제 5장 ANOVA

§5.1 서론

ANOVA는 "analysis of variance"의 약자인데, 이를 직역하면 "분산분석"이 된다. 그러나, 관심의 대상은 모분산이 아니라 모평균이다. 다만, 모평균에 대한 가설을 검정할 때 표본의 분산을 도구로 사용할 따름이다.

ANOVA는 다수의 모집단이 있을 때 사용하는데, 기본적인 가정은 첫째로 모분포들이 모두 정규분포이고, 둘째로 모분산들이 (알려지지는 않았지만) 모두 동일하다는 것이다. 그러니까, 모평균 $\mu_1, \mu_2, \mu_3, \cdots$ 는 서로 다를 수도 있는데, ANOVA는 바로

$$\begin{aligned} H_0 &: \mu_1=\mu_2=\mu_3=\cdots \\ H_a &: Not\ H_0 \end{aligned} \tag{5.1.1}$$

를 검정하는 것이다. (비고: "$\sigma_1^2=\sigma_2^2=\sigma_3^2=\cdots$"를 H_0와 H_a에 포함시켜도 무방함.) 즉, ANOVA는 §4.4.3의 뒷부분에서 간단히 다루었던

$$\begin{aligned} H_0 &: \mu_1=\mu_2 \ (\text{그리고, } \sigma_1^2=\sigma_2^2) \\ H_a &: \mu_1\neq\mu_2 \ (\text{그리고, } \sigma_1^2=\sigma_2^2) \end{aligned} \tag{5.1.2}$$

경우를 확장한 것이다 (<비고 4.4.7> 참조).

§4.6.2와 §4.6.5에서, 모비율에 대한 TTT인 Z-test를 확장하면 UTT인 카이제곱 검정이 되는 것을 보았다. 이와 같이, 식 (5.1.2)에 대한 TTT인 T-test를 확장하면 식 (5.1.1)에 대한 UTT인 F-test가 된다 (<비고 4.4.6> 참조). T-test가 F-test로 확장되는 것은 식 (2.5.7)에 의한 것이다. 그리고, F-test가 UTT인 이유는 이미 등장한 식 (4.4.10)에서도 찾을 수 있다. 즉, 식 (4.4.10)에서 "$t\geq\sqrt{k'}$ 또는 $t\leq\sqrt{k'}$"은 "$t^2\geq k'$"과 동일한데, 전자는 T-test의 기각역이고 후자는 F-test의 기각역이다.

§5.2에서는 식 (5.1.2)에 대한 검정을 정식으로 다루고, §5.3에서는 식 (5.1.1)에 대한 검정을 다룬다. 그리고, §5.4 이후에서는 보다 효과적으로 식 (5.1.1)을 검정하는 방법을 소개한다. 사실, 이 책에서 ANOVA라고 부르는 것은 소위 (통계적) 실험계획법(experimental design)의 범주에 속하는 것인데, 이 책에서는 효과적인 검정을 위한 실험계획법 중에서 가장 기본적인 경우만 소개한다. 또한, ANOVA는 소위 선형(통계)모형(linear model)의 틀에 속하기도 하는데, 6장에서 다룰 Linear Regression 역시 선형모형이다.

§5.2 T-test for Independent Samples

ANOVA를 거론하기에 앞서 준비작업 삼아 식 (5.1.2)에 대한 T-test를 정식으로 다룬다.

젖소용 사료가 두 종류가 있는데, 사료 i를 먹인 젖소의 우유생산량을 Y_i라 하자 ($i=1,2$). Y_i에 대한 가정은

$$Y_i \sim N(\mu_i, \sigma^2) \tag{5.2.1}$$

이다 (비고: $\sigma_1^2=\sigma_2^2\equiv\sigma^2$). 식(5.1.2)의 가설을 검정하기 위해서 모집단 i에서 크기가 n_i인 표본 $\{Y_{i1}, Y_{i2}, \cdots, Y_{i,n_i}\}$를 추출한다고 하자 ($i=1,2$).

<비고 5.2.1> 이 절의 제목에 "independent samples"라는 표현이 사용된 이유는 서로 다른 모집단에서 따로따로 추출된 두 표본은 서로 독립이기 때문이다.

두 표본의 관찰치를 $\{y_{i1}, y_{i2}, \cdots, y_{i,n_i}\}$, $i=1,2$ 라 하면 LF는 식 (3.2.1)을 확장한 형태인

$$L(\mu_1,\mu_2,\sigma^2)=\left\{\prod_{i=1}^{2}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n_i}\right\}\cdot e^{-\frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2/\sigma^2} \tag{5.2.2}$$

이다. 즉, LF는 식 $\{Y_{11}, \cdots, Y_{1,n_1}\}$의 결합 밀도함수와 $\{Y_{21}, \cdots, Y_{2,n_2}\}$의 결합 밀도함수의 곱이다 (<비고 5.2.1> 참조).

귀무가설 하에서는 "$\mu_1=\mu_2$"이므로, 식 (5.2.2)에서 μ_1과 μ_2를 μ로 대체한 다음 (μ에 대해서 편미분하여) μ에 대한 최우추정치를 구하면

$$\hat{\mu}=\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}y_{ij}}{\sum_{i=1}^{2}n_i}\ (=\bar{y}) \tag{5.2.3}$$

을 얻는데, 이는 두 표본을 합친 평균 관찰치이다. 다음, 식 (5.2.2)에 식 (5.2.3)을 대입(하고 나서 σ^2에 대해서 편미분)하면, σ^2에 대한 최우추정치로

$$\widehat{\sigma_0^2}=\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2}{n},\ n=\sum_{i=1}^{2}n_i \tag{5.2.4}$$

를 얻는다. 반면에, (귀무가설의 제약이 없는) 전체 모수공간에서의 최우추정치는 식 (5.2.2)를 μ_1, μ_2, σ^2에 대해서 편미분해서 얻는데, 결과는 다음과 같다 (§3.2.1 참조).

$$\widehat{\mu}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} (\equiv \overline{y_i}),\ i=1,2 \tag{5.2.5}$$

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i} (y_{ij} - \overline{y_i})^2}{n},\ n = \sum_{i=1}^{2} n_i \tag{5.2.6}$$

위에서 구한 최우추정치들을 LF인 식 (5.2.2)에 대입하여 LRT의 기각역을 구하면 다음과 같다.

$$\lambda = \left(\frac{\widehat{\sigma^2}}{\widehat{\sigma_0^2}} \right)^{\frac{n}{2}} \le k,\ n = \sum_{i=1}^{2} n_i \tag{5.2.7}$$

<비고 5.2.2> λ의 형태는 식 (4.4.9)와 같다. 다만, $\widehat{\sigma_0^2}$과 $\widehat{\sigma^2}$의 내용이 복잡해졌을 뿐이다.

λ를 손질하면 (자세한 내용은 §5.3.1 참조), 다음과 같이 TTT의 기각역을 얻는다.

$$t^2 = \left\{ \frac{\overline{y_1} - \overline{y_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right\}^2 \ge k' = (n-2)(k^{-\frac{2}{n}} - 1) \tag{5.2.8}$$

그리고, $\overline{y_i}$를 $\overline{Y_i}$로 대체하고 s^2을 식 (3.7.3)의 S^2으로 대체하면, 검정통계량으로

$$T_{n-2} = \frac{\overline{Y_1} - \overline{Y_2}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},\ n = n_1 + n_2 \tag{5.2.9}$$

을 얻는데, 이는 귀무가설 하에서 자유도가 $n-2$인 t분포를 따른다 (<비고 3.7.2> 참조). (비고: 식 (3.7.4)에 귀무가설 "$\mu_1 = \mu_2$"를 대입하면 식 (5.2.9)를 얻음. 단, 지금은 $(\overline{X}, \overline{Y})$ 대신에 $(\overline{Y_1}, \overline{Y_2})$로 표본평균을 표기하고 있음.)

예를 들어, $\{y_{11}, y_{12}, y_{13}, y_{14}\} = \{3,2,1,2\}$이고 $\{y_{21}, y_{22}, y_{23}, y_{24}\} = \{5,2,4,5\}$라 하자. 먼저, 표본평균과 표본분산을 따로따로 구하면 다음과 같다.

$$\overline{y_1}=\frac{3+2+1+2}{4}=\frac{8}{4}=2,\quad \overline{y_2}=\frac{5+2+4+5}{4}=\frac{16}{4}=4$$

$$s_1^2=\frac{(3-2)^2+(2-2)^2+(1-2)^2+(2-2)^2}{4-1}=\frac{2}{3} \tag{5.2.10}$$

$$s_2^2=\frac{(5-4)^2+(2-4)^2+(4-4)^2+(5-4)^2}{4-1}=\frac{6}{3}$$

그리고, 이로부터 $\overline{y}$와 s^2을 다음과 같이 얻는다.

$$\overline{y}=\frac{4\overline{y_1}+4\overline{y_2}}{4+4}=\frac{8+16}{4+4}=3$$

$$s^2=\frac{(4-1)s_1^2+(4-1)s_2^2}{(4-1)+(4-1)}=\frac{2+6}{3+3}=\frac{8}{6} \tag{5.2.1}$$

다음, 식 (5.2.8)에 $\overline{y_1}=2$, $\overline{y_2}=4$, $s=\sqrt{8/6}$, $n_1=n_2=4$를 대입하면 $t=-2.449$를 얻는다. 그런데, 자유도가 $n-2=6$인 T-test에서 $\alpha=0.05$에 대한 TTT의 기각역은 "$t\geq 2.447$ 또는 $t\leq -2.447$"이므로 (또는, $t^2>-5.99$), 유의수준 5%에서 귀무가설을 (가까스로나마) 기각한다. (비고: 가까스로 기각하므로 p-value는 0.05보다 약간 작은 값이 될 것임. §4.6.3 참조.)

§5.3 One-Way ANOVA

5.3.1 One-Way ANOVA에 대한 LRT

§5.2의 " T-test for Independent Samples"를 모집단이 셋 이상인 경우로 확장한 것을 "One-Way ANOVA"라 하는데, 이는 §5.5에 등장할 "Two-Way ANOVA"와 구별하기 위한 명칭이다.

모집단의 개수를 I라 하자. 그러면, 귀무가설은 "$\mu_1 = \mu_2$"에서 "$\mu_1 = \mu_2 = \cdots = \mu_I$"로 확장되는데, 이때 식 (5.2.1)~(5.2.7)에서 달라지는 것은 "$i = 1, 2$"가 "$i = 1, 2, \cdots, I$"로 달라지는 것 한가지 뿐이다. 즉, $\prod_{i=1}^{2}$와 $\sum_{i=1}^{2}$를 각각 $\prod_{i=1}^{I}$와 $\sum_{i=1}^{I}$로 바꾸기만 하면 된다. 이제 식 (5.2.7)을 손질해서 편리한 형태로 고치겠는데, 손질한 후에 $I = 2$를 대입하면 식 (5.2.8)이 된다.

먼저, 앞으로 사용할 용어를 다음과 같이 정의한다.

SS : sum of squares

$$TSS \text{ (total } SS) = n\,\widehat{\sigma}_0^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 \qquad (5.3.1)$$

$$SSE \text{ (SS for error)} = n\,\widehat{\sigma}^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2 \qquad (5.3.2)$$

$$SSTr \text{ (SS for treatments)} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\overline{y_i} - \overline{y})^2 = \sum_{i=1}^{I} n_i(\overline{y_i} - \overline{y})^2 \qquad (5.3.3)$$

<비고 5.3.1> §3.3에서는 $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$을 SS라 불렀음.

<비고 5.3.2> 식 (5.3.1)~(5.3.3)에서 y_{ij} (및 $\overline{y_i}$와 $\overline{y}$)를 확률변수 Y_{ij} (및 $\overline{Y_i}$와 $\overline{Y}$)로 대체하더라도 여전히 동일한 호칭 (TSS, SSE, $SSTr$)을 사용함 (<비고 3.1.1> 참조).

<비고 5.3.3> "$TSS = SSE + SSTr$"가 성립할 뿐더러 (증명은 생략함), 이들이 확률변수일 때 (<비고 5.3.2> 참조) SSE와 $SSTr$은 서로 독립이다.

SSE와 $SSTr$이 독립인 이유는 정규 모분포가 하나일 때 $\sum_{i=1}^{n}(Y_i-\overline{Y})^2$와 $\overline{Y}$가 독립인 (<비고 2.15.2> 참조) 이유와 동일하다. 또한, $I=2$인 경우에는

$$SSE=(n_1+n_2-2)S^2 \tag{5.3.4}$$

$$SSTr=(\overline{Y_1}-\overline{Y_2})^2\Big/\left(\frac{1}{n_1}+\frac{1}{n_2}\right) \tag{5.3.5}$$

가 되는데, 이들은 각각 식 (5.2.9)의 분모와 분자에 등장한다. (비고: SSE와 $SSTr$이 독립이면 S와 $(\overline{Y_1}-\overline{Y_2})$도 독립. <비고 2.11.2> 참조.)

식 (5.3.1)과 (5.3.2)를 식 (5.2.7)에 대입하면

$$\lambda=\left(\frac{SSE/n}{TSS/n}\right)^{\frac{n}{2}}\le k,\ n=\sum_{i=1}^{I}n_i \tag{5.3.6}$$

가 되는데, 이를 SSE와 SST_r로 표현되도록 손질하면 (<비고 5.3.3> 참조),

$$f\equiv\frac{SST_r/(I-1)}{SSE/(n-I)}\ge k'=\frac{n-I}{I-1}(k^{-\frac{2}{n}}-1) \tag{5.3.7}$$

를 얻는다. 그리고, 이에 식 (5.3.4)와 (5.3.5)를 대입하면 식 (5.2.8)이 된다.

식 (5.3.7)에서 SST_r과 SSE를 각각 $(I-1)$과 $(n-I)$로 나눈 이유는 다음과 같다. SST_r과 SSE가 확률변수인 경우에 (즉, y_{ij}를 Y_{ij}로 대체했을 때: <비고 5.3.2> 참조), SST_r/σ^2과 SSE/σ^2은 귀무가설 하에서 카이제곱 분포를 따르는데 자유도는 각각 $(I-1)$과 $(n-I)$ (이고 물론 서로 독립)이다. 따라서,

$$\frac{SST_r/(I-1)}{SSE/(n-I)}\sim F(I-1,n-I) \tag{5.3.8}$$

이다 (식 (2.15.14) 참조). 그리고, $SST_r/\sigma^2\sim\chi^2(I-1)$과 $SSE/\sigma^2\sim\chi^2(n-I)$에 대한 증명 방법은 §2.15.3에서 식 (2.15.11)을 증명한 방법과 동일한데 복잡하므로 생략한다.

<비고 5.3.4> 귀무가설 하에서 $TSS/\sigma^2\sim\chi^2(n-1)$이다. 그러나, 이는 SST_r/σ^2 및 SSE/σ^2과 독립이 아니다.

5.3.2 One-way ANOVA 예제

§5.2에 등장한 예제를 $I=4$ 경우로 다음과 같이 확장한다.

j \ i	1	2	3	4	
1	3	5	2	4	y_{ij}
2	2	2	2	2	
3	1	4	4	5	
4	2	5	4	1	
$\overline{y_i}$	2	4	3	3	$\overline{y}=3$

식 (5.3.1)과 (5.3.3) 그리고 <비고 5.3.3>에 의해서

$$TSS=\sum_{i=1}^{4}\sum_{j=1}^{4}(y_{ij}-\overline{y})^2=30$$

$$SST_r=\sum_{i=1}^{4}4(\overline{y_i}-\overline{y})^2=8$$

$$SSE=TSS-SST_r=22$$

인데, 이를 식 (5.3.7)에 대입하면

$$f=\frac{SST_r/(4-1)}{SSE/(16-4)}=\frac{2.6\dot{6}}{1.8\dot{3}}=1.4\dot{5}$$

를 얻는다. 그런데, 분자와 분모 자유도가 각각 (4-1)과 (16-4)인 F-test에서 $\alpha=0.05$ 에 대한 UTT의 기각역은 $f\geq 3.49$ 이므로 귀무가설을 (기각하지 못하고) 채택한다. (비고: $\alpha=0.10$ 에 대해서도 기각역이 $f\geq 2.61$ 이므로 귀무가설을 채택함.)

5.3.3 ANOVA Table

ANOVA 결과를 일목요연하게 표로 만들면 편리하다. 이때 추가되는 용어는 MS (mean square) 인데, 이는 SS 를 자유도로 나눈 것이다. §5.3.2의 예제에 대한 ANOVA

Table 은 다음과 같다. (비고: 통계 패키지는 p-value 도 제공함. §4.6.3 참조.)

Source (of Variation)	SS	자유도	MS	f
Treatment	8	3 (= I - 1)	2.6̇	1.4̇5̇
Error	22	12 (= n - I)	1.8̇3̇	
Total	30	15 (= n - 1)	(2)	

참고로, §5.2의 예제에 대한 ANOVA Table은 다음과 같다.

Source	SS	자유도	MS	f
Treatment	8	1	8	6
Error	8	6	1.3̇	
Total	16	7	(2.286)	

그런데, 분자와 분모 자유도가 각각 1 과 6 인 F-test에서 $\alpha = 0.05$ 에 대한 UTT의 기각역은 $f \geq 5.99$ 이므로 귀무가설을 (가까스로) 기각한다. (비고: f 값 6 과 경계치 5.99는 §5.2에서 얻은 t 값 -2.449 와 경계치 2.447 을 각각 제곱한 것임.)

§5.4 실험계획

5.4.1 신호와 잡음

3 장에서는 추정을 다루었고 4장 이후에는 검정을 다루고 있는데, 지금까지는 주로 LF를 미분해서 MLE를 얻고 또한 LR을 손질해서 검정통계량을 얻는데에만 급급했다. 그래도, 추정량은 대부분 그 자체로 의미가 있어서 실감이 났다. (예: 모평균에 대한 추정량은 표본평균.) 이제 검정통계량에 대해서도 의미를 부여해 보기로 한다.

신호 (signal)와 잡음 (noise) 의 비율을 SN비라고 한다. 신호가 어느정도 강해도 잡음이 더 강하면 신호를 감지하기 어렵다. 반면에, 잡음이 거의 없으면 약한 신호라도 감지할 수 있다. 검정통계량도 SN비로 해석하면 이해하기 쉽다. 예를 들어, 식 (5.2.9)와 (5.3.8)에서 분자는 신호에 그리고 분모는 잡음에 해당된다. 그리고, §5.3.3의 ANOVA Table에 있는 MS들의 비율인 f값은 바로 관찰된 SN비에 해당된다.

신호가 강하고 잡음이 약한 효과적인 검정법을 찾는 것이 바로 실험계획이다. 이제, §5.3의 One-way ANOVA 보다 더 효과적인 검정법을 찾기 위한 준비작업으로 먼저 지금까지 얻은 결과를 분석한다.

5.4.2 Source of Variation

§5.3.3의 ANOVA Table 에서 Treatment 와 Error 를 Source of Variation 이라 불렀다. 여기에서, Variation 이란 $(y_{ij}-\overline{y})^2$ 를 의미하는데, Variation 의 총량(total)은 TSS 인 $\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2$ 이다.

§5.2에서 예로 들었듯이, 사료 i를 먹인 젖소 n_i마리 중에서 j번째 젖소의 우유생산량을 y_{ij}라 하자. 그러면, 어떤 젖소는 우유생산량이 (평균치 $\overline{y}$보다) 많고 또 어떤 젖소는 우유생산량이 적은데, 이러한 Variation의 총량을 TSS라 하는 셈이다. 그리고, TSS중에서 서로 다른 사료를 먹인데 기인한 부분이 SST_r인 $\sum_i\sum_j(\overline{y_i}-\overline{y})^2$ 이고, 나머지 부분을 SSE라 하는 셈이다. 이때, SST_r을 사료 (또는, Treatment)에 의해서 설명된(explained) Variation 이라 하고, SSE를 설명안된 (unexplained) Variation 이라 한다.

SST_r이 클수록 신호가 강하고 SSE가 작을수록 잡음이 약하다. 그러나, SST_r과 SSE 그 자체가 아니라 각각에 대응된 자유도로 나누어서 표준화시킨 것을 신호와 잡음으

로 사용하는데, 이들이 바로

$$MST_r \equiv SST_r/(I-1) \tag{5.4.1}$$

$$MSE \equiv SSE/(n-I) \tag{5.4.2}$$

이다.

5.4.3 MS의 정체 (이 절은 생략해도 무방함)

ANOVA의 기본 가정인 "분산이 같은 I개의 정규 모분포"에 귀무가설인 " $\mu_1 = \mu_2 = \cdots = \mu_I$" 를 합치면

$$Y_{ij} \sim N(\mu,\sigma^2) \tag{5.4.3}$$

이 된다. (비고: Y_{ij}, $i=1,\cdots,I$, $j=1,\cdots,n_i$, 는 iid 확률변수임.) 즉, Y_{ij} 는 정규분포를 따르는데 평균 μ 와 분산 σ^2 은 사료의 종류인 i와 무관하다. 따라서, I개의 표본을 하나로 합치면 μ 에 대한 MLE로 $\overline{Y}$를 얻는데 이는 또한 MVUE이다 (식 (5.2.3) 참조). 또한, σ^2 에 대한 MLE는 TSS/n이고 (식 (5.2.4) 참조), MVUE는

$$TMS \equiv TSS/(n-1) \tag{5.4.4}$$

이다.

<비고 5.4.1> TMS(total mean square)는 ANOVA Table에 잘 등장하지 않음. 또한, " $TMS \neq MST_r + MSE$" 임 (<비고 5.3.3>참조).

그런데, MST_r과 MSE도 (귀무가설 하에서는) σ^2 에 대한 불편추정량이다. 다만, 이들의 분산은 TMS의 분산보다 크다. (즉, MST_r과 MSE는 $MVUE$가 아니다.) 구체적으로, $E(TMS) = E(MST_r) = E(MSE) = \sigma^2$ 이지만, $V(TMS) = 2\sigma^4/(n-1)$, $V(MST_r) = 2\sigma^4/(I-1)$, $V(MSE) = 2\sigma^4/(n-I)$ 이다 (식 (2.15.12) 참조).

반면에, 귀무가설의 제약이 없으면, μ_i에 대한 MLE 겸 MVUE로 $\overline{Y_i}$를 얻고 (식 (5.2.5) 참조), σ^2 에 대한 MLE로 SSE/n을 얻으며 (식 (5.2.6) 참조), σ^2에 대한 MVUE로는 바로 MSE를 얻는다. (비고: $I=2$ 일때의 MSE는 식 (3.7.3)과 (5.3.4)의 S^2 임.) 그런

데, 귀무가설은 단순가설이므로 이는 전체 모수공간에서 극히 일부분에 지나지 않는다. (비고; 모수공간은 $I+1$ 차원 공간인데, 이 속에서 귀무가설은 2차원을 차지함.) 따라서, 귀무가설의 제약없이 구한 추정량들은 사실상 "대립가설 하에서" 구한 것이라 해도 별로 무리가 없다. 즉, MSE 는 사실상 "대립가설 하에서" σ^2 에 대한 MVUE 인 셈이다.

그렇다면, 귀무가설 하에서 MSE 와 같이 σ^2 에 대한 불편추정량이던 MST_r 은 대립가설 하에서는 무엇이 되는가? 이를 $I=2$ 인 경우에 대해서 간단히 살펴보자 (자세한 내용은 실험계획법 교재 참조). $I=2$ 이면 $I-1=1$ 이므로, 식 (5.3.5)는 SST_r 인 동시에 MST_r 이다. 그런데, 기대치를 구하면 (과정 생략)

$$E(MST_r)=\sigma^2+\frac{(\mu_1-\mu_2)^2}{\frac{1}{n_1}+\frac{1}{n_2}} \tag{5.4.5}$$

이 되므로, MST_r 속에는 σ^2 에 대한 불편추정량 외에 다른 것들이 추가로 포함되어 있음을 알 수 있다. (비고: 귀무가설 하에서는 $\mu_1-\mu_2=0$ 임.)

$I=2$ 인 경우에는 식 (5.4.5)에서 $(\mu_1-\mu_2)^2$ 이 클수록 $E(MST_r)$ 또는 E(신호의 크기) 가 커진다. 그런데, $(\mu_1-\mu_2)^2$ 은 우리가 제어(control)할 수 있는 것이 아니다. 따라서 우리는 E(신호의 크기) 를 증가시키는 대신에 E(잡음의 크기) 를 감소시키려고 노력한다.

§5.5 Two-Way ANOVA

5.5.1 Two-way ANOVA

잡음의 크기를 감소시키기 위한 대표적인 방법은 TWA (Two-Way ANOVA)이다. §5.3.2의 OWA(One-Way ANOVA) 예제에서 Variation의 총량인 TSS 는 30 인데 그 중에서 사료의 차이에 의해서 설명된 부분인 SST_r 은 8 이고 설명안된 나머지인 SSE 는 22 라고 했다. (§5.4.2 참조). TWA란 OWA에서 설명안된 부분 중에서 일부를 추가로 설명하는 것이다.

§5.3.2의 예제에서는 모두 16 마리의 젖소가 동원되었다. 즉 16 마리를 4 마리씩 4 무리로 나누어서 각각 다른 사료를 먹인 것이다.

<비고 5.5.1> OWA를 □Completely Randomized Design□ 이라고도 부르는데, 그 이유는 전체 실험대상을 I 개의 무리로 나누는 방법이 □무작위□이기 때문이다.

예를 들어, 전체 16 마리 중에서 가장 어린 4 마리에게는 사료 1 을 먹이고 가장 잘 자란 4 마리에게는 사료 2 를 먹인다면 이는 공정한 실험이 아니다. 물론, 그 이유는 잘 자란 젖소의 우유생산량이 어린 젖소보다 많을 것이기 때문이다. 따라서, OWA에서는 16 마리를 무작위로 4 마리씩 4 무리로 나눈다(<비고 5.5.1> 참조).

반면에, 젖소의 (나이, 체중, 품종 등의) 차이에 의해서 발생하는 Variation을 아예 제거하는 방법이 바로 TWA 이다. 이 경우 젖소는 4 마리만 필요하지만 시간은 4 배가 소요된다. 예를 들어, 처음 한달간은 4 마리 모두에게 사료 1 을 먹인다음 우유생산량을 측정한다. 그리고, 다음 한달 간은 4 마리 모두에게 사료 2 를 먹인 다음 우유생산량을 측정하는 식으로 실험을 하는 것이다. (비고: 실제로는 사료의 순서를 무작위로 결정함. <비고5.5.3> 참조.)

OWA와 TWA의 차이점을 쉽게 파악하기 위해서 편의상 §5.3.2 의 y_{ij} 를 다음과 같이 순서만 바꾸어서 사용한다.

젖소 j \ 사료 i	1	2	3	4 (= I)	$\overline{y_{.j}} = \sum_{i=1}^{I} y_{ij}/I$
1	3	5	4	4	4
2	2	5	4	5	4
3	1	4	2	1	2
4 (= J)	2	2	2	2	2
$\overline{y_{i.}} = \sum_{j=1}^{J} y_{ij}/J$	2	4	3	3	$\overline{y} = 3$

y_{ij}의 순서만 바꾸었으므로 TSS는 여전히 30이다. 그리고, $\overline{y_{i.}}$는 §5.3.2의 표에서 $\overline{y_i}$와 일치하므로 SST_r 또한 여전히 8이다. 즉, Variation의 총량 30 중에서 사료의 차이에 의해서 설명된 Variation은 여전히 8이다.

이제, 젖소의 차이에 의해서 설명된 Variation을 구한다. 이를 SSB(SS for Blocks)라 부르는데, 구하는 방법은 SST_r과 동일하다(식 (5.3.3) 참조). 즉,

$$SSB = \sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{y_{.j}} - \overline{y})^2 = I\sum_{j=1}^{J}(\overline{y_{.j}} - \overline{y})^2 = 16 \tag{5.5.1}$$

이다. 그리고, SST_r에 대응된 자유도가 ($I-1$) 이듯이 (식 (5.4.1) 참조), SSB에 대응된 자유도는 ($J-1$)이다.

TWA에서는 SSE가 다음과 같다.

$$SSE = TSS - SST_r - SSB \tag{5.5.2}$$

즉, SSE는 전체 Variation 중에서 사료의 차이와 젖소의 차이에 의해서 설명된 부분들을 제외하고 남은 (설명안된) Variation이다. 또한, 식 (5.5.2)는 SS에 대응된 자유도 간에도 성립한다. 따라서, SSE에 대응된 자유도는 다음과 같다.

$$(IJ-1)-(I-1)-(J-1) = (I-1)(J-1) \tag{5.5.3}$$

이상의 결과를 ANOVA Table로 정리하면 다음과 같다.

Source (of Variation)	SS	자유도	MS	f
Treatment	8	3	2.6	4
Block	16	3		
Error	6	9	0.6	
Total	30	15		

§5.3.3 의 ANOVA Table과 비교하면, 관찰된 신호의 크기는 같지만 (MST_r=2.6), 관찰된 잡음의 크기인 MSE는 1.83 에서 0.6 으로 줄어들었다. 따라서 SN비인 f값은 1.45 에서 4로 늘어났다.

분자와 분모의 자유도가 각각 3과 9인 F-test에서 $\alpha=0.05$ 에 대한 UTT 의 기각역은 $f \geq 3.86$ 이므로, 이제는 귀무가설 "$\mu_1=\mu_2=\mu_3=\mu_4$"를 기각할 수 있다. (비고: $\alpha=0.01$ 이면 기각역은 $f \geq 6.99$ 이므로 귀무가설을 채택함.)

5.5.2 TWA에 대한 LRT

OWA에서는 모집단이 I개인 반면에, TWA 에서는 모집단이 IJ개이다. 그러니까, 예제에서는 사료별로 그리고 젖소별로 모집단이 다르다. 그리고, 각 모집단에서 크기가 1인 표본을 하나씩 추출하는 셈이다.

<비고 5.5.2> 이책에서 다루는 TWA는 □TWA with one observation per cell□에 해당된다.

<비고 5.5.3> TWA 를 □Randomized Block Design□ 이라고도 하는데, 이는 예를 들어 사료를 먹이는 순서 (예: 사료 3→1→4→사료2) 를 젖소 별로 무작위로 결정하는 것이다.

Y_{ij} 를 젖소 j에게 사료 i를 먹일 때의 우유생산량이라 하자. TWA 의 기본적인 가정은 OWA와 같다. 즉, Y_{ij} 는 정규분포를 따르고 모분산이 (i와 j에 무관하게) 모두 같다는 것이다. 그러나, 모평균 μ_{ij} 는 i뿐만 아니라 j에 따라서도 다를 수 있다. 편의상,

$$\mu_{ij}=\mu+\tau_i+\beta_j \qquad (5.5.4)$$

라 하자. (단, $\sum_{i=1}^{I} \tau_i = 0$, $\sum_{j=1}^{J} \beta_j = 0$.) 그러면,

$$Y_{ij} \tilde{} N(\mu + \tau_i + \beta_j, \sigma^2) \tag{5.5.5}$$

가 TWA의 가정이다. 이를 식 (5.2.1)과 비교하면, OWA에서는 "$\mu_i = \mu + \tau_i$" 이다. 즉, OWA에서는

$$\beta_1 = \beta_2 = \cdots = \beta_J (=0) \tag{5.5.6}$$

을 가정하는 셈이다. 그리고, 식 (5.1.1)의 가설은 이제 다음과 같이 표현된다.

$$\begin{aligned} H_0 &: \tau_1 = \tau_2 = \cdots = \tau_I (=0) \\ H_a &: Not\ H_0 \end{aligned} \tag{5.5.7}$$

<비고 5.5.4> 이 책에서는 미지의 모수인 τ_i와 β_i가 (확률변수가 아니라) 상수인 경우만 취급한다.

LRT의 결과를 OWA 경우과 비교해서 요약하면 다음과 같다. 첫째, 식 (5.5.7)의 H_0와 H_a에서 공통으로 μ와 β_j에 대한 최우추정량(겸 MVUE)은 각각 $\overline{Y}$와 ($\overline{Y_{\cdot j}} - \overline{Y}$)이다. 둘째로, τ_i에 대한 최우추정량(겸 MVUE)은 H_a하에서 (엄격히 하자면 "$H_o \cup H_a$" 하에서) ($\overline{Y_{i\cdot}} - \overline{Y}$)이다. (비고: OWA의 H_a하에서 $\mu_i \equiv \mu + \tau_i$에 대한 최우추정량(겸 MVUE)는 $\overline{Y} + (\overline{Y_{i\cdot}} - \overline{Y}) = \overline{Y_{i\cdot}}$임.) 셋째로, σ^2에 대한 최우추정량은 H_0와 $H_a(\cap H_0)$ 하에서 각각

$$\widehat{\sigma_0^2} = \frac{SSE + SST_r}{IJ} \left(= \frac{TSS - SSB}{IJ}\right) \tag{5.5.8}$$

$$\widehat{\sigma^2} = \frac{SSE}{IJ} \tag{5.5.9}$$

이다. (비고: OWA에서는 $TSS = SSE + SST_r$.)

LR인 λ의 형태는 식 (5.2.7)과 같은데 (<비고 5.2.2> 참조), 식 (5.5.8)과 (5.5.9)를 대

입해서 손질하면

$$\frac{MST_r}{MSE} \equiv \frac{SST_r/(I-1)}{SSE/(I-1)(J-1)} \sim F(I-1,(I-1)(J-1)) \qquad (5.5.10)$$

을 얻는다 (식 (5.3.8) 참조). 참고로, 식 (5.5.2)에서 $\{SSE,\ SST_r,\ SSB\}$ 는 서로 독립이고, 귀무가설 하에서 $SSE/\sigma^2 \sim \chi^2((I-1)(J-1))$, $SST_r/\sigma^2 \sim \chi^2(I-1)$, $SSB/\sigma^2 \sim \chi^2(J-1)$ 이다. 또한 식 (5.5.8)과 (5.5.9)를 불편추정량이 되도록 손질하면 MVUE로 $(SSE+SST_r)/(IJ-J)$ 와 MSE를 얻는다. 마지막으로, $H_a(\cap H_0)$ 하에서 $E(MST_r)$ 과 $E(MSB)$ 는 다음과 같다 (식 (5.4.5) 참조).

$$E(MST_r) = \sigma^2 + \frac{J}{I-1}\sum_{i=1}^{I}\tau_i^2$$

$$E(MSB) = \sigma^2 + \frac{I}{J-1}\sum_{j=1}^{J}\beta_j^2$$

5.5.3 ANOVA 뒷처리

§5.5.1 의 예제에서는 H_0 인 "$\tau_1=\tau_2=\tau_3=\tau_4(=0)$"을 (유의수준 5%에서) 기각했다. 그러니까 H_a 인 □Not H_0□을 채택한 셈인데 (<비고 4.4.1>참조), 이때 유의할 점은 대립가설이 □모든 τ_i 가 0 이 아님□이 아니라 □ τ_i 중에서 최소한 하나는 0 이 아님□이라는 점이다. 이에 따라, 어느 τ_i 가 0 이 아닌지 알아보는 뒷처리가 필요한데, 이를 Post-ANOVA Analysis 라 한다.

τ_i 에 대한 최우추정량(겸 MVUE)는 $(\overline{Y_{i\cdot}}-\overline{Y})$ 인데, 예제에서는 τ_3 와 τ_4 에 대한 추정치인 $(\overline{y_{3\cdot}}-\overline{y})$와 $(\overline{y_{4\cdot}}-\overline{y})$ 가 모두 0 이므로 τ_3 와 τ_4 에 대한 검정은 해 볼 필요조차 없다. (비고: 관찰된 신호의 크기가 0 임.) 반면에, τ_1 과 τ_2 에 대한 검정은 해볼만한데, 유의할 점은 이 두 가지 검정이 서로 독립이 아니라는 점이다. 이는 "$\sum_{i=1}^{I}\tau_i=0$"이라는 제약 때문인데, 예를 들어 "$H_0:\tau_1=0$"를 기각하면 "$H_0:\tau_2=0$"도 기각하게 된다. 이러한 경우에는 둘을 묶어서 "$H_0:\tau_1=\tau_2$, $H_a:\tau_1\neq\tau_2$"를 검정하면 효과적이다. (비고:검정통계량도 더 간단하거니와 검정력도 더 강하다.)

"$H_0:\tau_1=\tau_2$"에 대한 검정 역시 LRT 이다. 그런데, 이번에는 검정통계량을 구하는

기계적인 과정은 생략하고 대신에 SN비의 관점으로 검정통계량을 해석하기로 한다. 귀무가설은 "$\tau_1 - \tau_2 = 0$"과 마찬가지이고, $(\tau_1 - \tau_2)$에 대한 최우추정량(겸 MVUE)는 $(\overline{Y_{1\cdot}} - \overline{Y_{2\cdot}})$이다. 신호의 역할을 하게될 $(\overline{Y_{1\cdot}} - \overline{Y_{2\cdot}})$의 분포는 다음과 같이 구한다. 식 (5.5.5)에서 Y_{ij}, $i = 1, \cdots, I$, $j = 1, \cdots, J$는 서로 독립이다. (비고: 평균이 다르므로 *iid* 확률변수는 아님.) 따라서, <비고 2.15.1>에 의해서 $\overline{Y_{i\cdot}} \sim N(\mu + \tau_i, \sigma^2/J)$이고, 또한 $(\overline{Y_{1\cdot}} - \overline{Y_{2\cdot}}) \sim N(\tau_1 - \tau_2, 2\sigma^2/J)$이다. (참고로, τ_i에 대한 추정량인 $(\overline{Y_{i\cdot}} - \overline{Y})$의 분포는 $N(\tau_i, \sigma^2(I-1)/IJ)$이다.) 그러니까, 만약 σ^2이 알려졌더라면 귀무가설 하에서 $N(0,1^2)$ 분포를 따르는

$$\frac{\overline{Y_{1\cdot}} - \overline{Y_{2\cdot}}}{\sqrt{2\sigma^2/J}} \tag{5.5.11}$$

를 검정통계량으로 사용했을 것이다. (비고: $\sqrt{2\sigma^2/J}$는 잡음의 역할을 함) 그런데, σ^2을 모르므로 이를 MVUE인 MSE로 대체하면 t분포를 따르는 검정통계량이 된다. 이때 유의할 점은 다음과 같다. 가설 "$\tau_1 = \tau_2$"가 τ_1과 τ_2에 대해서만 언급했다고 해서 $i = 1, 2$에 해당되는 표본만 사용하는 것은 아니다. (비고: MSE는 $i = 1, 2$ 뿐만 아니라 $i = 3, 4$에 해당되는 표본까지 사용해서 얻은 것임.) 따라서, 식 (5.5.11)의 σ^2을 MSE로 대체하면

$$\frac{\overline{Y_{1\cdot}} - \overline{Y_{2\cdot}}}{\sqrt{MSE(2/J)}} \sim t((I-1)(J-1)) \tag{5.5.12}$$

가 되는데, 이는 귀무가설 "$\tau_1 = \tau_2$"하에서 t분포를 따르며 자유도는 MSE의 자유도와 같다.

식 (5.5.12)에 $\overline{y_{1\cdot}} = 2$, $\overline{y_{2\cdot}} = 4$, MSE = 0.6, $I = J = 4$를 대입하면 $t = -3.464$를 얻는데, 자유도가 9인 T-검정에서 $\alpha = 0.05$에 대한 TTT의 기각역은 "$t > 2.262$ 또는 $t < -2.262$" 이므로 귀무가설 "$\tau_1 = \tau_2$"를 기각한다. 또한, $\alpha = 0.01$에 대해서도 $t = -3.464 < -3.250$ 이므로 귀무가설을 기각한다. (참고로, "$H_0 : \tau_1 = 0$, $H_a : \tau_1 \neq 0$"에 대해서는 $t = -2.83$을 얻으므로 $\alpha = 0.05$로는 H_0를 기각하지만 $\alpha = 0.01$로는 H_0를 채택한다.)

아울러, $(\tau_1 - \tau_2)$에 대한 95%와 99% 신뢰구간으로

$$(\overline{y_{1.}} - \overline{y_{2.}}) \pm 2.262\sqrt{MSE(2/J)} = -2 \pm 1.306$$
$$(\overline{y_{1.}} - \overline{y_{2.}}) \pm 3.250\sqrt{MSE(2/J)} = -2 \pm 1.876 \qquad (5.5.13)$$

를 얻는다.

<비고 5.5.5> 식 (5.5.12)와 신뢰구간에서 동일한 MSE를 사용하므로, □귀무가설 □$\tau_1 - \tau_2 = 0$ □을 기각함□과 □($\tau_1 - \tau_2$)에 대한 신뢰구간이 0을 포함하지 않음□은 동치이다.

신뢰구간은 TTT의 대용품이 되기도 하고 (<비고 5.5.5> 참조), 또한 그 자체로도 의미가 있으므로 몇 가지 더 구해보기로 한다. 첫째, $\mu_i = \mu + \tau_i$에 대한 추정량은

$$\overline{Y} + (\overline{Y_{i.}} - \overline{Y}) = \overline{Y_{i.}} \sim N(\mu + \tau_i,\ \sigma^2/J)$$

이므로, 예를 들어 $\mu_1 \equiv \mu + \tau_1$에 대한 95% 신뢰구간은

$$\overline{y_{1.}} \pm 2.262\sqrt{MSE/J} = 2 \pm 0.923 \qquad (5.5.14)$$

이다. 둘째로

$$\mu_1 - \mu_2 \equiv (\mu + \tau_1) - (\mu + \tau_2) = \tau_1 - \tau_2$$

이므로 ($\mu_1 - \mu_2$)에 대한 신뢰구간은 ($\tau_1 - \tau_2$)에 대한 신뢰구간인 식 (5.5.13)과 같다.

참고로, 식 (5.5.13)과 (5.5.14)는 OWA에서도 유효한데, 다만 MSE의 값이 달라질 뿐이다. 물론, OWA에서는 $n_1 \neq n_2$일 수 있으므로 식 (5.5.14)의 J는 n_1으로 고치고, 식 (5.5.13)의 $2/J$는 $(1/n_1 + 1/n_2)$로 고치면 된다.

마지막으로, TWA의 주 목적은 식 (5.5.7)을 검정하는 것이지만, 부산물로 식 (5.5.6)에 대한 검정결과도 얻는다. 식 (5.5.6)에 대한 검정통계량은 다음과 같다 (식 (5.5.10) 참조).

$$\frac{MSB}{MSE} = \frac{SSB/(J-1)}{SSE/(I-1)(J-1)} \sim F(J-1, (I-1)(J-1))$$

이에, $MSB = 16/3 = 5.3$과 $MSE = 6/9 = 0.6$을 대입하면 $f = 8 > 6.99$이므로 $\alpha = 0.01$에서도 식 (5.5.6)을 기각할 수 있다.

5.5.4 T-test for Matched Samples

OWA 에서 $I=2$ 인 경우를 □T-test for Independent Samples□라 불렀듯이 (<비고 5.2.1> 참조), TWA 에서는 $I=2$ 인 경우를 □T-test for Matched Samples□ 또는 □Pairwise T-test□라 부른다.

먼저, §5.5.1의 예제에서 $I=1,2$ 에 대해서만 TWA 를 적용한 결과는 다음과 같다.

Source (of Variation)	SS	자유도	MS	f
Treatment	8	1	8	8
Block	5	3		
Error	3	3	1	
Total	16	7		

유의수준을 5%로 잡으면, 분자와 분모 자유도가 각각 1과 3인 F-test에서 UTT의 기각역은 $f > 10.13$ 이므로 귀무가설 "$\tau_1 = \tau_2$"를 기각하지 못한다.

이와 동일한 결과를 T-test로도 얻을 수 있는데, 흥미로운 점은 (실제로는) 모집단이 8개인 문제를 마치 모집단이 하나인 문제처럼 처리할 수 있다는 것이다 (§4.4.2 참조). 먼저, Y_{1j} 와 Y_{2j} 를 짝(match)을 지어서 그 차이(difference)를 D_j 라 하자. 예를 들어, $D_j = Y_{1j} - Y_{2j}$ 는 젖소 j 의 우유생산량의 차이인데, 이 차이가 두 종류의 사료의 차이에 기인한다는 것이 대립가설의 입장이다. 다음, $\{D_1,\cdots,D_J\}$ 를 크기가 J 인 표본으로 간주하는데, 모집단의 분포는 $N(\delta,\sigma_D^2)$ 이라 가정한다. 그리고는

$$H_0 : \delta = 0,\ H_a : \delta \neq 0 \tag{5.5.15}$$

를 검정하는 것이다. 따라서, 기각역은 식 (4.4.10)의 형태가 된다.

예제에서 관찰된 표본 $\{d_j = y_{1j} - y_{2j}\ ;\ j=1,2,3,4\}$ 는 $\{-2,-3,-3,0\}$ 이므로, 이로부터

$$\overline{d} = \sum_{j=1}^{4} d_j/4 = -2$$

$$\sigma_D^2 = \sum_{j=1}^{4} (d_j - \overline{d})^2/(4-1) = 2$$

를 얻은 다음, t 값을 구하면

$$t \equiv \frac{\overline{d}-0}{\widehat{\sigma}_D/\sqrt{J}} = \frac{-2}{\sqrt{2}/\sqrt{4}} = -\sqrt{8}(\approx 2.828)$$

을 얻는다. 유의수준을 5%로 잡으면, 자유도가 $J-1=3$ 인 T-test 에서 TTT의 기각역은 "$t>3.182$ 또는 $t<-3.182$" 이므로 귀무가설 "$\delta = 0$" 를 기각하지 못한다. (비고: $f=8=t^2$, $10.13=(3.182)^2$.)

§5.6 선형모형

이제 ANOVA를 마무리하는 동시에 6장의 회귀분석을 대비할 때가 되었다.

TWA의 기본가정은 다음과 같다. 모집단이 IJ개 있는데, 모분포는

$$Y_{ij} \sim N(\mu + \tau_i + \beta_j, \sigma^2)$$

이다 (식 (5.5.5) 참조). 단, τ_i와 β_j는 확률변수가 아니며 (<비고 5.5.4> 참조), $\sum_{i=1}^{I} \tau_i = 0$ 과 $\sum_{j=1}^{J} \beta_j = 0$ 를 만족시킨다.

TWA의 기본가정을 선형모형(linear model)으로 표현하면 다음과 같다.

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}\ ,\ \varepsilon_{ij} \sim iid\ N(0,\sigma^2) \tag{5.6.1}$$

즉, 확률변수 Y_{ij}는 (미지의) 상수 μ, τ_i, β_j에 iid 확률변수인 ε_{ij}를 합친 것인데, 정규분포를 따르는 ε_{ij}의 평균은 0이고 분산은 (i, j와 무관하게) σ^2이다 . 한편, 귀무가설 "$\tau_1 = \cdots = \tau_I$" 하에서의 선형모형은 다음과 같다.

$$Y_{ij} = \mu + \beta_j + \varepsilon_{ij}\ ,\ \varepsilon_{ij} \sim iid\ N(0,\sigma^2) \tag{5.6.2}$$

<비고 5.6.1> 식 (5.6.1)을 CM (complete model)이라 하고 식 (5.6.2)를 RM (reduced model)이라 한다.

반면에, OWA에 대한 CM과 RM은 다음과 같다(식 (5.2.1) 참조).

$$CM\ :\ Y_i = \mu + \tau_i + \varepsilon_i$$
$$RM\ :\ Y_i = \mu + \varepsilon_i,\ \varepsilon_i \sim iid\ N(0,\sigma^2)$$

사실 TTT인 T-test들도 모두 선형모형의 틀에 속한다. 예를 들어, §4.4.2에서 다룬 "$H_0 : \mu = \mu_0$, $H_a : \mu \neq \mu_0$" 경우는

$$\begin{aligned} &CM:\ Y = \mu + \varepsilon \\ &RM:\ Y = \mu_0 + \varepsilon,\ \varepsilon \sim N(0,\sigma^2) \end{aligned} \tag{5.6.3}$$

에 해당된다. (비고: μ 는 미지의 상수이고 μ_0 는 구체적인 수치임.) 또한, §5.5.4의 예제는 TWA에 해당되기도 하지만

$$CM: D=\delta+\varepsilon$$

$$RM: D=0+\varepsilon,\ \varepsilon \sim iid\ N(0,\sigma_D^2)$$

에 해당되기도 한다(식 (5.5.15) 참조).

6장에서 다룰 회귀모형 (regression model) 역시 선형모형이다. 그런데, 가장 큰 차이점은 다음과 같다. OWA에서는 Y_i 의 i 가 그리고 TWA 에서는 Y_{ij} 의 i 와 j 가 자연수인 반면에, 회귀모형에서는 Y_x 의 x 가 연속적인 값을 가질 수 있다. 구체적으로, 단순(simple) 회귀모형에서는

$$Y_x=\beta_0+\beta_1 x+\varepsilon_x,\ \varepsilon_x \sim iid\ N(0,\sigma^2) \tag{5.6.4}$$

이고, 다중(multiple) 회귀모형에서는

$$Y_x=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\varepsilon_x,\ \varepsilon_x \sim iid\ N(0,\sigma^2) \tag{5.6.5}$$

이다. (단, 식 (5.6.5)에서는 x 가 벡터 $(x_1, x_2, \cdots)$ 를 의미함.)

예를 들어, 식 (5.6.4)에서 x 는 학부과정의 성적 (또는 평점)이고, Y_x 는 학부성적이 x 인 학생이 대학원에 진학할 경우에 대학원에서 취득할 성적을 의미한다. 또한, 식 (5.6.5)에서도 Y_x 가 대학원에서 취득할 성적이라 하고 x_1 을 학부성적이라 하면, $x_2, x_3, \cdots$ 는 대학원 성적에 영향을 미칠 수 있는 다른 변수들을 의미한다. 예를 들면

$$x_2=\begin{cases} 1\ , \text{ if 모교출신} \\ 0\ , \text{ if 타교출신} \end{cases} \tag{5.6.6}$$

인데, 이 경우 x_2 를 dummy variable 이라 부른다.

<비고 5.6.2> 단순회귀모형은 다중회귀모형의 RM이다.

<비고 5.6.3> ANOVA 는 dummy variable 만 있는 다중회귀모형으로 표현할 수 있다. 예를 들어, OWA 의 CM 은 " $Y_x=\mu+\tau_1 x_1+\tau_2 x_2+\cdots+\tau_I x_I+\varepsilon_x$ " 에서

$$x_i = \begin{cases} 1, & \text{if 사료 } i\text{를 먹인 젖소} \\ 0, & otherwise \end{cases}$$

인 경우에 해당된다. (단, $\sum_{i=1}^{I} \tau_i = 0$ 이므로 τ_I대신에 $\square - \sum_{i=1}^{I-1} \tau_i \square$를 대입함.)

<비고 5.6.4> 이 책의 예제(§5.3.2,§5.5.1 참조)에서와 같이 동일한 표본을 사용하여 OWA와 TWA를 하는 경우에는 OWA의 CM이 TWA의 RM이 된다.