

6강

통계추론 2

한림대학교 금융정보통계학과 심송용교수

목 차

1. 분산에 대한 추론
2. 상관계수에 대한 추론
3. 도수분포표와 교차표
4. 회귀분석
5. 분산분석



1

분산에 대한 추론



분산에 대한 추론

- X_1, X_2, \dots, X_n 이 정규분포 $N(\mu, \sigma^2)$ 에서의 확률표본이라 하고 표본평균 \bar{X} 와 표본분산 S^2 을 각각

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}, \quad S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

이라고 하면

$$\chi^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2}$$

은 자유도 $(n - 1)$ 인 카이제곱분포.

- 따라서 σ^2 에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1; \alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1; 1-\alpha/2}} \right)$$

분산에 대한 추론

- 귀무가설 $H_0 : \sigma^2 = \sigma_0^2$ 에 대해서 검정통계량은

$$\chi_0^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma_0^2} = \frac{(n-1)s^2}{\sigma_0^2}$$

을 사용하여 각각의 대립가설에 대한 기각역 및 유의확률은

대립가설	기각역	유의확률
$H_1: \sigma^2 > \sigma_0^2$	$\chi_0^2 > \chi_{n-1;\alpha}^2$	$\Pr[\chi_{n-1}^2 > \chi_0^2]$
$H_1: \sigma^2 < \sigma_0^2$	$\chi_0^2 < \chi_{n-1;1-\alpha}^2$	$\Pr[\chi_{n-1}^2 < \chi_0^2]$
$H_1: \sigma^2 \neq \sigma_0^2$	$\chi_0^2 > \chi_{n-1;\alpha/2}^2$ 또는 $\chi_0^2 < \chi_{n-1;1-\alpha/2}^2$	$\chi_0^2 < 1$ 이면 $\Pr[\chi_{n-1}^2 < \chi_0^2]$ $\chi_0^2 > 1$ 이면 $\Pr[\chi_{n-1}^2 > \chi_0^2]$

분산에 대한 추론

예제 3.14. R에서 분산이 4인 정규분포에서 난수 100개를 발생하여 분산이 4보다 큰지 유의수준 5%에서 검정하고 95% 신뢰구간을 구해보자.

```
> nn <- 100
> x <- rnorm(nn, sd =2); vx <- var(x)
> chi0 <- (nn-1)*vx / 2^2
> chi0
[1] 103.5662
> p.val <- 1-pchisq(chi0, nn-1)
> p.val
[1] 0.3568574
> ci <- c( (nn-1)*vx /qchisq(0.975, nn-1), (nn-1)*vx /qchisq(0.025, nn-1) )
> ci
[1] 3.225810, 5.646932
```

분산에 대한 추론(이표본)

1. X_1, X_2, \dots, X_m 이 정규분포 $N(\mu_1, \sigma_1^2)$ 에서의 확률표본이고
2. Y_1, Y_2, \dots, Y_n 이 정규분포 $N(\mu_2, \sigma_2^2)$ 에서의 확률표본이며
3. 두 확률표본이 독립인 경우

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$

의 분포는 자유도가 $(m - 1, n - 1)$ 인 F-분포

- 따라서 $\frac{\sigma_1^2}{\sigma_2^2}$ 에 대한 $100(1-\alpha)\%$ 신뢰구간은

$$\left(\frac{1}{F_{m-1, n-1; \alpha/2}} \frac{s_1^2}{s_2^2}, \frac{1}{F_{m-1, n-1; 1-\alpha/2}} \frac{s_1^2}{s_2^2} \right)$$

분산에 대한 추론(이표본)

- 귀무가설 $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = r$ 에 대해서 검정통계량은

$$F_0 = \frac{S_1^2}{rS_2^2}$$

을 사용하여 각각의 대립가설에 대한 기각역 및 유의확률은

대립가설	기각역	유의확률
$H_1: \sigma_1^2 / \sigma_2^2 > r$	$F_0 > F_{m-1, n-1; \alpha}$	$\Pr[F_{m-1, n-1} > F_0]$
$H_1: \sigma_1^2 / \sigma_2^2 < r$	$F_0 < F_{m-1, n-1; 1-\alpha}$	$\Pr[F_{m-1, n-1} < F_0]$
$H_1: \sigma_1^2 / \sigma_2^2 \neq r$	$F_0 > F_{m-1, n-1; \alpha/2}$ 또는 $F_0 < F_{m-1, n-1; 1-\alpha/2}$	$F_0 < 1$ 이면 $\Pr[F_{m-1, n-1} < F_0]$ $F_0 > 1$ 이면 $\Pr[F_{m-1, n-1} > F_0]$

분산에 대한 추론(이표본)

```
var.test(x, y, ratio = 1, alternative = c("two.sided", "less", "greater"),  
         conf.level = 0.95, ...)
```

또는

```
var.test(formula, data, subset, na.action, ...)
```

- x, y : 분산비에 대한 추론에 적용될 두 변수의 값을 저장한 벡터의 이름.
- alternative : 대립가설의 방향을 설정함. 기본값은 양측검정을 설정하는 "two.sided"이며 '크다' 또는 '작다'인 경우 각각 "greater" 또는 "less"를 설정할 수 있으며 첫 글자만 사용 가능. 여기서 '크다'와 '작다' 먼저 나온 x의 **분산**이 두 번째인 y의 **분산**보다 '크다' 또는 '작다'를 의미함.
- ratio : 귀무가설하에서의 값(r)을 설정한다. 기본값은 1.

분산에 대한 추론(이표본)

- conf.level : 신뢰구간의 신뢰도를 설정.
- formula : 독립인 두 그룹을 설정하는 방법으로 dep ~ indep 로 설정하며 dep는 종속변수, indep는 그룹(이 경우 두 개의 그룹)을 표시하는 변수를 설정.
- data : 사용할 데이터 프레임의 이름을 설정.
- subset : data에 설정한 데이터 프레임에서 일부의 자료만 얻고자 할 때 해당하는 조건식을 설정.
- na.action : 자료에 NA가 있을 때 NA를 어떻게 처리할지 설정.
설정은 NA의 값을 처리할 함수를 설정하며,
기본값은 `getOption("na.action")`에 설정된 값.
R 기본 설치시 설정된 값은 NA를 제외하는 것임.

분산에 대한 추론(이표본)

예제 3.16. 예제 3.2의 자료를 var.test 함수에 적용하면 다음과 같은 결과를 얻을 수 있다. 이 결과는 모두 예제 3.15의 결과와 일치함을 확인할 수 있다.

```
> x <- c(21.6,20.8,17.6,20.1,20.1,21.9,20.6,19.4,21.5,26.1)
> y <- c(20.6,20.4,20.2,20.2,18.0,19.8,20.9,19.7,20.3,19.7,22.7)
> var.test(x,y)
```

F test to compare two variances

data: x and y

F = 3.8723, num df = 9, denom df = 10, p-value = 0.04617

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

1.024708 15.349414

sample estimates:

ratio of variances

3.872335

2

상관계수에 대한 추론



상관계수에 대한 추론

- $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$: 독립인 이변량 정규분포에서의 확률벡터.
- 공분산은

$$Cov(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

- S_x^2 과 S_y^2 을 각각 X 와 Y 의 분산이라고 하면 상관계수는

$$r = Corr(X, Y) = \frac{Cov(X, Y)}{S_x S_y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

상관계수에 대한 추론

- Fisher 변환은 모상관계수가 ρ 이고 표본상관계수가 r 일 때 다음의 변환이 근사적으로 정규분포를 따른다는 것을 말함.

$$v = \frac{1}{2} \log \frac{1+r}{1-r} \sim N\left(\frac{1}{2} \log \frac{1+\rho}{1-\rho}, \frac{1}{n-3}\right) \quad (\text{교재의 기댓값 } \frac{1}{2} \log \frac{1+r}{1-r} \text{ 는 오타임})$$

- 따라서 귀무가설 $H_0 : \rho = \rho_0$ 에 대한 가설검정의 검정통계량은

$$Z = \frac{\frac{1}{2} \log \frac{1+r}{1-r} - \frac{1}{2} \log \frac{1+\rho_0}{1-\rho_0}}{\sqrt{\frac{1}{n-3}}}$$

를 사용하며 (근사) 기각역 및 유의확률은 다음과 같음.

상관계수에 대한 추론

대립가설	기각역	유의확률
$H_1: \rho > \rho_0$	$z_0 > z_\alpha$	$\Pr[Z > z_0]$
$H_1: \rho < \rho_0$	$z_0 < -z_\alpha$	$\Pr[Z < z_0]$
$H_1: \rho \neq \rho_0$	$ z_0 > z_{\alpha/2}$	$2\Pr[Z > z_0]$

- 상관계수에 대한 신뢰구간은
 1. Fisher 변환의 근사분포를 이용하여 $v = \frac{1}{2} \log \frac{1+\rho}{1-\rho}$ 에 대한 신뢰구간을 구함.
 2. Fisher 변환의 역변환 $\rho = \frac{e^{2v}-1}{e^{2v}+1}$ 으로 구함.
- 즉

상관계수에 대한 추론

- 표본상관계수 r 을 구해 $v = \frac{1}{2} \log \frac{1+\rho}{1-\rho}$ 에 대한 신뢰상한과 하한을 구함. 즉,

$$\text{신뢰하한: } v_L = \frac{1}{2} \log \frac{1+r}{1-r} - z_{\alpha/2} \sqrt{\frac{1}{n-3}}$$

$$\text{신뢰상한: } v_U = \frac{1}{2} \log \frac{1+r}{1-r} + z_{\alpha/2} \sqrt{\frac{1}{n-3}}$$

- Fisher 변환의 역변환 $\rho = \frac{e^{2v} - 1}{e^{2v} + 1}$ 으로 ρ 의 신뢰구간 구함.

$$r_L = \frac{e^{2v_L} - 1}{e^{2v_L} + 1}, \quad r_U = \frac{e^{2v_U} - 1}{e^{2v_U} + 1}$$

- 구간 (r_L, r_U) 가 $100(1-\alpha)\%$ 신뢰구간임.

상관계수에 대한 추론

- 회귀분석에서 유도된 추론 : 귀무가설 $H_0: \rho = 0$ 을 검정하기 위한 검정통계량으로

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (\text{주의 : 교재 오타})$$

을 검정통계량으로 사용하며 기각역과 유의확률은 다음과 같음.

대립가설	기각역	유의확률
$H_1: \rho > 0$	$t_0 > t_{n-2;\alpha}$	$\Pr[T_{n-2} > t_0]$
$H_1: \rho < 0$	$t_0 < -t_{n-2;\alpha}$	$\Pr[T_{n-2} < t_0]$
$H_1: \rho \neq 0$	$ t_0 > t_{n-2;\alpha/2}$	$2\Pr[T_{n-2} > t_0]$

상관계수에 대한 추론

예제 3.17. 아이의 재능에 대한 부모의 평가와 교사의 평가를 조사하였더니 다음과 같았다. 상관계수가 0인지 검정하고 상관계수에 대한 95% 신뢰구간을 얻어 보아라.

부모평가 : 35 35 33 34 31 35 35 35 35 33 35 35 35 35 31 32 35

교사평가 : 25 31 33 33 34 33 34 33 29 33 35 35 35 35 35 35 32

먼저 Fisher 변환을 사용하는 방법으로 검정통계량, 유의확률 및 신뢰구간을 구해보면 다음과 같다.

```
> x <- c(35,35,33,34,31,35,35,35,35,33,35,35,35,35,31,32,35)
> y <- c(25,31,33,33,34,33,34,33,29,33,35,35,35,35,35,35,32)
> rr <- cov(x, y)/(sd(x)*sd(y)) # 또는 rr <- cor(x, y)
> z0 <- 0.5*log((1+rr)/(1-rr)) / sqrt(1/14) # 17개의 자료
> 2*(1-pnorm(abs(z0))) # 양측 유의확률
[1] 0.1479776
```

상관계수에 대한 추론

```
> nul <- 0.5*log((1+rr)/(1-rr)) -qnorm(0.975)*sqrt(1/14)
> nuu <- 0.5*log((1+rr)/(1-rr)) +qnorm(0.975)*sqrt(1/14)
> lbd <- (exp(2*nul)-1)/(exp(2*nul)+1)
> ubd <- (exp(2*nuu)-1)/(exp(2*nuu)+1)
> c(lbd, ubd)
[1] -0.7213589 0.1363185
```

t-검정에 기초한 방법 : 유의확률은 Fisher 변환 때와 큰 차이가 나지 않으며 귀무가설을 기각할 수 없음.

```
> t0 <- rr*sqrt(15)/sqrt(1-rr^2)
> t0 # 검정통계량 인쇄
[1] -1.535081
> 2*(1-pt(abs(t0), 15)) # 유의확률
[1] 0.1455849
```

상관계수에 대한 추론

```
cor.test(x, y, alternative = c("two.sided", "less", "greater"),  
         method = c("pearson", "kendall", "spearman"), conf.level = 0.95, ...)
```

또는

```
cor.test(formula, data, subset, na.action, ...)
```

신뢰구간은 3.7.1절의 Fisher 변환을 적용한 신뢰구간을 구함.

가설검정은 3.7.2절의 t 검정을 사용. 따라서 귀무가설은 항상 $H_0: \rho = 0$ 임.

- x, y : 상관계수 추론을 하기 위해 얻은 두 변수의 값을 저장한 벡터. 두 벡터의 길이는 같아야 함.
- Alternative : 대립가설의 방향을 설정하며 양측검정인 경우 "two.sided", 대립가설이 0보다 크다면 "greater" 작다면 "less"를 설정.
양측검정이 기본값.

상관계수에 대한 추론

- method : 상관계수의 종류를 설정한다. 기본값은 피어슨 상관계수인 "pearson"이며 켄달의 계수는 "kendall"로, 스피어만의 순위상관계수는 "spearman"을 설정할 수 있음.
- conf.level : 신뢰구간의 신뢰도를 설정한다.
method가 "pearson"인 경우만 적용됨.
- formula : $\sim x + y$ 형태로 설정하여 x와 y의 상관계수에 대한 추론 결과를 얻음.
- data : formula에 포함된 변수가 있는 데이터 프레임의 이름을 설정.
- subset : data에서 설정한 데이터 프레임의 일부 자료만 얻고자 할 때 설정.
- na.action : NA 값을 어떻게 처리할지 설정.

상관계수에 대한 추론

예제 3.18. 예제 3.17의 자료를 사용하여 cor.test 함수를 적용해보자.

```
> x <- c(35,35,33,34,31,35,35,35,35,33,35,35,35,31,32,35)  
> y <- c(25,31,33,33,34,33,34,33,29,33,35,35,35,35,35,32)  
> cor.test(x,y)
```

Pearson's product-moment correlation

data: x and y

t = -1.5351, df = 15, p-value = 0.1456

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.7213589 0.1363185

sample estimates:

cor

-0.3684686

상관계수에 대한 추론

Data frame을 사용하는 경우

```
> x <- c(35,35,33,34,31,35,35,35,35,33,35,35,35,35,31,32,35)  
> y <- c(25,31,33,33,34,33,34,33,29,33,35,35,35,35,35,35,32)  
> cor.data <- data.frame(x,y)  
> cor.test(~ x+y, data = cor.data)
```

- 위 결과는 모두 앞의 계산결과 같음

3

도수분포표와 교차표



도수분포표와 교차표

- k 개의 범주에 대해서 i 번째 범주에 속할 확률이 p_i 인지 검정하는 문제.
- 얻은 자료를 요약하면 다음과 같이 표로 얻음.

범주	1	2	...	k	합
빈도	O_1	O_2	...	O_k	n

- i 번째 범주에 속할 확률이 p_i 라면 i 번째 범주에 대한 기대 뜻수 E_i 는 $E_i = np_i$.
- 귀무가설 H_0 : ‘ i 번째 범주의 확률은 p_i 이다’ 대 대립가설 H_1 : ‘ p_i 가 아니다’에 대한 검정통계량은

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

도수분포표와 교차표

- 이 검정통계량은 귀무가설이 참일 때 근사적으로 자유도가 $(k - 1)$ 인 카이제곱 분포를 따름.
- 근사적으로 카이제곱분포를 따르는 기준으로 보통 모든 기대도수가 5 이상인 경우.

교차표에서의 독립성

- 각각 r 개와 c 개의 범주를 가진 두 개의 범주형자료의 독립성 검정.
- 두 변수는 편의상 행변수 및 열변수로 부르며 행변수의 i 번째 범주 및 열변수의 j 번째 범주의 관측빈도수를 O_{ij} 로 표현.
- 또한 $O_{i..}$, $O_{.j}$ 로 각각 i 번째 행의 빈도합, j 번째 열의 빈도합이라고 하며 전체 자료수는

$$n = O_{..} = \sum_{i=1}^r O_{i..} = \sum_{j=1}^c O_{.j} = \sum_{i=1}^r \sum_{j=1}^c O_{ij}$$

도수분포표와 교차표

- 로 표시하면 전체 자료는 다음의 표와 같은 교차표를 얻음.

	1	2	...	c	합
1	O_{11}	O_{12}	...	O_{1c}	$O_{1.}$
2	O_{21}	O_{22}	...	O_{2c}	$O_{2.}$
...
r	O_{r1}	O_{r2}	...	O_{rc}	$O_{r.}$
합	$O_{.1}$	$O_{.2}$...	$O_{.c}$	$O_{..}$

도수분포표와 교차표

- 행변수와 열변수가 독립일 때 번째 (i, j) 칸의 기대빈도를 E_{ij} 라고 하면

$$E_i = \frac{O_i \cdot O_j}{O_{..}}$$

- 귀무가설 H_0 : ‘행변수와 열변수가 독립이다’ 가 참일 때 검정통계량

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

의 분포는 근사적으로 자유도가 $(r - 1)(c - 1)$ 인 카이제곱분포가 된다.

적합도 및 독립성 검정을 위한 R 함수

```
chisq.test(x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)),  
           rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)
```

도수분포표와 교차표

- x : 빈도수를 저장한 행렬, table 객체 또는 벡터를 설정하거나 x가 범주형 값을 갖는 변수로 설정.
- y : x가 빈도수를 저장한 행렬인 경우 y는 생략되며 x가 범주형자료를 저장한 벡터일 경우 y도 범주형 자료를 저장한 변수로 설정.
 - > `x <- c(1,1,1,2,2,2) ; y <- c(1,2,1,2,1,2)`
 - > `chisq.test(x,y)`
 - > `chisq.test(table(x,y))`는 같은 결과를 얻음.
- correct : 교차표에서 Yates의 연속형 수정을 할지 설정한다.
TRUE로 설정하면 모든 $|O - E|$ 에서 1/2만큼 뺀 값을 사용함.
`simulate.p.value = TRUE`이면 연속성수정은 하지 않음.
- p : 적합도 검정을 시행할 때 사용하는 것으로 i번째 칸의 확률이 `p[i]`인지 검정함.

도수분포표와 교차표

- rescale.p : TRUE로 설정된 경우 p의 합이 1이 아니면 p를 합이 1이 되도록 보정함.
- simulate.p.value : 논리값을 설정하며 유의확률의 계산을
Monte Carlo simulation으로 할지 설정.
FALSE이면 카이제곱분포를 이용한 근사유의확률값을 계산.
- B : 유의확률계산을 Monte Carlo simulation으로 할 때 시뮬레이션 회수를 정함.

예제 3.21. 주사위를 100번 던져 눈금을 기록하였더니 다음과 같았다.
이 주사위의 각 눈금이 나올 확률이 $1/6$ 이라고 할 수 있겠는가?

눈금	1	2	3	4	5	6	합
빈도	19	16	19	18	14	14	100

도수분포표와 교차표

```
> frq <- c(19, 16, 19, 18, 14, 14)  
> chisq.test(frq)
```

Chi-squared test for given probabilities

```
data: frq  
X-squared = 1.64, df = 5, p-value = 0.8964
```

예제 3.20. 성별과 정당지지도를 조사하였더니 다음과 같은 결과를 얻었다.
성별과 정당지지도가 독립이라고 할 수 있겠는가?

도수분포표와 교차표

	정당 A	정당 B	정당 C	합
남	20	30	15	65
여	30	20	15	65
합	50	50	30	130

```
> obs <- matrix(c(20,30,15,30,20,15), ncol=3, byrow=T)
> chisq.test(obs)
Pearson's Chi-squared test
data: obs
X-squared = 4, df = 2, p-value = 0.1353
```

4

회귀분석



회귀분석

- x의 값이 주어졌을 때 확률변수 Y의 기댓값이 μ 이고, 분산이 σ^2 인 정규분포. 즉, n개의 짹 (x_i, Y_i) 에 대해서

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

- 최소제곱법에 의한 회귀계수 β_0, β_1 의 추정 :

$$S = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

을 최소로 하는 b_0, b_1 을 구함.

- $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})$ 라 하면

$$b_1 = \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad b_0 = \bar{Y} - b_1 \bar{x}$$

- x_i 에서 Y의 예측값 \hat{Y}_i 는

$$\hat{Y}_i = b_0 + b_1 x_i$$

회귀분석

- 제곱합의 분해 :

$$\text{SST} = \sum_{i=1}^n (Y_i - \bar{Y})^2, \text{SSR} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2, \text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \text{ 라고 하면}$$
$$\text{SST} = \text{SSR} + \text{SSE}$$

- 각각의 자유도는

$$(n - 1) = 1 + (n - 2)$$

- 분산분석표

요인	제곱합	자유도	평균제곱	F	유의확률
회귀	SSR	1	MSR=SSR/1	$F_0 = \frac{\text{MSR}}{\text{MSE}}$	$\Pr[F_{1,n-2} > F_0]$
잔차	SSE	$n - 2$	$\text{MSE} = \text{SSE}/(n - 2)$		
전체	SST	$n - 1$			

회귀분석

- 분산분석표에서 $F_0 > F_{1,n-2; \alpha}$ 또는 유의확률 $\Pr[F_{1,n-2} > F_0]$ 이 α 보다 작으면 귀무가설 $H_0: \beta_1 = 0$ 를 기각하고 대립가설 $H_0: \beta_1 \neq 0$ 을 받아들임.

❖ 회귀계수에 대한 추론

- $b_1 \sim N(\beta_1, \frac{\sigma^2}{\text{MSE}/S_{xx}})$, $b_0 \sim N\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\sigma^2}{S_{xx}}\right]\right)$
이며
- $\frac{b_1 - \beta_1}{\sqrt{\text{MSE}/S_{xx}}} \sim t_{n-2}$ 및 $\frac{b_0 - \beta_0}{\sqrt{\text{MSE}\left[\frac{1}{n} + \frac{\sigma^2}{S_{xx}}\right]}} \sim t_{n-2}$
이다.

회귀분석

- 따라서 회귀계수의 $100(1-\alpha)\%$ 신뢰구간은

$$\beta_1 = b_1 \pm t_{n-2;\alpha/2} \sqrt{\text{MSE}/S_{xx}} \text{ 와}$$

$$\beta_0 = b_0 \pm t_{n-2;\alpha/2} \sqrt{\text{MSE} \left[\frac{1}{n} + \frac{\sigma^2}{S_{xx}} \right]}$$

- $H_0 : \beta_1 = \beta_1^0$ 에 대한 검정통계량은
- $t_0 = \frac{b_1 - \beta_1^0}{\sqrt{\text{MSE}/S_{xx}}}$ 이며 각 대립가설에 따른 기각역 및 유의확률은 다음과 같음.

대립가설	기각역	유의확률
$H_1: \beta_1 > \beta_1^0$	$t_0 > t_{n-2;\alpha}$	$\Pr[T_{n-2} > t_0]$
$H_1: \beta_1 < \beta_1^0$	$t_0 < -t_{n-2;\alpha}$	$\Pr[T_{n-2} < t_0]$
$H_1: \beta_1 \neq \beta_1^0$	$ t_0 > t_{n-2;\alpha/2}$	$2\Pr[T_{n-2} > t_0]$

회귀분석

- $H_0 : \beta_0 = \beta_0^0$ 에 대한 검정통계량은
- $t_0 = \frac{b_0 - \beta_0^0}{\sqrt{\text{MSE} \left[\frac{1}{n} + \frac{\sigma^2}{s_{xx}} \right]}}$ 이며 각 대립가설에 따른 기각역 및 유의확률은 다음과 같음.

대립가설	기각역	유의확률
$H_1: \beta_0 > \beta_0^0$	$t_0 > t_{n-2;\alpha}$	$\Pr[T_{n-2} > t_0]$
$H_1: \beta_0 < \beta_0^0$	$t_0 < -t_{n-2;\alpha}$	$\Pr[T_{n-2} < t_0]$
$H_1: \beta_0 \neq \beta_0^0$	$ t_0 > t_{n-2;\alpha/2}$	$2\Pr[T_{n-2} > t_0]$

회귀분석

예제 3.22. 다음은 11명의 나이와 혈중 콜레스테롤 농도를 조사한 자료이다. 나이를 독립변수, 콜레스테롤 수치를 종속변수로 회귀분석을 하여 F 검정통계량, 계수의 신뢰구간, 나이에 따른 종속변수의 기댓값에 대한 신뢰구간, 종속변수의 예측구간을 구해보아라.

```
> age <- c( 54, 69, 43, 39, 64, 52, 47, 34, 73, 37, 45) # lm.test.r  
> c.level <- c(181, 235, 193, 177, 197, 191, 213, 167, 212, 183, 190)  
> cdata <- data.frame(age, c.level)  
# 자료의 수, Sxx, Sxy 등을 계산하여 제곱합, 평균제곱합, 계수의 추정치  
> nn <- length(age) ; my <- mean(c.level); mx <- mean(age)  
> Sxx <- sum(age^2) - nn*mx^2  
> Sxy <- sum(age*c.level) - nn*mx*my  
> SST <- sum(c.level^2) - nn*my^2  
> b1 <- Sxy/Sxx; b0 <- my -b1*mx
```

회귀분석

```
> coef <- c(b0, b1)
> yhat <- b0 + b1* age
> SSE <- sum( (c.level - yhat)^2 )
> SSR <- b1^2*Sxx ; MSR <- SSR
> MSE <- SSE / (nn-2)
> cat("SST = ", SST, "SSR = ", SSR, "SSE = ", SSE, "\n")
SST = 3706.727 SSR = 2098.843 SSE = 1607.884
> R.square <- SSR/SST
cat("R^2 = ", R.square, "\n")
R^2 = 0.5662253
> F0 <- MSR/MSE ; p.val <- 1-pf(F0, 1, nn-2)
> cat("F = ", F0, "P-value = ", p.val, "\n")
F = 11.7481 P-value = 0.007537384
```

회귀분석

각 계수의 검정통계량 계산

```
> se.b1 <- sqrt(MSE / Sxx) ; se.b0 <- sqrt(MSE*(1/nn + mx^2/Sxx))
> cat("SE b1 = ", se.b1, "SE b0 = ", se.b0, "\n")
SE b1 = 0.3213029 SE b0 = 16.76131
```

```
# beta0 =0, beta1 =0에 대한 가설검정
> t0.b1 <- b1/se.b1; t0.b0 <- b0/se.b0
> p.b1 <- 2*(1-pt(t0.b1, nn-2)); p.b0 <- 2*(1-pt(t0.b0, nn-2))
> cat("t for b1 = ", t0.b1, "p.val = ", p.b1, "\n" )
t for b1 = 3.42755 p.val = 0.007537384
> cat("t for b0 = ", t0.b0, "p.val = ", p.b0, "\n" )
t for b0 = 8.274392 p.val = 1.689008e-05
```

회귀분석

각 계수의 신뢰구간 계산

```
# beta0, beta1의 신뢰구간
```

```
> ci.b0 <- c(b0-qt(.975,nn-2)*se.b0, b0+qt(.975,nn-2)*se.b0)
> cat("beta0의 신뢰구간: ", ci.b0, "₩n")
beta0의 신뢰구간: 100.7729 176.6064
```

```
> ci.b1 <- c(b1-qt(.975,nn-2)*se.b1, b1+qt(.975,nn-2)*se.b1)
> cat("beta1의 신뢰구간: ", ci.b1, "₩n")
beta1의 신뢰구간: 0.37444441 1.828119
```

회귀분석

회귀분석을 위한 R 함수

`lsfit(x, y, intercept = TRUE, ...)`

`lm(formula, data, ...)`

- x : $y = b_0 + b_1 * x$ 에서 독립변수 x의 벡터 또는 행렬
- y : 종속변수 y가 저장된 벡터의 이름을 설정한다.
- intercept : b_0 를 0으로 강제로 설정할지 정하는 값이다.
기본값은 `False`이다.
- formula : $y \sim x$ 형태로 모형을 설정한다. 만일 둘 이상의 독립변수를 설정한다면 $y \sim x_1 + x_2$ 등과 같이 설정한다.
- data : formula에 설정한 변수들이 저장된 데이터 프레임의 이름을 설정한다.

회귀분석

```
> lsfit(age, c.level)
```

```
$coefficients
```

```
Intercept      X
```

```
138.689641  1.101282
```

```
(이하 출력 생략)
```

```
> summary(lm(c.level ~ age, data=cdata))
```

```
Call:
```

```
lm(formula = c.level ~ age, data = cdata)
```

```
Residuals:
```

	Min	1Q	Median	3Q	Max
-17.159	-8.108	-4.640	5.259	22.550	

회귀분석

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	138.6896	16.7613	8.274	1.69e-05 ***
age	1.1013	0.3213	3.428	0.00754 **

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

Residual standard error: 13.37 on 9 degrees of freedom

Multiple R-squared: 0.5662, Adjusted R-squared: 0.518

F-statistic: 11.75 on 1 and 9 DF, p-value: 0.007537

5

분산분석



분산분석

- g 개의 정규분포에서 얻은 자료의 평균이 모두 같은지 검정하는 문제.
- i 번째 그룹에서 n_i 개의 자료를 얻고 전체 자료의 수를 n 이라고 하면

$$n = n_1 + n_2 + \dots + n_g$$

- i 번째 그룹의 j 번째 자료를 Y_{ij} 라고 표현하면 Y_{ij} 들은 모두 독립이며
$$Y_{ij} \sim N(\mu_i, \sigma^2)$$
- \bar{Y}_i 는 i 번째 그룹의 표본평균 $\bar{Y}_{..}$ 은 전체 자료의 평균이라고 하면 자료를 다음과 같이 표로 정리.

그룹	자료	평균
1	$Y_{11}, Y_{12}, \dots, Y_{1n_1}$	$\bar{Y}_{1..}$
2	$Y_{21}, Y_{22}, \dots, Y_{2n_2}$	$\bar{Y}_{2..}$
⋮	⋮	⋮
g	$Y_{g1}, Y_{g2}, \dots, Y_{gn_g}$	$\bar{Y}_{g..}$
전체	⋮	$\bar{Y}_{..}$

분산분석

- g 개의 정규분포에서 얻은 자료의 평균이 모두 같은지 검정하는 문제.
- 즉, 귀무가설 $H_0: \mu_1 = \mu_2 = \dots = \mu_g$ 대 대립가설 $H_1: '적어도 하나의 평균은 나머지와 다르다'$ 의 검정.
- 제곱합의 분해

$$\sum_{i=1}^g \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2 = \sum_{i=1}^g \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2 + \sum_{i=1}^g \sum_{j=1}^{n_i} (\bar{Y}_{i.} - \bar{Y}_{..})^2$$
$$\begin{array}{rcl} SST & = & SSE + SST_{\text{Trt}} \\ (n-1) & = & (n-g) + (g-1) \end{array}$$

- 순서대로 전체제곱합, 오차제곱합, 처리제곱합
- 해당 자유도로 나눈 값 : $MSE = SSE/(n-g)$, $MST_{\text{Trt}} = SST_{\text{Trt}}/(g-1)$
- 결과를 분산분석표로 요약.

분산분석

요인	제곱합	자유도	평균제곱	F	유의확률
처리	SSTrt	$g - 1$	$MSTrt = SSTrt/(g - 1)$	$\frac{MSTrt}{MSE}$	$Pr[F_{g-1,n-g} > F]$
오차	SSE	$n - g$	$MSE = SSE/(n - g)$		
전체	SST	$n - 1$			

- 분산분석표의 F값이 $F_{g-1,n-g;\alpha}$ 보다 크거나
- 유의확률이 α 보다 작은 경우 귀무가설 $H_0: \mu_1 = \mu_2 = \dots = \mu_g$ 을 기각

분산분석

R언어의 함수

```
oneway.test(formula, data, subset, na.action, var.equal = FALSE)
```

- formula : response ~ group 과 같은 형식으로 response는 종속변수(반응변수)를 group에는 그룹을 표시하는 독립변수(설명변수)를 설정.
- data : formula에 들어 있는 respons와 group 변수가 있는 데이터 프레임 또는 행렬의 이름을 설정.
- subset : data에서 설정한 데이터 프레임 또는 행렬의 일부 자료만 사용할 경우 이 자료의 조건을 설정.
- var.equal : 등분산성을 가정할지 설정. 기본값은 False이며 일반적으로 분산 분석에서 적용되는 검정은 등분산성을 가정하므로 이 값은 True로 설정함. 등분산성이 가정되지 않는 경우는 Welch 검정을 수행(Welch(1951)). Welch 검정은 심송용(2012) 참고.

분산분석

예제 3.25. 네 가지 비료를 사용하여 얻은 수확량이 다음과 같았다. 비료에 따라 수확량이 차이가 난다고 할 수 있는지 분산분석표를 작성하여 검정해보자.

비료1	11	11	10	10	10	11	10	8	10	9
비료2	12	11	11	13	12	11	11	11	12	9
비료3	12	13	13	11	10	13	11	12	13	11
비료4	9	10	8	10	13	10	10	10	10	8

분산분석

```
y <- c( 11, 11, 10, 10, 10, 11, 10, 8, 10, 9,  
      12, 11, 11, 13, 12, 11, 11, 11, 12, 9,  
      12, 13, 13, 11, 10, 13, 11, 12, 13, 11,  
      9, 10, 8, 10, 13, 10, 10, 10, 10, 8)  
x <- c(rep(1,10), rep(2,10), rep(3,10), rep(4,10))  
fert <- data.frame(y,x)  
oneway.test(y ~x, data = fert, var.equal=T)
```

One-way analysis of means

data: y and x

F = 7.9571, num df = 3, denom df = 36, p-value = 0.0003382

- 유의확률이 0.00034 < 0.05 이므로 평균이 모두 같은 귀무가설을 기각.



수고하셨습니다.

