Least Squares Fits and Anova

Two Groups

We have seen already that when we have two groups, the two group means are the best predictors of the response, given the group membership. We have also seen that the minimzation of the squared errors can be viewed as a special case of simple linear model.

Begin with some notation. Suppose there are n_1 subjects in group 1 and n_2 subjects in group 2, where $n_1 + n_2 = n$. Also suppose a response is measured on each subject, where: $y_k =$ the response of the kth subject, $k = 1, \ldots, n$. Define the following averages,

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{k=1}^{n} y_k \\ \text{------} &= \frac{1}{\text{---}} \sum_{k \text{ in group } 1} y_k \\ \text{-------} &= \frac{1}{\text{---}} \sum_{k \text{ in group } 2} y_k \end{aligned}$$

Sum of squares

If our goals is to find minimize the squared error in predicting y_k with only the knowledge of which group the kth subject belongs to then we would minimize the following sum of squares:

$$\sum_{k \text{ in group } 1} (y_k - c_1)^2 + \sum_{k \text{ in group } 2} (y_k - c_2)^2$$

From this representation of the sum of squared errors, it is easy to see (WHY?) that

$\hat{y}_k =$ ------- if the kth observation belongs to group 1, and

$\hat{y}_k =$ ------- if the kth observation belongs to group 2.

Indicator Variable

Notice that we can introduce the indicator variable e, where $e_k = 1$ for those subjects in group 1, and $e_k = 0$ for those subjects in group 2. Using this variable we can express these means through a least squares minimization of a linear model. To see this, first note that

$$\begin{aligned} n_1 &= \sum_{k=1}^{n} e_k \\ \text{-----} &= n - \sum_{k=1}^{n} e_k \\ \text{------} &= \frac{1}{\sum e_k} \sum_{k=1}^{n} y_k e_k \\ \text{------} &= \frac{1}{\sum (1 - e_k)} \sum_{k=1}^{n} y_k (1 - e_k) \end{aligned}$$

Next consider fitting the model: $y_k = a + be_k + E_k$, where the E_k are mean 0 constant variance errors, to our data.

$$\sum_{k=1}^{n} (y_k - (a + be_k))^2$$

We know already that the minimizing value for a and b, are

$\hat{a} = \bar{y} - \hat{b}\bar{e}$
$\hat{b} = \frac{\sum (e_k y_k) - \bar{y} \sum e_k}{\sum e_k^2 - \bar{e} \sum e_k}$

But, it's easier to figure out what $\hat{a}$ and $\hat{b}$ represent in terms of $\bar{y}$, $\bar{y}_1$ and $\bar{y}_2$, by doing the minimzation directly.

$$\sum_{k=1}^{n}(y_k - (a + be_k))^2 = \sum_{k \text{ in group } 1} (y_k - (a + b))^2 + \sum_{k \text{ in group } 2} (y_k - a)^2$$

The first derivative with respect to b is

$$-2 \sum_{k \text{ in group } 1} (y_k - (a + b))$$

Solving for $\hat{b}$

$\hat{b} =$

Plug this back in and differentiate with respect to a to find

$\hat{a} =$

Testing the hypothesis that $b = 0$ reduces to the two sample test that we saw earlier.

Other parametrizations

We chose the parameterization

$$y_k = a + be_k + E_k$$

Using the geometric perspective, we see that an equivalent parameterization is

$$y_k = ce_{1,k} + de_{2,k} + E_k$$

where $e_{1,k} = 1$ if the kth observation belongs to group 1, and 0 otherwise, and where $e_{2,k} = 1$ if the kth observation belongs to group 2, and is 0 otherwise. That is, $e_{1,k}$ indicates membership in group 1, and $e_{2,k}$ indicates membership in group 2. The parameterization is equivalent because the space spanned by $\mathbf{1}$ and $\mathbf{e}$ is the same as the space spanned by $\mathbf{e}_1$ and $\mathbf{e}_2$. In fact, $\mathbf{e} =$ _____ and $\mathbf{1} =$ _____.

So, the fitted values will be the same, that is

$\hat{y}_k =$ _____, if the kth observation belongs to group 1, and
$\hat{y}_k =$ _____, if the kth observation belongs to group 2.

Note that in this case, the estimated coefficients themselves are

$\hat{c} =$ _____
$\hat{d} =$ _____.

Often, the parameterization used is

$$y_k = \alpha + \beta_1 e_{1,k} + \beta_2 e_{2,k} + E_k$$

Or
$y_k =$ _____ if the kth observation belongs to group 1, and
$y_k =$ _____ if the kth observation belongs to group 2.

What's strange about this parameterization? Note there is no unique solution to the minimization, To find one, we add the constraint that $\beta_1 + \beta_2 = 0$, In this case,

$\hat{y}_k =$ _____ if the kth observation belongs to group 1, and
$\hat{y}_k =$ _____ if the kth observation belongs to group 2.

So $\hat{\alpha} =$ _____
$\hat{\beta}_1 =$ _____
$\hat{\beta}_2 =$ _____.

New notation

The notation is often simplified by giving y a double subscript, that is, $y_{i,j}$ represents the measurement on the jth subject in the ith group. Then we often write the model as follows:

$$y_{i,j} = \alpha + \beta_i + E_{i,j}$$

So, when $i = 1$, $y_{1,j} =$ ------ $+ E_{1,j}$
and when $i = 2$, $y_{2,j} =$ ------ $+ E_{2,j}$.
Again, we need that constraint that $\beta_1 + \beta_2 = 0$ in order to have a unique parameterization.

Anova Table No matter what the parameterization, the ANOVA table looks the same. It is based on the decomposition of the sums of squares:

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i \sum_j (\bar{y}_i - \bar{y})^2$$

Fill in the table:

Source	DF	Sum of Squares	Mean Square	F-statistic
Labs				
Error		0.231		
Total		0.356		