Last course on 17 H.

"report" until end of July

VIII Analysis of variance and regression

VIII.1 "One way" ANOVA

Idea: Consider data like

	Treatments → 1	2	⋯	k
Observation ↓	y_{11}	y_{21}		y_{k1}
	y_{12}	y_{22}		y_{k2}
	⋮	⋮		⋮
	y_{1n_1}	y_{2n_2}		y_{kn_k}

with $\{n_i\}$ independent.

We shall assume that the corresponding r.v. follow (θ_i: unknown, ε_{ij}: noise or error)

$$Y_{ij} = \theta_i + \varepsilon_{ij}$$
$$i = 1, \cdots, k$$
$$j = 1, \cdots, n_i$$

Remark: we could consider

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

but usually impose that $\sum_{i=1}^{k} \tau_i = 0$ (additional parameter) which fixes one parameter.

Def. The model $Y_{ij} = \theta_i + \varepsilon_{ij}$ is called a oneway ANOVA if

1) ε_{ij} is a r.v. following $n(0, \sigma^2)$ (σ is independent of i, j) for any i, j.

2) ε_{ij} and $\varepsilon_{i'j'}$ are independent for any $(i,j) \neq (i',j')$.

Remark: These assumptions can also be weakened if necessary.

Def. ANOVA null assumption is $H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$. (very strong assumption)

$\Rightarrow H_1 : \theta_j \neq \theta_k$ for ≥ 1 pair (θ_j, θ_k) with $j \neq k$

Let us set

$$\mathcal{A} = \mathcal{A}_k = \{\underline{a} = (a_1, a_2, \cdots, a_k) \in \mathbb{R}^k \text{ with } \sum_{j=1}^{k} a_j = 0\}$$

Examples: $a = (1, -1, 0, \cdots, 0)$ or $a = (1, -\frac{1}{2}, -\frac{1}{2}, 0, \cdots, 0)$, etc

Def. Consider $\underline{t} := (t_1, \cdots, t_k)$ a set of k parameters or of k r.v.

If $\underline{a} \in \mathcal{A}_k$, then $\sum_{j=1}^{k} a_j t_j \equiv \underline{a} \cdot \underline{t}$ is called a contrast.

Lemma: $H_0 \Leftrightarrow \forall \underline{a} \in \mathcal{A} : \underline{a} \cdot \underline{\theta} = 0$ ($\Rightarrow$ easy, $\Leftarrow$ as an exercise)

Corollory: $H_1 \Leftrightarrow \exists\, \underline{a} \in \mathcal{A} : \underline{a} \cdot \underline{\theta} \neq 0$

Under the ANOVA assumption one has

$$Y_{ij} \sim n(\theta_i, \sigma^2) \quad \forall i = 1, \cdots, k, \ \forall j = 1, \cdots, n_i$$

$$\Rightarrow \overline{Y_i} := \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} \sim n\left(\theta_i, \frac{\sigma^2}{n_i}\right)$$

Then for any $a \in \mathbb{R}^k$ consider $\sum_{i=1}^{k} a_i \overline{Y_i} =: \underline{a} \cdot \overline{\underline{Y}}$ one has

$$\underline{a} \cdot \overline{\underline{Y}} \sim n\left(\sum_{i=1}^{n} a_i \theta_i, \sigma^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}\right) \quad \text{(Ex. 11.8)}$$

If σ^2 is not known, then we can define the sample variance

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2 \sim \chi^2_{n_i - 1} \quad (\S \text{II}.2)$$

Since σ^2 is the same for all experiments, one can set

$$s^2 = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1)\, s_i^2 = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2 \text{ and}$$

$(N-k)\, s^2/\sigma^2 \sim \chi^2_{N-k}$ (sum of $\chi^2_{n_i-1}$) (Lemma 5.3.2) with

$$N := \sum_{i=1}^{k} n_i$$

Then

$$\frac{\sum_{i=1}^{k} a_i \overline{Y}_i - \sum_{i=1}^{k} a_i \theta_i}{\sqrt{s^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}} \sim t_{N-k}$$

Now for any $\underline{a} \in \mathbb{R}^k$ fixed, a hypothesis test could be

Reject $H_0 : \sum_{i=1}^{k} a_i \theta_i = 0$ if $\left| \frac{\underline{a} \cdot \underline{\overline{Y}} - \underline{a} \cdot \underline{\theta}}{\sqrt{s^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}} \right| > t_{N-k, \frac{\alpha}{2}}$ for a given α.

Equivalentally,

a confident interval with confidence coefficient $1-\alpha$ is given by

$$\sum_{i=1}^{k} a_i \overline{y_i} - t_{N-k, \frac{\alpha}{2}} \sqrt{s^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}} \leq \sum_{i=1}^{k} a_i \theta_i \leq \sum_{i=1}^{k} a_i \overline{y_i} + t_{N-k, \frac{\alpha}{2}} \sqrt{s^2 \sum_{i=1}^{k} \frac{a_i^2}{n_i}}$$

Key fact: If we choose $\underline{a} = (1, -1, 0, \cdots, 0)$ then we are testing $H_0 : \theta_1 - \theta_2 = 0$ or for $\underline{a} = (1, -\frac{1}{2}, -\frac{1}{2}, 0, \cdots, 0)$ then we are testing $H_0 : \theta_1 = \frac{1}{2}(\theta_2 + \theta_3)$.

Observe that all data are used for the computation of s^2.

⚠ If we choose 2 different $\underline{a}, \underline{a}' \in \mathbb{R}^k$, the 2 computations for $H_0 : \underline{a} \cdot \underline{\theta} = 0$ and $H_0' : \underline{a}' \cdot \underline{\theta} = 0$ are not independent, which implies that the 2 confidence coefficients are not both $1-\alpha$.

Remark: The test for the ANOVA null assumption $H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$ can be obtained by the hypothesis union-intersection

$$H_0 \Leftrightarrow \underline{\theta} \in \bigcap_{\underline{a} \in A} \{\underline{\xi} \in \mathbb{R}^k \mid \underline{a} \cdot \underline{\xi} = 0\}$$

To obtain a criterion for this intersection corresponds to a maximization problem. See 11.2.4

An α-level test for rejecting H_0 is given by

$$\frac{\sum_{i=1}^{k} n_i (\overline{Y}_i - \overline{\overline{Y}})^2}{s^2} > (k-1)\, F_{k-1, n-k; \alpha} \quad \text{with } \overline{\overline{Y}} := \frac{1}{N} \sum_{i,j} Y_{ij}$$

↑ F distribution

from which one can obtain a $(1-\alpha)$ confidence interval. Called ANOVA F-test

VIII.2 Simple Linear Regression

Idea: Consider relation of the form: $f(x_i)$

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

Y_i: response; x_i: predictor; ε_i: noise or error; α, β: 2 unknown coeff.

Remark: It is called **linear regression** because it is linear in the coefficients α, β (and not because of x_i)

Remark: Once some data are collected, just evaluating α and β corresponds to "data fitting" but there is no statistical inference.

i	x	Y
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
n	x_n	y_n

For any random variables $\{(X_i, Y_i)\}_{i=1}^n$, let us set

$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \qquad S_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i \qquad S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2 \qquad S_{xY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

About data fitting

Lemma: (least squares) One has

$$\min_{\alpha,\beta} \sum_{i=1}^n \left(y_i - (\alpha + \beta x_i)\right)^2 = \sum_{i=1}^n \left(y_i - (a + b x_i)\right)^2 \text{ for}$$

$b = S_{xy}/S_{xx}$ and $a = \bar{y} - b\bar{x}$

Remark: The previous computation is quite natural if we think that y_i is a function of x_i with relation $y = \alpha + \beta x$, but if we just collect (x_i, y_i), it is not clear that this is the best choice.

In order to do some statistical inference let us assume that

$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

with $\{\varepsilon_i\}$ indep. r.v., with $E(\varepsilon_i) = 0$ and $\text{var}(\varepsilon_i) = \sigma^2$ (independent of i)

Then $E(Y_i) = \alpha + \beta x_i$ and $\text{var}(Y_i) = \sigma^2$.