8장. 통계적 모형론

통계적 모형(模型, model)은 관측 자료를 보는 틀입니다. 통계적 모형의 요건은 단순성과 정합성입니다. 단순하지 않은 모형은 현상을 명쾌하게 설명하지 못합니다. 그러나 모형이 너무 단순하여 현상과 정합적이지 않으면 존립 근거가 없습니다. 일반적으로 단순성과 정합성은 서로 상충적이므로 통계적 모형은 단순성과 정합성의 조화를 통하여 설정됩니다. 그러므로 통계적 모형은 관측 자료와 정합(整合)하는 범위 내에서 가장 단순한 틀이라고 하겠습니다.

이 장에서는 관심 변수를 다른 변수들과 관련시킨 통계적 모형을 소개하고 이에 부수된 통계적 추론을 제시하고자 합니다. 통계적 모형의 기본 철학과 양태를 학습하게 될 것입니다.

8.1절에서는 1원 분산분석 모형과 분석 방법, 그리고 선형대비(線形對比)와 다중비교(多重比較) 등 동시추론에 대하여 설명합니다. 구체적으로 본페로니(Bonferroni), 쉐페(Scheffé), 튜키(Tukey)의 방법 등을 살펴봅니다.

8.2절에서는 선형회귀모형의 설정, 추정, 검증 및 예측에 대하여 간단히 설명합니다. 이어서 8.3절에서는 선형회귀모형을 확장한 일반화선형모형을 제시합니다. 일반화의 과정에서 지수족 확률분포의 역할과 특성을 살펴보고, 특히 일반화선형모형의 추정・검증 등이 어떻게 전개되는지를 보도록 하겠습니다.

이제까지 공부한 수리통계적 도구들이 이 장에서 총동원됩니다. 여러분의 실력을 마음껏 발휘해보십시오.

차례 :

8.1 분산분석모형과 동시추론

이제까지는 거의 대부분의 경우에서 표본전체가 독립적으로 동일분포를 따르는 i.i.d. (independently identically distributed) 상황을 다루었습니만, 앞으로는 그렇지 않습니다. 가장 단순한 경우로서 I개의 모집단에서 독립적으로 표집된 정규분포 표본을 생각하기로 합시다. 즉,

$$Y_{i1}, \cdots, Y_{i\,n_i} \sim \text{i.i.d. } \mathrm{N}(\theta_i, \sigma^2), \ i = 1, \cdots, I \tag{1}$$

이면서 I개 표본들이 독립인 경우를 생각하겠습니다 (총 관측 수는 $\sum_{i=1}^{I} n_i = N$임). 이것은 I개 처리의 평균을 비교하기 위한 실험에서 가장 기본적인 세팅이지요. 물론

$$Y_{i\,j}, \ i = 1, \cdots, I, \ j = 1, \cdots, n_i$$

가 i.i.d.인 것은 아닙니다. $Y_{i\,j}$의 평균이 θ_i로서 i에 의존하니까요.

I개의 모집단에 대한 확률모형으로서 꼭 (1)의 경우만 생각할 수 있는 것은 아닙니다. 그것보다는

$$Y_{i1}, \cdots, Y_{i\,n_i} \sim \text{i.i.d. } \mathrm{N}(\theta_i, \sigma_i^2), \ i = 1, \cdots, I \tag{2}$$

가 더 포괄적이지요.[1] (1)과 (2)를 비교해봅시다.

- (1)이 (2)에 비하여 더 단순하므로 설명하기가 쉽습니다. (1)에서는 집단간 차이가 평균에만 있으나 (2)에서는 집단간 차이가 평균에 뿐만 아니라 분산에도 있기 때문입니다.
- 그러나 (1)은 너무나 단순하여 실제 상황과 정합적이지 않을 수 있습니다. 실제로는 집단의 평균이 클수록 분산도 따라서 큰 경우도 많습니다.

하여튼, 뚜렷한 반증이 없는 한 단순한 모형을 기본으로 합니다 (모형설정 단계에서 반증의 유무를 보기 위하여 유의성 검증을 쓰는 것은 합당하지 않습니다. 기술적(記述的, descriptive) 분석이 더 주효한 도구입니다. 있는 그대로를 묘사하기 위해서는 스케치를 충실히 해야 하는 것과 마찬가지입니다).

여기서는 (1)을 I개의 집단을 비교하기 위한 기본모형으로 생각하기로 합니다. 잘 알려져 있듯이, 이 모형이 바로 1원 분산분석 모형(one-way ANOVA model)입니다. 일단, 기본모형이 정립되면 모형을 구성하는 파라미터에 대한 점추정이 따르게 됩니다. 가장 정통적인 점 추정방법은 최대가능도추정(mle)입니다.

---

1) 다른 방향으로 모형 (1)보다 포괄적인 경우는 $Y_{i1}, \cdots, Y_{i\,n_i}$가 각각 $\mathrm{N}(\theta_i, \sigma^2), \ i = 1, \cdots, I$를 따르되 Y_{ij}와 $Y_{i\,k}$가 상관되어 있는 경우입니다.

자료가 $\{ y_{ij};\ j = 1, \cdots, n_i, i = 1, \cdots, I \}$로 관측된 경우, (1)에 대하여 가능도는

$$L(\theta_1, \cdots, \theta_I, \sigma^2) = \prod_{i=1}^{I} \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi}} (\sigma^2)^{-1/2} \exp\left\{ - \frac{(y_{ij} - \theta_i)^2}{2\sigma^2} \right\}$$

이며 로그 가능도가

$$l(\theta_1, \cdots, \theta_I, \sigma^2) = -\frac{N}{2} \log_e \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \theta_i)^2 + constant$$

로 표현됩니다. 따라서

$$\frac{\partial l}{\partial \theta_i} = \frac{1}{\sigma^2} \sum_{j=1}^{n_i} (y_{ij} - \theta_i) = 0 \quad \Rightarrow \ \hat{\theta}_i = \sum_{j=1}^{n_i} y_{ij} / n_i = \bar{y}_{i.},\ i = 1, \cdots, I$$

를 얻습니다. 상식적인 결과입니다. 한편 σ^2에 대하여는

$$\frac{1}{\sigma^2} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 \ \sim \ \chi^2(n_i - 1), \quad i = 1, \cdots, I\ ;\ 독립적으로$$

$$\Rightarrow \ \frac{1}{\sigma^2} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 \ \sim \ \chi^2 \left(\sum_{i=1}^{I} (n_i - 1) \right)$$

$$\Rightarrow \ \bar{\sigma}^2 = \frac{1}{N - I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 \ (= s^2 \text{으로 표기})$$

을 비편향 추정치로 얻습니다. (이것은 σ^2에 대한 단순한 mle인

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

과는 약간 다릅니다.) 또한 어렵지 않게 $\bar{Y}_i, i = 1, \cdots, I$와 s^2은 독립적으로 분포함을 보일 수 있습니다 (연습문제 8.1).

1원 분산분석 자료는 구체적으로 I개의 처리를 비교하는 실험에서 얻어지는 것이 보통입니다. 그러한 실험에서, 많은 경우, 연구자는 구체적인 연구가설을 갖고 있습니다. 그것은 처리 모평균의 선형결합인 선형대비로 표현됩니다.

선형대비(線形對比, linear contrast; 또는 대비) : 정의

$$\psi = c_1 \theta_1 + \cdots + c_I \theta_I \quad (\text{단, } c_1 + \cdots + c_I = 0).$$

예를 들어 $\psi = \theta_1 - \theta_2$, $\frac{1}{2}(\theta_1 + \theta_2) - \theta_3$, $-3\theta_1 - \theta_2 + \theta_3 + 3\theta_4$ 등은 모두 선형대비입니다.

예를 들어 연구자의 영가설과 연구가설(대안가설)이

$$\mathrm{H}_0 : c_1\theta_1 + \cdots + c_I\theta_I = 0 \quad \text{대} \quad \mathrm{H}_1 : c_1\theta_1 + \cdots + c_I\theta_I > 0 \tag{3}$$

이라고 합시다 (단, $c_1 + \cdots + c_I = 0$). 여기서 $\psi = c_1\theta_1 + \cdots + c_I\theta_I$는

$$L = c_1\bar{y}_{1\cdot} + \cdots + c_I\bar{y}_{I\cdot}$$

로 추정가능한데,

$$L \sim \mathrm{N}\left(c_1\theta_1 + \cdots + c_I\theta_I, \left(\frac{c_1^2}{n_1} + \cdots + \frac{c_I^2}{n_I}\right)\sigma^2\right)$$

을 따르므로

$$\frac{L-\psi}{s\sqrt{\frac{c_1^2}{n_1} + \cdots + \frac{c_I^2}{n_I}}} \sim \mathrm{t}(N-I)$$

라는 t 추축량을 얻게 됩니다. 이것에 근거하여 선형대비 ψ에 관한 가설검증과 신뢰구간을 얻을 수 있습니다. 즉, $t_0 = L/\left(s\sqrt{c_1^2/n_1 + \cdots + c_I^2/n_I}\right)$으로 놓을 때, (3)의 가설문제에 대한 p 값은

$$p-\text{값} = P\{\mathrm{t}(\mathrm{N}-\mathrm{I}) \geq \mathrm{t}_0\}$$

이고 ψ에 대한 신뢰구간은

$$\left(L - t_{N-I,\alpha/2}\, s\sqrt{\frac{c_1^2}{n_1} + \cdots + \frac{c_I^2}{n_I}}\,, L + t_{N-I,\alpha/2}\, s\sqrt{\frac{c_1^2}{n_1} + \cdots + \frac{c_I^2}{n_I}}\right)$$

입니다. (상황에 따라서는 양측 신뢰구간 대신 ψ에 관한 단측(one-sided) 신뢰구간

$$\left(L - t_{N-I,\alpha}\, s\sqrt{\frac{c_1^2}{n_1} + \cdots + \frac{c_I^2}{n_I}}\,, \infty\right)$$

을 활용할 수도 있겠습니다.)

대부분의 경우, 연구자의 관심은 단 하나의 대비인 $\psi = c_1\theta_1 + \cdots + c_I\theta_I$에 그치지 않습니다. 이왕이면 다른 대비들

$$\psi' = c_1'\theta_1 + \cdots + c_I'\theta_I \quad (\text{단, } c_1' + \cdots + c_I' = 0),$$

$$\psi'' = c_1''\theta_1 + \cdots + c_I''\theta_I \quad (\text{단, } c_1'' + \cdots + c_I'' = 0),$$

$$\vdots$$

에 대하여도 동시에 알 수 있으면 좋겠지요. 물론 앞의 절차를 반복하여 가설검증(또는 신뢰구간 추론)을 할 수 있겠지만, 이렇게 다중적 추론을 하다보면 원래 의도되었던 유의수준 또는 신뢰수준을 유지할 수 없게 된다는 문제가 생깁니다. 예컨대,

$I = 4$ 개의 처리를 비교하는 실험에서 3개의 대비

$$\begin{aligned} \psi &= \theta_1 - \theta_2 \\ \psi' &= \theta_3 - \theta_4 \\ \psi'' &= 0.5\,\theta_1 + 0.5\,\theta_2 - 0.5\,\theta_3 - 0.5\,\theta_4 \end{aligned} \tag{4}$$

를 생각해봅시다. 만약 모든 처리 평균들이 같은 경우 (즉, $\theta_1 = \theta_2 = \theta_3 = \theta_4$), 3개의 영가설

$$\mathrm{H}_0 : \psi = 0,\ \mathrm{H}_0' : \psi' = 0,\ \mathrm{H}_0'' : \psi'' = 0$$

중 적어도 하나가 기각될 확률은 1개의 영가설이 기각될 확률인 α 보다 크게 될 것입니다. 그러므로 각 대비를 유의수준 α 로 검증하게 되면 개별적으로는 제1종 오류의 크기가 α 이지만 전체적인 제1종 오류의 크기가 α 보다 훨씬 크게 될 가능성이 있습니다 (또는, ψ, ψ', ψ'' 에 관한 신뢰구간에 대한 신뢰수준이 개별적으로는 $1-\alpha$ 이지만 전체 동시적으로는 $1-\alpha$ 보다 훨씬 작을 수 있습니다).

$K(\geq 2)$ 개의 대비를 동시에 다루면서 전체적 유의수준 (또는 신뢰수준)을 제어하는 방법 두 가지를 소개하기로 하겠습니다. 첫째는 본페로니 방법입니다.

본페로니(Bonferroni) 부등식 : 정리

$A_1, \cdots, A_K$를 각각 $1-\alpha_0$의 확률을 갖는 임의구간이라고 합시다. 즉,

$$P\{A_1\} = 1-\alpha_0, \cdots, P\{A_K\} = 1-\alpha_0.$$

그러면

$$P\left\{\cap_{i=k}^{K} A_k\right\} \geq 1 - K\alpha_0 \ . \tag{5}$$

부등식 (5)의 증명은 다음과 같습니다. $A_1, \cdots, A_K$의 여집합에 대하여

$$P\{\overline{A}_1\} = \alpha_0, \cdots, P\{\overline{A}_K\} = \alpha_0$$

이므로, 부등식

$$P\left\{\cup_{k=1}^{K} \overline{A}_k\right\} \leqq \sum_{k=1}^{K} P\{\overline{A}_k\} = K\alpha_0$$

이 성립합니다. 따라서

$$P\left\{\cap_{k=1}^{K} A_k\right\} = 1 - P\left\{\cup_{k=1}^{K} \overline{A}_k\right\} \geq 1 - K\alpha_0 \ . \qquad \blacksquare$$

본페로니 부등식에 의하여, $K(\geq 2)$ 개의 대비에 관한 동시 신뢰구간이 수준 $1-\alpha$를 확보하기 위하여는 개별 신뢰구간의 수준이 최소한 $1-\alpha/K$여야 함을 알 수 있습니다. 수치 예로서 $I=4,\ n_1 = n_2 = n_3 = n_4 = 5$인 경우를 생각해봅시다. 그리고

대비가 $K=3$ 개라고 합시다. 본페로니 보정에 의하면, 동시 신뢰구간이 수준 95%가 되기 위하여 필요한 개별 신뢰구간은

$$\frac{|\,L-\psi\,|}{s\,\sqrt{\dfrac{c_1^2}{n_1}+\cdots+\dfrac{c_I^2}{n_I}}} \le t_{16,0.025/3}\,(=3.0045)$$

있니다. 참고로, $t_{16,0.025}=2.1199$ 입니다. 그러므로 본페로니 보정에 의하여 구간 폭이 약 1.42배(=3.0045/2.1199)로 커집니다.

다음은 다른 한 방법인 쉐페의 방법입니다.

쉐페(Scheffé)의 동시추론 : 정리

일반적인 코쉬-슈바르츠(Cauchy-Schwarz) 부등식

$$\left(\sum_{i=1}^{I} a_i\,b_i\right)^2 \le \left(\sum_{i=1}^{I} a_i^2\right)\left(\sum_{i=1}^{I} b_i^2\right)$$

을 적용해봅니다.

$$\left(\sum_{i=1}^{I} c_i\,\overline{Y}_{i\cdot}-\sum_{i=1}^{I} c_i\,\theta_i\right)^2 = \left(\sum_{i=1}^{I} c_i\,(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}-\theta_i+\overline{\theta}_{\cdot})\right)^2$$

$$\because\ \sum_{i=1}^{I} c_i = 0 \text{ 이므로, 여기서}$$

$$\overline{Y}_{\cdot\cdot}=\sum_{i=1}^{I} n_i\overline{Y}_{i\cdot}/N,\overline{\theta}_{\cdot}=\sum_{i=1}^{I} n_i\overline{\theta}_i/N\ .$$

$$=\left(\sum_{i=1}^{I}\frac{c_i}{\sqrt{n_i}}\cdot\sqrt{n_i}\,(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}-\theta_i+\overline{\theta}_{\cdot})\right)^2$$

$$\le \left(\frac{c_1^2}{n_1}+\cdots+\frac{c_I^2}{n_I}\right)\cdot\left(\sum_{i=1}^{I} n_i\,(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}-\theta_i+\overline{\theta}_{\cdot})^2\right).$$

따라서

$$\frac{\left(\sum_{i=1}^{I} c_i\,\overline{Y}_{i\cdot}-\sum_{i=1}^{I} c_i\,\theta_i\right)^2}{\dfrac{c_1^2}{n_1}+\cdots+\dfrac{c_I^2}{n_I}} \le \sum_{i=1}^{I} n_i\,(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}-\theta_i+\overline{\theta}_{\cdot})^2\ .$$

$$\therefore\quad \frac{\left(\sum_{i=1}^{I} c_i\,\overline{Y}_{i\cdot}-\sum_{i=1}^{I} c_i\,\theta_i\right)^2}{s^2\left(\dfrac{c_1^2}{n_1}+\cdots+\dfrac{c_I^2}{n_I}\right)} \le \frac{1}{s^2}\sum_{i=1}^{I} n_i\,(\overline{Y}_{i\cdot}-\overline{Y}_{\cdot\cdot}-\theta_i+\overline{\theta}_{\cdot})^2\ .$$

위 식의 우변이 다름이 아닌 $(I-1)\,\mathrm{F}\,(I-1,N-I)$ 분포를 따르므로, 수준

$1-\alpha$ 의 동시신뢰구간은

$$\left(\sum_{i=1}^{I} c_i \overline{Y}_{i\cdot} - \sum_{i=1}^{I} c_i \theta_i\right)^2 \leq (I-1)\cdot F_{I-1,N-I,1-\alpha} \cdot s^2 \left(\frac{c_1^2}{n_1} + \cdots + \frac{c_I^2}{n_I}\right),$$

즉

$$\frac{|L-\psi|}{s\sqrt{\frac{c_1^2}{n_1} + \cdots + \frac{c_I^2}{n_I}}} \leq \sqrt{(I-1)\cdot F_{I-1,N-I,1-\alpha}}$$

로부터 생성됩니다. ■

수치 예로서 $I=4,\ n_1=n_2=n_3=n_4=5$ 인 경우, 대비의 수 K에 관계없이 동시 신뢰구간이 수준 95%가 되기 위하여 필요한 개별 신뢰구간은

$$\frac{|L-\psi|}{s\sqrt{\frac{c_1^2}{n_1} + \cdots + \frac{c_I^2}{n_I}}} \leq \sqrt{3\cdot F_{3,16,0.95}}\ (=3.165)$$

입니다. 따라서 앞의 본페로니 구간에 비하여 약간 느슨한 것처럼 보입니다. 그렇지만 본페로니 구간은 대비의 수 K가 커짐에 따라 늘어나는데 반하여 쉐페 구간은 그렇지 않습니다. 이 점을 감안하면, $I=4,\ n_1=n_2=n_3=n_4=5$ 인 경우에서, K가 3보다 크면 쉐페 구간이 본페로니 구간보다 낫다는 것을 알 수 있습니다.

많은 연구에서 연구자들이 관심을 갖는 대비는 모든 쌍의 처리간 비교에서 나옵니다. 즉, $I=4$ 인 경우 이런 대비는

$$\theta_1-\theta_2,\ \theta_1-\theta_3,\ \theta_1-\theta_4,\ \theta_2-\theta_3,\ \theta_2-\theta_4,\ \theta_3-\theta_4$$

로 모두 6개입니다 (일반적으로 이런 대비의 수는 $I\cdot(I-1)/2$ 개입니다).

튜키의 다중비교(Tukey's Multiple Comparison) : 절차

I개의 모집단에서 독립 표집된 정규분포 표본

$$Y_{i1},\cdots,Y_{in_i} \sim \text{i.i.d. } N(\theta_i,\sigma^2),\ i=1,\cdots,I \tag{6}$$

에서 모집단별 표본크기가 모두 같다고 합시다. 즉,

$$n_1 = \cdots = n_I (= n).$$

이 때 표본평균 $\overline{Y}_{i\cdot}, i=1,\cdots,I$들의 모든 쌍 차이에 관한 최대값인 Q를 다음과 같이 정의합니다:

$$Q = \max_{i,i'} \frac{|\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot} - (\theta_i-\theta_{i'})|}{\frac{s}{\sqrt{n}}}.$$

그러면, Q의 $1-\alpha$ 분위수 $Q_{I,N-I,1-\alpha}$를 활용함으로써 모든 처리 쌍 $i \neq i'$의 평균 차에 관한 동시 신뢰구간으로

$$| \overline{Y}_{i.} - \overline{Y}_{i'.} - (\theta_i - \theta_{i'}) | \leq Q_{I,N-I,1-\alpha} \frac{s}{\sqrt{n}}$$

를 얻을 수 있습니다. ■

Q의 분포는 수리적으로 상당히 복잡한 적분으로 표현되어 여기서 그것을 재현하기는 어렵습니다. 그렇다고 해서 그냥 통계표에서 찾으면 된다고 하면 학문적으로 진지한 자세가 아닙니다. 어떤 방법이 있을까요?

몬테칼로(Monte Carlo)는 어떨까요? Q의 분포가 $\theta_1, \cdots, \theta_I$ 및 σ^2에 관계없다는 것을 쉽게 알 수 있습니다. 따라서 이것들을 모두 $0, \cdots, 0$ 및 1로 두고 컴퓨터를 활용하여 (6)을 만들어내고 Q를 계산해보는 것입니다. 그리고 이런 과정을 N_{repeat}(=999)번 반복해보는 것입니다. 그러면 Q의 경험적 분포가 도출될 테니까 그것으로부터 $1-\alpha$ 분위수 $Q_{I,N-I,1-\alpha}$가 나오겠지요. <그림 1>을 참고해 보십시오.

<표 1>은 수치 예로서, $I=4, n_1=n_2=n_3=n_4=5$인 경우, Q의 상위 5% 분위수를 산출하기 위한 SAS/IML 프로그램입니다. 이번에는 충분히(?) 정확하게 구하기 위하여 N_{repeat} = 99,999번의 몬테칼로 시행으로부터 Q의 상위 5% 분위수를 구해보았습니다. 그랬더니 $Q_{4,16,0.95}=4.043$이 나오는군요. 참고로 통계표에서 찾은 정확한 값은 $Q_{4,16,0.95}=4.046$입니다. 이 정도면 몬테칼로의 힘을 인정할 수 있지 않습니까?

따라서, $I=4, n_1=n_2=n_3=n_4=5$인 경우, 튜키의 다중비교법에 따른 모평균의 쌍 차이(pairwise difference)에 관한 동시 신뢰구간이

$$\frac{| \overline{Y}_{i.} - \overline{Y}_{i'.} - (\theta_i - \theta_{i'}) |}{s \cdot \sqrt{\frac{1}{n} + \frac{1}{n}}} \leq \frac{Q_{I,N-I,1-\alpha}}{\sqrt{2}} \ (= 2.861) \qquad (7)$$

로부터 얻어집니다 (여기서는 통계표 수치인 $Q_{4,16,0.95}=4.046$을 썼습니다). (7)의 우변 수치는 쉐페(Scheffé)로 얻은 3.165에 비교하여 약 10% 정도 작습니다. 일반적으로, 모든 쌍의 평균을 비교하는 경우에서는 튜키의 방법이 쉐페의 방법에 비하여 더 효율적입니다.

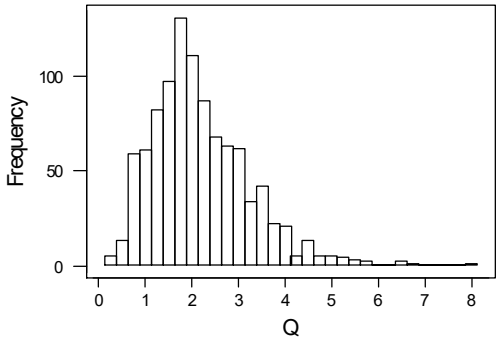
<표 1> Q의 상위 5% 분위수 계산을 위한 SAS/IML 프로그램

```
/* Tukey's Multiple Comparison */
/* tukey.iml                   */

proc iml;
   ll = 4;  n = 5;  Nrepeat = 99999;
   Y = j(ll,n,0);
   PairDiff = j(ll,ll,0);
   Q = j(Nrepeat,1,0);   Qsort = j(Nrepeat,1,0);

   do repeat = 1 to Nrepeat;
      do i =1 to ll;  do j = 1 to n;
         Y[i,j] = rannor(0);
      end;  end;
      Ybar = Y[,+]/n;
      Residual = Y - Ybar*j(1,n,1);
      s = sqrt(ssq(Residual)/(ll*n-ll));
      do i = 1 to ll;   do j = 1 to ll;
         PairDiff[i,j] = abs(Ybar[i] - Ybar[j]);
      end;  end;
      Q[repeat] = max(PairDiff)/s*sqrt(n);
   end;

   R = rank(Q);
   do repeat = 1 to Nrepeat;
      Qsort[R[repeat]] = Q[repeat];
   end;
   Q05 = Qsort[(Nrepeat+1)*0.95];
   print Nrepeat Q05[format=8.3];
quit;
```



<그림 1> Q의 몬테칼로 분포 (반복수 999)

8.2 선형회귀모형

우리는 복잡계(complex system)에 살고 있습니다. 어떤 종류의 결과가 나왔을 때 그것에 대한 원인이 하나인 경우는 없다고 봐야 합니다 (여기서 영향을 주는 변수를 '원인'으로 생각하기로 하겠습니다). 경우에 따라서는 수십, 수백 가지가 원인이 된다고 봐야 합니다. 예로서 <표 2>의 자료를 봅시다. 포도의 수확량 (10월 측정, Y)에 무엇이 원인이 되겠습니까? 한 원인으로서 포도송이 수 (7월 측정, X)가 그것에 영향을 줄 것입니다만, Y에 영향을 주는 것이 꼭 X만은 아니지요. 7월과 10월 사이의 각종 일기(日氣)도 원인이 될 것입니다.

복잡계에서 관심 변수를 Y라고 하고, 이에 영향을 주는 변수로서 $X_2, \cdots, X_p$를 측정하고 모형 안에 고려하기로 합시다. 그리고 $X_2, \cdots, X_p$ 이외에 Y에 영향을 주는 변수를 $Z_{p+1}, Z_{p+2}, Z_{p+3}, \cdots$로 나타내기로 합시다. 그러면

$$Y = f(X_2, \cdots, X_p, Z_{p+1}, Z_{p+2}, Z_{p+3}, \cdots)$$

에 의하여 반응변수 Y가 $X_2, \cdots, X_p$ 및 $Z_{p+1}, Z_{p+2}, Z_{p+3}, \cdots$로부터 결정된다고 할 수 있습니다. 모든 영향 변수들($X_2, \cdots, X_p$ 및 $Z_{p+1}, Z_{p+2}, Z_{p+3}, \cdots$)의 효과가 선형적이라면, 즉 영향변수들의 단위량 변화가 반응변수의 상수적 증가효과를 유발한다면

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \gamma_1 Z_{p+1} + \gamma_2 Z_{p+2} + \gamma_3 Z_{p+3} + \cdots \quad (X_1 \equiv 1)$$

로 쓸 수 있습니다. 실제로 $Z_{p+1}, Z_{p+2}, Z_{p+3}, \cdots$는 측정되지 않기 때문에 위에서와 같이 모형 식에 써넣는 것은 의미가 없습니다. 그런데, 만약

$$\gamma_1 Z_{p+1}, \gamma_2 Z_{p+2}, \gamma_3 Z_{p+3}, \cdots$$

등이 서로 연관되지 않고 동일한 분포로부터 생성된다면 (어느 것이 다른 것을 압도하지 않을 만큼 효과크기들이 비교적 균일하다면) 이것들의 합 ϵ은 대략 정규분포를 따르게 될 것입니다 (중심극한정리). 이에 따라

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2)$$

이라는 통계적 모형이 만들어집니다. 즉, n 개의 관측개체로부터

$$(y_i, x_{i2}, \cdots, x_{ip}),\ i = 1, \cdots, n$$

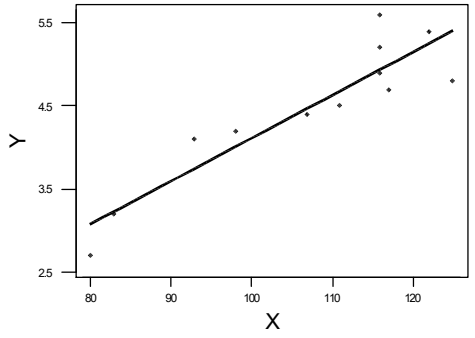
을 얻을 때, 이 자료에 대한 틀로서

$$Y_i \sim \text{독립적으로 } \mathrm{N}(\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \sigma^2),\ i = 1, \cdots, n$$

을 생각하자는 것입니다. 이 모형을 선형회귀(linear regression) 모형이라고 하지요.

<표 2> 포도 자료 : 10월의 산출량(Y)과 7월의 송이 수(X)

Y	X
5.6	116
3.2	83
4.5	111
4.2	98
5.2	116
2.7	80
4.8	125
4.9	116
4.7	117
4.1	93
4.4	107
5.4	122



이 모형에서 로그 가능도가

$$l\,(\beta_0,\beta_1,\cdots,\beta_p,\sigma^2\,)\;=\;-\frac{n}{2}\log_e\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\,y_i-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})^2$$

이므로

$$\frac{\partial\,l}{\partial\beta_j}\;=\;\frac{1}{\sigma^2}\sum_{i=1}^{n}(\,y_i-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})\,x_{ij},\quad j=1,\cdots,p$$

가 됩니다. 따라서

$$\frac{\partial\,l}{\partial\beta_j}\;=\;0\;\;,\;\;j=1,\cdots,p$$

로부터 p 개의 미지수 $\beta_1,\beta_2,\cdots,\beta_p$에 관한 선형체계

$$\begin{aligned}
&\sum_{i=1}^{n}x_{i1}x_{i1}\,\beta_1+\sum_{i=1}^{n}x_{i2}x_{i1}\,\beta_2+\cdots+\sum_{i=1}^{n}x_{ip}x_{i1}\,\beta_p\;=\;\sum_{i=1}^{n}y_i\,x_{i1}\;\;,\\
&\sum_{i=1}^{n}x_{i1}x_{i2}\,\beta_1+\sum_{i=1}^{n}x_{i2}x_{i2}\,\beta_2+\cdots+\sum_{i=1}^{n}x_{ip}x_{i2}\,\beta_p\;=\;\sum_{i=1}^{n}y_i\,x_{i2}\;\;,\qquad(8)\\
&\quad\vdots\qquad\qquad\vdots\qquad\qquad\vdots\qquad\qquad\vdots\\
&\sum_{i=1}^{n}x_{i1}x_{ip}\,\beta_1+\sum_{i=1}^{n}x_{i2}x_{ip}\,\beta_2+\cdots+\sum_{i=1}^{n}x_{ip}x_{ip}\,\beta_p\;=\;\sum_{i=1}^{n}y_i\,x_{ip}
\end{aligned}$$

가 파생됩니다. 이것을 풀어 한꺼번에 $\hat\beta_1,\hat\beta_2,\cdots,\hat\beta_p$을 얻습니다. 또한

$$\frac{\partial l^2}{\partial\beta_j\partial\beta_k} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ij}x_{ik}, \quad j=1,\cdots,p;\ k=1,\cdots,p$$

이므로 $(\beta_1,\beta_2,\cdots,\beta_p)$에 관한 피셔 정보행렬로

$$I(\beta_1,\beta_2,\cdots,\beta_p) = \left(-E\left\{\frac{\partial l^2}{\partial\beta_j\partial\beta_k}\right\}\right)_{j,k} = \left(\frac{1}{\sigma^2}\sum_{i=1}^{n} x_{ij}x_{ik}\right)_{j,k}$$

를 얻습니다.

이후, 분석 및 표현 도구로 요긴한 것이 행렬입니다. 즉,

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} x_{11}\,x_{12}\cdots x_{1p} \\ \vdots \\ x_{i1}\,x_{i2}\cdots x_{ip} \\ \vdots \\ x_{n1}\,x_{n2}\cdots x_{np} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

로 놓으면 선형체계 (8)이 쉽게

$$(X^tX)\boldsymbol{\beta} = X^t\boldsymbol{y}$$

로 표현가능하고 이로부터

$$\hat{\boldsymbol{\beta}} = (X^tX)^{-1}X^t\boldsymbol{y}$$

을 얻게 됩니다 (X^tX에 대한 역행렬이 존재하는 경우). 또한 피셔 정보행렬이

$$I(\boldsymbol{\beta}) = \frac{1}{\sigma^2}X^tX \tag{9}$$

로 표현됩니다. 다음은 선형회귀모형의 분석에서 잘 알려진 몇가지 사실을 정리한 것입니다 (지면의 제한을 구실로 증명을 생략합니다).

선형회귀분석에서의 몇가지 사실 : 정리

1) 오차 ϵ의 분산 σ^2에 대한 비편향 추정치 s^2:

- $s^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_px_{ip})^2$.
- $(n-p)\frac{s^2}{\sigma^2} \sim \chi^2(n-p)$.

2) $\hat{\boldsymbol{\beta}}$의 표집분포 :

- $\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta},\sigma^2(X^tX)^{-1})$.
- $\frac{1}{s^2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^t X^tX(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \sim p\cdot F(p,n-p)$.

3) $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)^t$ 에서의 Y의 기대값인 $\eta(\boldsymbol{x}) \equiv \boldsymbol{x}^t\boldsymbol{\beta}$에 대하여:

- $\hat{\eta}(\boldsymbol{x}) = \boldsymbol{x}^t\hat{\boldsymbol{\beta}}$.
- $\dfrac{\hat{\eta}(\boldsymbol{x}) - \eta(\boldsymbol{x})}{s\sqrt{\boldsymbol{x}^t(X^tX)^{-1}\boldsymbol{x}}} \sim t(n-p)$.

4) $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)^t$ 에서의 Y에 대한 예측값 $Y(\boldsymbol{x})$에 대하여:

- $\hat{Y}(\boldsymbol{x}) = \boldsymbol{x}^t\hat{\boldsymbol{\beta}}$.
- $\dfrac{\hat{Y}(\boldsymbol{x}) - Y(\boldsymbol{x})}{s\sqrt{1+\boldsymbol{x}^t(X^tX)^{-1}\boldsymbol{x}}} \sim t(n-p)$. ■

이 정리에 포함되지 않은 것 중에서 가장 중요한 것은 $\beta_1, \beta_2, \cdots, \beta_p$ 중에서 하나인 β_p에 관한 가설 검증 또는 신뢰구간입니다. 한 방법은

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}, \sigma^2(X^tX)^{-1})$$

를 이용하는 것입니다. 즉

$$X^tX = A = \begin{pmatrix} A_{11} & \boldsymbol{a_1} \\ \boldsymbol{a_1^t} & a_{pp} \end{pmatrix}, \quad (X^tX)^{-1} = C = \begin{pmatrix} C_{11} & \boldsymbol{c_1} \\ \boldsymbol{c_1^t} & c_{pp} \end{pmatrix}$$

라고 놓으면, $AC = I_p$여야 하므로 C의 한 요소 c_{pp}가 다음과 같이 A의 요소들로 표현됩니다 : $c_{pp} = (a_{pp} - \boldsymbol{a_1^t}A_{11}^{-1}\boldsymbol{a_1})^{-1}$. 따라서

$$\hat{\beta}_p \sim \mathrm{N}(\beta_p, \sigma^2 c_{pp}), \quad \text{즉 } \mathrm{N}(\beta_p, \sigma^2(a_{pp} - \boldsymbol{a_1^t}A_{11}^{-1}\boldsymbol{a_1})^{-1}) .$$

그러므로

$$\frac{\hat{\beta}_p - \beta_p}{\dfrac{s}{\sqrt{a_{pp} - \boldsymbol{a_1^t}A_{11}^{-1}\boldsymbol{a_1}}}} \sim \mathrm{t}(n-p) \tag{10}$$

가 유도됩니다. 이 추축량을 이용하여 β_p에 관한 신뢰구간을 구하거나

$$\mathrm{H}_0 : \beta_p = 0 \quad \text{대} \quad \mathrm{H}_1 : \beta_p > 0$$

에 대한 가설검증을 할 수 있습니다. 추축량 (10)은 근사적으로 (큰 n에 대하여)

$$\hat{\beta}_p \sim \mathrm{N}(\beta_p, i^{pp}(\hat{\boldsymbol{\beta}}))$$

표현될 수 있음을 유의하기 바랍니다. 여기서 $i^{pp}(\hat{\boldsymbol{\beta}})$는 $\{\mathrm{I}(\hat{\boldsymbol{\beta}})\}^{-1}$의 (p,p) 요소 (식 (9) 참조). 이것이 선형회귀모형의 일반화(8.3절)에서 중요한 역할을 하게 됩니다.

8.3* 일반화선형모형

선형회귀모형이 많은 경우에 유용하긴 하지만 모든 경우에 적용 가능한 것은 결코 아닙니다. 대표적인 제약은 관심변수 Y를 정규분포로 모형화한다는 데 있습니다. Y가 이항형 반응이라든가 도수(度數, count) 등이라면 정규분포는 안성맞춤이지 않습니다. 그러므로 $(X_2, \cdots, X_p)$에 조건화한 Y의 분포를 베르누이 분포라든가 포아송 분포 등으로 확장할 필요가 있습니다. 정규분포를 포함, 베르누이 분포와 포아송 분포 등은 모두 지수족에 속하는 분포들입니다. 지수족에 대하여는 4.5절에서 다루었습니다만 여기서 다루기 편한 형태로 다시 정의하겠습니다.

지수족(指數族, exponential family) : 정의

$$f_Y(y;\eta,\phi_0) = \exp\left\{\frac{\eta \cdot y - b(\eta)}{a(\phi_0)} + c(y,\phi_0)\right\}, \qquad (11)$$

여기서 $a(\cdot), b(\cdot), c(\cdot,\cdot)$는 모두 알려진 함수이고, ϕ_0는 알려진 파라미터 또는 장애모수(nuisance parameter)입니다.

포아송 분포 Poisson(θ)의 확률함수는

$$f_Y(y;\theta) = \frac{\theta^y}{y!} e^{-\theta} = \exp\{\log_e \theta \cdot y - \theta - \log_e y!\}, \quad \theta > 0$$

이기 때문에 $\eta = \log_e \theta$로 놓으면

$$f_Y(y;\eta) = \exp\{\eta \cdot y - e^{\eta} - \log_e y!\}, \quad -\infty < \eta < \infty$$

가 됩니다. 포아송 분포의 경우에는 $\eta = \log_e \theta$, $b(\eta) = e^{\eta}$ 입니다 ($a(\phi_0) \equiv 1$).

베르누이 분포 Bernoulli(θ)의 확률함수는

$$\begin{aligned} f_Y(y;\theta) &= \theta^y (1-\theta)^{1-y} \\ &= \exp\{\log_e \theta \cdot y + \log_e (1-\theta) \cdot (1-y)\} \\ &= \exp\{[\log_e \theta - \log_e (1-\theta)] \cdot y + \log_e (1-\theta)\}, \ 0 < \theta < 1 \end{aligned}$$

입니다. 따라서 $\eta = \log_e [\theta/\{1-\theta\}]$로 놓으면

$$f_Y(y;\eta) = \exp\{\eta \cdot y - \log_e (1+e^{\eta})\}, \quad -\infty < \eta < \infty$$

가 됩니다. 즉 이 경우에는 $b(\eta) = \log_e (1+e^{\eta})$ 입니다. 그리고 $a(\phi_0) \equiv 1$.

두 경우 모두, 원 파라미터 θ는 제한된 범위의 값만 취하지만 변환 파라미터 η는 실수값 전체를 취할 수 있음에 유의하십시오.

마지막으로, 정규분포 $N(\theta, \sigma_0^2)$의 경우는

$$f_Y(y;\theta,\sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{1}{2\sigma_0^2}(y-\theta)^2\right\}$$

$$= \exp\left\{\frac{\theta\cdot y - \frac{1}{2}\theta^2}{\sigma_0^2} - \frac{y^2}{2\sigma_0^2} - \frac{1}{2}\log_e[2\pi\sigma_0^2]\right\},\ -\infty < \theta < \infty$$

이므로, $\eta = \theta$ 이고, $b(\eta) = \frac{1}{2}\eta^2$ 입니다. 그리고 $\phi_0 = \sigma_0^2$ 이고 $a(\phi_0) \equiv \phi_0$.

지수족 분포의 적률생성함수와 평균 : 정리

1) $m_Y(t;\eta,\phi_0) = \exp\left\{\frac{b(\eta + t\,a(\phi_0)) - b(\eta)}{a(\phi_0)}\right\}$.

2) $E(Y:\eta,\phi_0) = b'(\eta)$.

증명은 다음과 같습니다.

$$m_Y(t;\eta,\phi_0) = E\{\exp(tY);\eta,\phi_0\}$$

$$= \int \exp(ty)\cdot\exp\left\{\frac{\eta y - b(\eta)}{a(\phi_0)} + c(y,\phi_0)\right\}dy$$

$$= \int \exp\left\{\frac{[a(\phi_0)t+\eta]y - b(\eta)}{a(\phi_0)} + c(y,\phi_0)\right\}dy$$

$$= \exp\left\{\frac{b(\eta + t\,a(\phi_0)) - b(\eta)}{a(\phi_0)}\right\}$$

$$\cdot\int \exp\left\{\frac{[a(\phi_0)t+\eta]y - b(a(\phi_0)t+\eta)}{a(\phi_0)} + c(y,\phi_0)\right\}dy$$

$$= \exp\left\{\frac{b(\eta + t\,a(\phi_0)) - b(\eta)}{a(\phi_0)}\right\}\cdot 1,$$

여기서 확률변수 Y가 이산형인 경우에는 $\int$ 을 $\sum$ 로 대치하면 됩니다. 한편, Y의 평균은 적률생성함수로부터 다음과 같이 구할 수 있습니다.

$$\frac{d}{dt}m_Y(t;\eta,\phi_0) = b'(\eta + t\,a(\phi_0))\cdot\exp\left\{\frac{b(\eta+t\,a(\phi_0)) - b(\eta)}{a(\phi_0)}\right\}.$$

$$\therefore\quad E(Y:\eta,\phi_0) = \frac{d}{dt}m_Y(t;\eta,\phi_0)\,|_{\,t=0} = b'(\eta).$$

참고로 Y의 분산은 $Var(Y:\eta,\phi_0) = b''(\eta)\,a(\phi_0)$ 입니다 (연습문제 8.4). ■

일반화 선형모형(generalized linear model)은 선형회귀모형의 일반화로서 지수족 분포 (11)을 확률적 기반으로 합니다. 그리고 파라미터 η를 설명변수 $X_1, \cdots, X_p$의 선형결합(= 선형예측식, linear predictor)으로 표현합니다. 즉

$$\eta = \beta_1 X_1 + \cdots + \beta_p X_p, \quad X_1 \equiv 1.$$

그런데 Y의 평균 $\mu \equiv E\{Y;\eta\}$가 $b'(\eta)$이므로 결국 η를 μ로, μ를 η로 표현할 수 있습니다. 그것을

$$\eta = g(\mu), \quad \mu = h(\eta) (\equiv b'(\eta))$$

로 표현하기로 합시다. 포아송 분포 Poisson(θ)의 경우는 $\mu = \theta$이기 때문에

$$\eta = \log_e \mu (= g(\mu)), \quad \mu = e^{\eta} (= h(\eta))$$

입니다. 베르누이 분포 Bernoulli(θ)의 경우도 $\mu = \theta$입니다. 따라서

$$\eta = \log_e [\mu/(1-\mu)] (= g(\mu)), \quad \mu = e^{\eta}/[1+e^{\eta}] (= h(\eta))$$

입니다. 마지막으로, 정규분포 $N(\theta, \sigma_0^2)$의 경우는 $\mu = \theta = \eta$입니다. 일반적으로 함수 $\eta = g(\mu)$를 연결함수(link function)라고 합니다.

일반화 선형모형(generalized linear model) : 정의

n개의 관측개체로부터 얻은 자료

$$(y_i, x_{i2}, \cdots, x_{ip}), \ i = 1, \cdots, n$$

에 대하여

$$Y_i \sim \text{독립적으로 특정 지수족}(\eta_i, \phi_0), \quad i = 1, \cdots, n.$$

여기서

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \ (x_{i1} \equiv 1).$$

더 일반적으로 확장할 수도 있습니다 (McCullagh and Nelder, 1989). ■

이제 일반화 선형모형에서의 p개 파라미터 $\beta_1, \beta_2, \cdots, \beta_p$를 추정해봅시다. 로그가능도 함수가

$$l = \sum_{i=1}^{n} \frac{\eta_i y_i - b(\eta_i)}{a(\phi_0)}, \quad \text{단 } \eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

입니다. 따라서 연쇄법칙을 활용하여

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{1}{a(\phi_0)} \sum_{i=1}^{n} \{ y_i - b'(\eta_i) \} x_{ij}, \quad j = 1, \cdots, p$$

를 얻습니다. 또한

$$\frac{\partial^2 l}{\partial\beta_j\,\partial\beta_k} = \frac{\partial}{\partial\beta_k}\left(\frac{\partial l}{\partial\beta_j}\right) = \sum_{i=1}^{n}\frac{\partial}{\partial\eta_i}\left(\frac{\partial l}{\partial\beta_j}\right)\cdot\frac{\partial\eta_i}{\partial\beta_k}$$

$$= \frac{1}{a(\phi_0)}\sum_{i=1}^{n}\frac{\partial}{\partial\eta_i}\left(\sum_{i=1}^{n}\{y_i - b'(\eta_i)\}x_{ij}\right)\cdot\frac{\partial\eta_i}{\partial\beta_k}$$

$$= -\frac{1}{a(\phi_0)}\sum_{i=1}^{n}b''(\eta_i)\cdot x_{ij}x_{ik}$$

가 됩니다. 따라서 $\beta_1,\beta_2,\cdots,\beta_p$ 의 mle를 구하기 위한 뉴튼-라프슨 알고리즘에서 반복식은 다음과 같게 됩니다.

$$\boldsymbol{\beta}_{\neq w} = \boldsymbol{\beta}_{\text{old}} + I(\beta_{\text{old}})^{-1}\,U(\boldsymbol{\beta}_{\text{old}})\,,$$

여기서

$$U(\beta) = \begin{pmatrix} U_1 \\ \vdots \\ U_p \end{pmatrix},\ p\times 1\ ; \qquad I(\beta) = \begin{pmatrix} I_{11}\cdots I_{1p} \\ \vdots \quad\ \vdots \\ I_{p1}\cdots I_{pp} \end{pmatrix},\ p\times p\ ;$$

$$U_j = \frac{\partial l}{\partial\beta_j} = \frac{1}{a(\phi_0)}\sum_{i=1}^{n}(y_i - b'(\eta_i))\,x_{ij}\,,$$

$$I_{jk} = -\frac{\partial^2 l}{\partial\beta_j\,\partial\beta_k} = \frac{1}{a(\phi_0)}\sum_{i=1}^{n}b''(\eta_i)\cdot x_{ij}x_{ik}\,,$$

$$\eta_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \qquad j,k = 1,\cdots,p.$$

최종적으로 얻게 되는 $\hat{\boldsymbol{\beta}}$ 은 최대가능도 이론에 따라

$$\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p(\boldsymbol{\beta}, [I(\hat{\boldsymbol{\beta}})]^{-1})\,, \quad \text{근사적으로}$$

분포합니다. 그리고 $\hat{\beta}_p$ 은 근사적으로

$$\hat{\beta}_p \sim \mathrm{N}(\beta_p, i^{pp}(\hat{\boldsymbol{\beta}}))$$

을 따르게 됩니다. 여기서 $i^{pp}(\hat{\boldsymbol{\beta}})$ 은 피셔 정보행렬 $[I(\hat{\boldsymbol{\beta}})]^{-1}$ 의 (p,p) 요소임.

예를 들어보기로 하지요. <표 3>은 딱정벌레를 대상으로 한 독성실험 자료입니다. 반응변수 Y는 1(사망) 또는 0(생존)이고 설명변수 X는 독성수준입니다. 이에 관한 모형으로

$$Y=1 \mid X=x \sim \text{Bernoulli}(\theta(x)), \text{ 독립적으로} \tag{12}$$

$$\text{여기서} \quad \eta(x) \equiv \log_e\frac{\theta(x)}{1-\theta(x)} = \beta_1 + \beta_2 x \tag{13}$$

<표 3> 독성실험 자료

X	Y	count
1.691	1	6
1.691	0	53
1.724	1	13
1.724	0	47
1.755	1	18
1.755	0	44
1.784	1	28
1.784	0	38
1.811	1	52
1.811	0	11
1.837	1	53
1.837	0	6
1.861	1	61
1.861	0	1
1.884	1	60
1.884	0	0

를 고려합시다. (12)와 (13)에 의한 일반화선형모형이 로짓(logit) 모형입니다. 즉,

$$\theta(x) = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)} \quad .$$

베르누이 분포의 경우에는 $b(\eta) = \log_e(1 + e^{\eta})$ 이므로

$$b'(\eta) = e^{\eta}(1 + e^{\eta})^{-1}, \quad b''(\eta) = e^{\eta}(1 + e^{\eta})^{-2}$$

입니다. <표 4>은 독성실험 자료의 로짓 모형 분석을 위한 SAS/IML 프로그램입니다. 그것을 통하여 계산된 β_1 와 β_2 의 최대가능도추정치와 근사적 표준오차는 다음과 같습니다.

$$\hat{\beta_1} = -60.44, \quad \text{s.e.}(\hat{\beta_1}) = 5.14,$$
$$\hat{\beta_2} = 34.06, \quad \text{s.e.}(\hat{\beta_2}) = 2.88.$$

따라서 $H_0 : \beta_2 = 0$ 대 $H_1 : \beta_2 > 0$ 에 대한 근사적 검증은

$$z = \frac{\hat{\beta_2}}{\text{s.e.}(\hat{\beta_2})} = \frac{34.06}{2.88} = 11.83$$

에 의하여 수행될 수 있습니다 (영가설 하에서 $z \sim N(0,1)$). (로짓 모형에서의 정확검증은 좀 복잡합니다. 그래도 관심이 있는 학생은 내게 e-mail을 보내기 바랍니다.)

<표 4> 독성실험 자료의 로짓 모형 분석을 위한 SAS/IML 프로그램

```
/* Logit Model Fitting    */
/* File Name : logit.iml  */

proc iml;
    X = {1  1.691,     1  1.691,
         1  1.724,     1  1.724,
         1  1.755,     1  1.755,
         1  1.784,     1  1.784,
         1  1.811,     1  1.811,
         1  1.837,     1  1.837,
         1  1.861,     1  1.861,
         1  1.884,     1  1.884};
    y = {1,  0,  1,  0,  1,  0,  1,  0,  1,  0,  1,  0,  1,  0,  1,  0};
    count = {6, 53, 13, 47, 18, 44, 28, 38, 52, 11, 53,  6, 61,  1, 60,  0};

    beta_old = {0, 0};   tol = 1;   repeat=1;

    do while (tol > 0.000001);
       eta = X* beta_old;
       bprime1 = exp(eta)/(1+exp(eta));
       bprime2 = bprime1/(1+exp(eta));
       I = X`*diag(count#bprime2)*X;
       Iinverse = inv(I);
       U = X`*diag(count)*(y-bprime1);
       beta_new = beta_old + Iinverse*X`*diag(count)*(y-bprime1);
       diff = beta_new - beta_old;
       tol = ssq(diff);
       print repeat beta_new[format=8.4];
       repeat = repeat+1;
       beta_old = beta_new;
    end;

   loglik = 2*(eta`*diag(count)*y - count`*log(1+exp(eta)));
   se = sqrt(vecdiag(Iinverse));
   print loglik[format=8.4] se[format=8.4];
quit;
```

8.A 연습문제

8.1 1원 분산분석 모형에서 $\overline{Y}_i, i = 1, \cdots, I$와 s^2은 독립적으로 분포함을 보이세요.

[힌트 : $\overline{Y}_i, i = 1, \cdots, I$와 $Y_{ij} - \overline{Y}_i, i = 1, \cdots, I, j = 1, \cdots, n_i$를 전체자료 $Y_{ij}, i = 1, \cdots, I, j = 1, \cdots, n_i$의 선형결합으로 표현함으로써 독립임을 보이고 이를 활용하여 본 정리를 증명해보세요.]

8.2 처리 수가 $I = 6$이고 처리당 반복 수가 6인, $n_1 = n_2 = n_3 = n_4 = n_5 = n_6 = 6$인 1원 분산분석 모형에서 Q의 상위 5% 분위수를 <표 1>의 SAS/IML 프로그램을 활용하여 산출해보십시오. 그리고 통계표에서 얻는 Q의 상위 5% 분위수 4.302와 비교해보십시오.

8.3 선형회귀모형을 분석할 수 있는 SAS/IML 프로그램을 작성하고 적절한 예제에 적용해보세요.

8.4* 지수족 분포에서 Y의 분산이 $Var(Y : \eta, \phi_0) = b''(\eta)\, a(\phi_0)$로 표현됨을 증명하세요.

8.5* 다음 자료에서 Y는 포아송 도수이고 X는 이와 관련 있는 것으로 생각되는 설명변수입니다. 적절한 일반화 선형모형을 세우고 분석결과를 제시하시오.

Y	X
42	2.7
37	3.1
1	2.0
101	2.6
73	2.7
14	2.5

[힌트 : <표 4>의 SAS/IML 프로그램을 수정하여 사용하면 됩니다. 포아송 파라미터 $\theta(x)$에 대하여, $\log_e \theta(x) = \beta_1 + \beta_2 x$를 자료에 적합해보십시오.

[답: $\hat{\beta}_1 = 0.989$, s.e.$(\hat{\beta}_1) = 0.551$; $\hat{\beta}_2 = 1.059$, s.e.$(\hat{\beta}_2) = 0.203$]

8.B 읽을만한 책

분산분석과 선형회귀 모형에 관한 참고문헌은 도저히 수를 헤아릴 수 없을 만큼 많지만 수리통계학과 관련하여 조금만 더 알고 싶다면 다음 책을 보시기 바랍니다.

- Hogg, R.V. and Craig, A.T. (1995) *Introduction to Mathematical Statistics*, 5th Edition. Prentice Hall. (Chapter 10)
- Bickel, P.J. and Duksum, K.A. (1977) *Mathematical Statistics*.. Holden-Day. (Chapter 7)
- Rice, J.A. (1995) *Mathematical Statistics and Data Analysis*, 2nd Edition. Duxbury Press. (Chapters 12 and 14)
- Casella, G. and Berger, R.L. (1990) *Statistical Inference*. Duxbury. (Chapters 11 and 12)

일반화 선형모형에 관하여는 다음 책을 권합니다.

- 허명회 (1993) 「선형모형방법론」 자유아카데미. (5장)
- McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, 2nd Edition. Chapman and Hall.