ESTIMATING PARAMETERS AND VARIANCE FOR ONE-WAY ANOVA

Least Squares Estimates

Typical purpose of experiment: Use the data to estimate or compare certain of the model parameters (or certain functions of the parameters).

Data: The values y_{it} for the random variables Y_{it} obtained in our experiment:

From treatment 1: $y_{11}, y_{12}, \ldots, y_{1n_1}$;
From treatment 2: $y_{21}, y_{22}, \ldots, y_{2n_2}$;
etc.

Method of least squares.

Example: means model: $Y_{it} = \mu_i + \varepsilon_{it}$.

We seek an estimate $\hat{\mu}_i$ for μ_i.

Idea: find $\hat{\mu}_i$'s with the property that when we apply the estimated model to the data, the errors are as small as possible.

In other words: Make the "estimated error terms' (*residuals*)

$e_{it} = y_{it} - \hat{\mu}_i$

as small as possible -- *collectively*.

Picture:

How to do this?

Minimize the sum of the errors?

Positives and negatives will cancel out

Minimize the sum of the absolute values of the errors?

Technical problems.

Minimize the sum of the squared errors?

Works reasonably well!

Details:

Least squares for the effects model:

(Some details might be homework.)

We get the following v+1 equations ("normal equations") in the estimates $\hat{\mu}$ and $\hat{\tau}_i$ for μ and the τ_i's, respectively:

$$y_{..} - n\hat{\mu} - \sum_{i=1}^{v} r_i\hat{\tau}_i = 0$$

$$y_{i.} - r_i\hat{\mu} - r_i\hat{\tau}_i = 0,\ i = 1, 2, \ldots, v.$$

Note: Adding the last v equations gives the first one.

Thus: only v independent equations in the v + 1 unknowns $\hat{\mu}$, $\hat{\tau}_1, \ldots, \hat{\tau}_v$ -- hence infinitely many solutions.

Solution to dilemma: Add the constraint

$$\sum_{i=1}^{v} \hat{\tau}_i = 0.$$

Result: v + 1 independent equations in the v + 1 unknowns, so a unique solution.

Comments:

1. (For students with regression background) This is related to using only v - 1 indicator variables for a categorical variable with v categories in regression.

2. Solve for $\hat{\mu} + \hat{\tau}_i$:

Solve for $\hat{\tau}_i - \hat{\tau}_j$:

Estimable functions: The functions of the parameters $\mu, \tau_1, \ldots, \tau_v$ that *do* have unique least squares estimates. Examples: $\mu + \tau_i$ and $\tau_i - \tau_j$.
(See Sections 3.4.1 and 3.4.4.)

3. Functions of the parameters that have the form $\sum_{i=1}^{v} c_i \tau_i$ where $\sum_{i=1}^{v} c_i = 0$ are called *contrasts*. For example, each difference of effects $\tau_i - \tau_j$ is a contrast. Other contrasts, such as differences of averages, may be of interest as well in certain experiments.

Example: An experimenter is trying to determine which type of non-rechargeable battery is most economical. He tests five types and measures the lifetime per unit cost for a sample of each. He also is interested in whether alkaline or heavy duty batteries are most economical as a group. He has selected two types of heavy duty (groups 1 and 2) and three types of alkaline batteries (groups 3, 4, and 5). So to study his second question, he tests the difference in averages, $(\tau_1 + \tau_2)/2 - (\tau_3 + \tau_4 + \tau_5)/3$. Note that this *is* a contrast, since the coefficient sum is

$$1/2 + 1/2 - 1/3 - 1/3 - 1/3 = 0.$$

(Similarly, we can show that every difference of averages is a contrast.)

Exercise (Possible homework): Every contrast $\sum_{i=1}^{v} c_i \tau_i$ is a linear combination of the effect differences $\tau_i - \tau_j$ and is estimable, with least squares estimate $\sum_{i=1}^{v} c_i \hat{\tau}_i = \sum_{i=1}^{v} c_i \bar{y}_{i\bullet}$.

4. Since each Y_{it} has the distribution of Y_i , and $Y_i \sim N(\mu_i, \sigma^2)$, it follows from standard properties of expected values that $E(\bar{Y}_{i\bullet}) = \mu_i$. Since the Y_{it}'s are independent, it follows from standard variance calculations and properties of normal random variables that $\bar{Y}_{i\bullet} \sim N(\mu_i, \sigma^2/r_i)$.

Exercise: Go through the details of comment (4). Also verify that the least squares estimator $\sum_{i=1}^{v} c_i \bar{Y}_{i\bullet}$ of the contrast $\sum_{i=1}^{v} c_i \tau_i$ (where $\sum_{i=1}^{v} c_i = 0$) has normal distribution with mean $\sum_{i=1}^{v} c_i \tau_i$ and variance $\sum_{i=1}^{v} \frac{c_i^2}{r_i} \sigma^2$. [Hint: You need to establish and use the fact that the $\bar{Y}_{i\bullet}$'s are independent.]

Variance Estimate

For the i^{th} treatment group, the sample variance is

$$s_i^2 = \frac{\sum_{t=1}^{r_i} (y_{it} - \bar{y}_{i\bullet})^2}{r_i - 1}.$$

The corresponding random variable

$$S_i^2 = \frac{\sum_{t=1}^{r_i} (Y_{it} - \bar{Y}_{i\bullet})^2}{r_i - 1}$$

is an unbiased estimator for the population variance σ^2:

$$E(S_i^2) = \sigma^2.$$

(Ross, Chapter 4 or Wackerly, Chapter 8)

As in our discussion of the two-sample t-test, the average of the S_i^2's will then also be an unbiased estimator of σ^2. To take into account different sample sizes we will take a weighted average:

$$S^2 \text{ (or } \hat{\sigma}^2) = \frac{\sum_i (r_i - 1) S_i^2}{\sum_i (r_i - 1)}$$

Note: denominator equals $\sum_i r_i - \sum_i 1$ = n - v.

Exercise (might be homework): Check that S^2 is an unbiased estimator of σ^2 -- that is, check that $E(S^2) = \sigma^2$.

Using the definition of S_i^2, see that the numerator of S^2 is

$\sum_{i=1}^{v} \sum_{t=1}^{r_i} (Y_{it} - \bar{Y}_{i\bullet})^2$ -- called SSE, the *sum of squares for error* or the *error sum of squares*.

So

S^2 = SSE/(n-v) -- called MSE, the *mean square for error* or *error mean square*.

The above are random variables. Their values calculated from the data are:

ssE = $\sum_{i=1}^{v}\sum_{t=1}^{r_i}(y_{it} - \bar{y}_{i\bullet})^2$

-- also called the *sum of squares for error* or the *error sum of squares*

msE = ssE/(n-v)

-- also called the *mean square for error* or *error mean square*

s^2 = msE -- the unbiased estimate of σ^2 -- also denoted $\hat{\sigma}^2$.

Note:

- $y_{it} - \bar{y}_{i\bullet}$ is sometimes called the it^{th} *residual*, denoted $\hat{e}_{it}$. So ssE = $\sum_{i=1}^{v}\sum_{t=1}^{r_i}\hat{e}_{it}^{\,2}$

- Many people use SSE and MSE for ssE and msE.

- This unbiased estimate of σ^2 is sometimes called the *within groups* (or *within treatments*) *variation*, since it calculates the sample variance within each group and then averages these estimates.

- *Exercise (might be homework)*:

$$\text{ssE} = \sum_{i=1}^{v}\sum_{t=1}^{r_i} y_{it}^{\,2} - \sum_{i=1}^{v} r_i \bar{y}_{i\bullet}^2$$