.066584 .
...				
day91	0.38232	0.31515	1.213	0.225077
day92	0.64803	0.29862	2.170	0.030004 *

최종모형

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 625.25 on 183 degrees of freedom
Residual deviance: 107.64 on 91 degrees of freedom
AIC: 1183.6
Number of Fisher Scoring iterations: 4

107.64/91=1.18 로 더 작아짐

AIC 값도 더 작아짐

로그선형모형의 해석 : 개수형자료분석

속도제한여부(limit)의 회귀계수 추정치 $\hat{\beta}_1$ 에 대한 해석

특정일 ($x_j, j=2, \dots, 92$)에

속도제한을 하지 않은 경우($x_1=0$)

평균사고건수($=\mu_0$)의 로그추정치 :

$$\log(\hat{\mu}_0) = \hat{\beta}_0 + 0 + \hat{\beta}_j x_j = 2.20 + 0 + \hat{\beta}_j x_j$$

속도제한을 한 경우($x_1=1$) 평균사고건수($=\mu_1$)의 로그추정치 :

$$\log(\hat{\mu}_1) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_j x_j = 2.20 - 0.30 + \hat{\beta}_j x_j$$

$$\Rightarrow \log\left(\frac{\hat{\mu}_1}{\hat{\mu}_0}\right) = \hat{\beta}_1 = -0.30$$

$$\Rightarrow \frac{\hat{\mu}_1}{\hat{\mu}_0} = e^{-0.30} = 0.74$$

⇒ 즉, 고속도로에서 속도제한을 했을 때의 평균사고건수는
속도제한을 하지 않았을 때의 평균사고건수의 74% 수준으로
감소

$\frac{\mu_1}{\mu_0} = e^{\beta_1}$ 의 추정치와 95% 신뢰구간

```
> exp(coef(log_m1, parm="limit"))  
(Intercept)  limityes    day2  
9.0000000    0.7435897    1.7222222  
...  
> exp(confint(log_m1, parm="limityes",  
  level=0.95))  
Waiting for profiling to be done...  
      2.5 %    97.5 %  
0.6877111 0.8037687
```

고속도로에서 속도제한을 실시하면 평균사고
건수가 26% 정도 감소하며 95% 신뢰수준에서
많게는 31% 적게는 20% 정도 감소하는 것으로
추정됨

한 학기 동안 수고했습니다.