One-way analysis of variance

Linear model

$$\underline{Y} = \mathbf{X}\underline{\theta} + \underline{\epsilon}$$

Now the design matrix $\mathbf{X}$ does not necessarily have to come from a regression model, but could be some other matrix

$\underline{\epsilon}$ is assumed to consist of i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables (errors)

Seen: maximum likelihood or least squares give the normal equations

$$\mathbf{X}^T\mathbf{X}\underline{\theta} = \mathbf{X}^T\underline{Y}$$

A one-way classification model

Goal: Compare population means when there are p groups, with n_i observations in group i, for $i = 1, \ldots, p$

Example: Four different fertilizers, applied to different plots of land; record the yields: is there a significant difference between the fertilizers?

Fertilizer here would be called a *factor*; one could think of more than one factor (soil type and fertilizer, e.g.)

Model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad j = 1, \ldots, n_i; i = 1, \ldots, p$$

where $\sum_{i=1}^{p} n_i \alpha_i = 0$

so μ is the typical value; we would like to test for $\alpha_i = 0, i = 1, \ldots, p$

If all n_i have the same size n, we call this a *balanced design*

Write as model

$$\underline{Y} = \mathbf{X}\underline{\theta} + \underline{\epsilon}$$

with $(N = \sum_{i=1}^{p} n_i$ times 1) - vector

$$\underline{Y} = (Y_{1,1}, Y_{1,2}, \ldots, Y_{1,n_1}, \ldots, Y_{p,1}, Y_{p,2}, \ldots, Y_{p,n_p})^T$$

and

$$\underline{\theta} = (\mu, \alpha_1, \ldots, \alpha_p)^T \quad ((p+1) \times 1) - vector$$

$$\underline{\epsilon} = (\epsilon_{1,1}, \ldots, \epsilon_{1,n_1}, \ldots, \epsilon_{p,1}, \ldots, \epsilon_{p,n_p})^T$$

which is a $(N \times 1)$-vector,

and $N \times (p+1)$-design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Estimation

We use the notation

$$\bar{Y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{i,j}$$
$$\bar{Y}_{\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{i,j}$$

So $\bar{Y}_{i\cdot}$ is the sample mean in group i, and $\bar{Y}_{\cdot\cdot}$ is the overall mean

Least squares: Calculate

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} N & n_1 & n_2 & n_3 & \cdots & n_p \\ n_1 & n_1 & 0 & 0 & \cdots & 0 \\ n_2 & 0 & n_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ n_p & 0 & 0 & 0 & \cdots & n_p \end{pmatrix}$$

which is *not* invertible

$$\begin{aligned} &\mathbf{X}^T\mathbf{X}\underline{\theta} \\ &= (N\mu + \sum_{i=1}^{p} n_i\alpha_i, n_1\mu + n_1\alpha_1, \ldots, n_p\mu + n_p\alpha_p)^T \end{aligned}$$

and

$$\begin{aligned} \mathbf{X}^T\underline{Y} &= (\sum_{i=1}^{p}\sum_{j=1}^{n_i} Y_{i,j}, \sum_{j=1}^{n_1} Y_{1,j}, \cdots, \sum_{j=1}^{n_p} Y_{p,j})^T \\ &= (N\bar{Y}_{\cdot\cdot}, n_1\bar{Y}_{1\cdot}, \cdots, n_p\bar{Y}_{p\cdot})^T \end{aligned}$$

Normal equations

$$\mathbf{X}^T\mathbf{X}\underline{\theta} = \mathbf{X}^T\underline{Y}$$

give

$$\begin{aligned}
N\mu + \sum_{i=1}^{p} n_i\alpha_i &= N\bar{Y}_{..} \\
n_1\mu + n_1\alpha_1 &= n_1\bar{Y}_{1.} \\
\vdots \quad & \quad \vdots \\
n_p\mu + n_p\alpha_p &= n_p\bar{Y}_{p.}
\end{aligned}$$

use that $\sum_i n_i\alpha_i = 0$: get

$$\begin{aligned}
\hat{\mu} &= \bar{Y}_{..} \\
\hat{\alpha}_i &= \bar{Y}_{i.} - \bar{Y}_{..}, \quad i = 1, \ldots, p
\end{aligned}$$

Intuitively: If the factor levels differ in their mean parameters, then the $\bar{Y}_{i.}$'s should differ significantly from one another; equivalently, $\sum_{i=1}^{p} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$ should be large

Sum of Squares Decomposition

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{i,j}-\bar{Y}_{\cdot\cdot})^2$$
$$= \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{i,j}-\bar{Y}_{i\cdot})^2 + \sum_{i=1}^{p} n_i(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2$$

(*Exercise*) or

$$SST = SSE + SSF$$

with the *total sum of squares*

$$SST = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{i,j}-\bar{Y}_{\cdot\cdot})^2$$

the *error sum of squares*

$$SSE = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{i,j}-\bar{Y}_{i\cdot})^2$$

the *factor sum of squares*

$$SSF = \sum_{i=1}^{p} n_i(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2$$

Chapter 2 Theorem 7 gives:

Proposition 1 *Under* $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_p = 0$, *we have that*

$$\frac{SSF}{\sigma^2} \sim \chi^2_{p-1}, \qquad \frac{SSE}{\sigma^2} \sim \chi^2_{N-p},$$

and these variables are independent of each other

Put

$$MSE = \frac{1}{N-p} SSE$$

the *mean square error*, then MSE is unbiased for σ^2

Put

$$MSF = \frac{1}{p-1} SSF$$

then under H_0 we have from Chapter 2 that

$$F = \frac{MSF}{MSE} \sim F_{p-1,N-p}$$

The analysis is ususally summarized in an **ANOVA Table**:

Source	SS	d.f.	MS	F
Factor	SSF	p-1	$\frac{1}{p-1}SSF$	$\frac{MSF}{MSE}$
Error	SSE	N-p	$\frac{1}{N-p}SSE$	
Total	SST	N-1		

Example: Yield (transformed) by 4 fertilizers

Fertilizer	Yields	n_i	$\bar{Y}_{i\cdot}$
1	8, 5, -1, 6, 5, 3	6	4.33
2	7, 12, 5, 3, 10	5	7.4
3	4, -2, 1	3	1
4	1, 6, 10, 7	4	6
Total	90	18	

so that

Fertilizer	$\sum_{ij}(Y_{ij} - \bar{Y}_{i\cdot})^2$	$n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$
1	47.33	2.67
2	53.2	28.8
3	18	48
4	42	4
Total	160.53	83.47

ANOVA-Table

Source	SS	d.f.	MS	F
Fertilizer	83.47	3	27.82	2.43
Error	160.53	14	11.47	
Total	244	17		

Compare 2.43 to $F_{3,14}$: not significant at 10 % level

(P-value is about 0.11)

Example: multiple regression

Multiple regression can be put in this framework; (but with a different design matrix)

have sum of squares decomposition

Test H_0 : all $\theta_i = 0$ for $i = 1, \ldots, p-1$

with $p-1$ explanatory variables we have the ANOVA-table

Source	SS	d.f.	MS	F
Regression	SSReg	p-1	$\frac{1}{p-1} SSReg$	$\frac{MSReg}{MSE}$
Error	SSE	N-p	$\frac{1}{N-p} SSE$	
Total	SST	N-1		

Here,

$$SSReg = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

and

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

and

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Put

$$R^2 = \frac{SSReg}{SST}$$

then R^2 measures the percentage of the variation explained by the regression

Sometimes one also used the

$$adjusted R^2 = R^2 - \frac{p-1}{n-p}(1 - R^2)$$

Exercise: In the case of simple linear regression, this reduces to the same R^2 as before.

Example: Polymer data, see lectures