Cell means model is a linear model

Illustrate why, using a case involving $r = 3$ treatments and two replicates per treatment.

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}.$$

First column of X is the Group 1 dummy variable; second column is the Group 2 dummy variable; third column is the Group 3 dummy variable.

The design matrix may look a little unusual to you, because there is no column consisting of all 1's, for the intercept term, as we usually have in regression analysis.

But, here, that is a feature, not a mistake. We need to do that, to make sure the design matrix is of "full rank."

Students should check for themselves, by matrix multiplication, that this matrix equation gives the correct model equations for all six observations.

Notation for cell and grand means

Let $Y_{i.} = \sum_{j=1}^{n_i} Y_{ij}$, be the sum of the observations in Group i, for $i = 1, \ldots, r$.

Then the mean of the i^{th} group is:

$$\overline{Y}_{i.} = \frac{1}{n_i} Y_{i.}$$

The group means are also called the *cell means*.

Let $Y_{..} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij}$ be the total of all the observations.

Then the *grand mean* is:

$$\overline{Y}_{..} = \frac{1}{n_T} Y_{..}$$

Exercise
Is $\overline{Y}_{..} = \frac{1}{r}\sum_{i=1}^{r}\overline{Y}_{i.}$?

Answer No, not in general. (Consider a case with unequal n_i's.)
What is true is that

$$\overline{Y}_{..} = \frac{\sum_{i=1}^{r} n_i \overline{Y}_{i.}}{n_T}$$

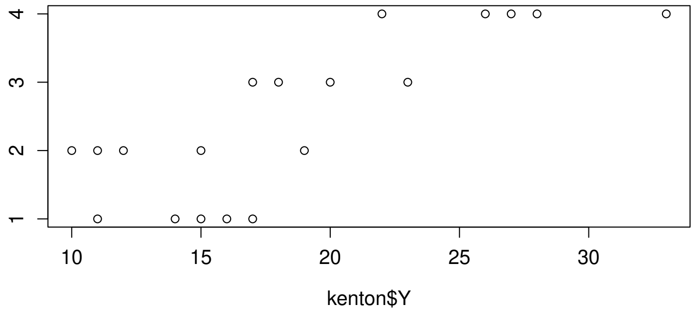
Now we will consider the very important topic of *least-squares* estimation. This is a general method of estimation.

Our model: $E(Y_{ij}) = \mu_i, \; i = 1, \ldots, r, \; j = 1, \ldots, n_i$.

GOAL: Estimate the μ_i's.

Fitting the one-way ANOVA model: Least Squares

Comparative dot plots for Kenton Example



Method of least squares in two steps:

- For each possible set of parameter values $(\mu_1, \mu_2, \mu_3, \mu_4)$, calculate the sum of the squared distances (SS) between the y-values and their means
- Find the values of $(\mu_1, \mu_2, \mu_3, \mu_4)$ that minimize SS; these are the "least squares" estimates. (if the minimization can be done)

Least Squares Estimators

The least-squares estimates of the $\mu_i, i = 1, \ldots, r$ are the values $\hat{\mu}_i$ which minimize the sum of squared deviations of Y's from their expected values. Call the quantity to be minimized Q; then Q is given by

$$Q = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2$$

KEY SIMPLIFICATION FOR ONE-WAY ANOVA:

To minimize Q, you can minimize each of the r terms above separately, since each term involves exactly one parameter. The answer for term i is $\hat{\mu}_i = \overline{Y}_{i.}$ The least-squares estimate is denoted mu_i-hat

A simple explanation, not requiring calculus, can be given in this case. Drop the index i for simplicity.

We want to minimize $\sum_{j=1}^{n} \left(Y_j - \mu\right)^2$ w.r.t. μ.

We want to minimize $Q = \sum_{j=1}^{n} (Y_j - \mu)^2$ w.r.t. μ.

use a "device" or "trick":

First, add and subtract $\overline{Y}$ to each term in the sum; then expand the quadratic. We have

$$Q = \sum_{j=1}^{n} \big(\overbrace{Y_j - \overline{Y}}^{a} + \overbrace{\overline{Y} - \mu}^{b}\big)^2 = \sum_{j=1}^{n} (Y_j - \overline{Y})^2 + \boxed{\sum_{j=1}^{n} 2(Y_j - \overline{Y})(\overline{Y} - \mu)} + \sum_{j=1}^{n} (\overline{Y} - \mu)^2$$

The second term on the right is 0.

$$" 2(\overline{Y}-\mu) \sum_{j=1}^{n} (Y_j - \overline{Y}) =$$

$$2(\overline{Y}-\mu)\left(\sum_{j=1}^{n} Y_j - \sum_{j=1}^{n} \overline{Y}\right) = 2(\overline{Y}-\mu)(n\overline{Y} - n\overline{Y})$$

$$= 2(\overline{Y}-\mu) \times 0$$

$$= 0$$

So we have

$$Q = \sum_{j=1}^{n} (Y_j - \overline{Y})^2 + \sum_{j=1}^{n} (\overline{Y} - \mu)^2$$

The first term is constant wrt μ, and the second term is obviously minimized when we take $\mu = \overline{Y}$. Therefore, $\hat{\mu} = \overline{Y}$.

QED. So: $\hat{\underset{\sim}{\mu}} = (\overline{Y}_{1.}, \overline{Y}_{2.}, \overline{Y}_{3.} \dots, \overline{Y}_{r.})$

Fitted values and residuals

Recall the model equation for the cell-means model:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \text{ i} = 1, \ldots, r; \text{ j} = 1, \ldots, n_i$$

$E(Y_{ij}) = \mu_{ij}$

$\hat{Y}_{ij} = \hat{\mu}_{ij}$

$= \bar{Y}_{i.}$

The fitted values $\hat{Y}_{ij}$ are defined to be:

$$\hat{Y}_{ij} = \overline{Y}_{i.}, \text{ i} = 1, \ldots, r; \text{ j} = 1, \ldots, n_i$$

The residuals:

$e_{ij} = Y_{ij} - \overline{Y}_{i.}$

$e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \bar{Y}_{i.}$

Fitted equation:

$$Y_{ij} = \overline{Y}_{i.} + e_{ij}, \text{ i} = 1, \ldots, r; \text{ j} = 1, \ldots, n_i$$

Special for one-way ANOVA

Property of residuals in one-way ANOVA is that within every group, $i = 1, \ldots r$, the residuals sum to zero:

$$\sum_{j=1}^{n_i} e_{ij} = 0$$

Ex. Kenton Food Company

	Store (j)					Number of			
Package Design	1	2	3	4	5	Stores	Total	Mean	SD
i	Y_{i1}	Y_{i2}	Y_{i3}	Y_{i4}	Y_{i5}	n_i	$Y_{i.}$	$\bar{Y}_{i.}$	s_i
1	11	17	16	14	15	5	73	14.6	
2	12	10	15	19	11	5	67		
3	23	20	18	17		4	78	19.5	
4	27	33	22	26	28	5	136		

$Y_{..} = 354$, $\bar{Y}_{..} = 18.63$, $n_T = 19$

$$s_i = \sqrt{\frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i.}\right)^2}$$

Residuals

$e_{11} = Y_{11} - \hat{Y}_{11} = Y_{11} - \bar{Y}_{1.} = 11 - 14.6 = -3.6$

$e_{34} = Y_{34} - \hat{Y}_{34} = Y_{34} - \bar{Y}_{3.} = 17 - 19.5 = -2.5$

			Store (j)			Number of			
Package	1	2	3	4	5	Stores	Total	Mean	SD
Design									
i	Y_{i1}	Y_{i2}	Y_{i3}	Y_{i4}	Y_{i5}	n_i	$Y_{i.}$	$\bar{Y}_{i.}$	s_i
1	11	17	16	14	15	5	73	14.6	2.302
2	12	10	15	19	11	5	67	13.4	3.647
3	23	20	18	17		4	78	19.5	2.646
4	27	33	22	26	28	5	136	27.2	3.962

Residuals:

Package Design			Store (j)			
i	1	2	3	4	5	Total
1	-3.6	2.4	1.4	$-.6$	.4	0
2	-1.4	-3.4	1.6	5.6	-2.4	0
3	3.5	.5	-1.5	-2.5		0
4	$-.2$	5.8	-5.2	-1.2	.8	0