1 Linear Models

Let's start with a brief introduction to fitting linear models. Suppose we have n data points $Y_i, i = 1, \ldots, n$ and we think there may be a linear relationship between the Y's and some p covariates, $X_{ij}, i = 1, \ldots, n, j = 1, \ldots, p$. For example the $Y's$ could represent CBF and the X's could represent different tasks, different dosage levels, etc.

This relationship can be expressed as:

$$\begin{aligned} Y_i &= \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \\ Y &= X\beta + \varepsilon \end{aligned}$$

where

- Y is an $n \times 1$ matrix
- X is an $n \times p$ matrix
- β is a $p \times 1$ matrix and
- ε is an $n \times 1$ matrix of independent random errors.

Now, in looking at the relationship between the Y's and the X's, we want to find anestimate $\widehat{\beta}$ for the coefficient matrix β. With our estimate, we get fitted values $\widehat{Y}$ as:

$$\widehat{Y} = X\widehat{\beta}$$

Usually, we are interested in the least squares solution, i.e. the solution that minimizes the sum of errors squared:

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^{n}(Y_i - x_i \cdot \widehat{\beta}) \\ &= (Y - X\widehat{\beta})^T(Y - X\widehat{\beta}) \end{aligned}$$

where x_i represents the i^{th} row of the design matrix X. Since the SSE is a quadratic, the solution will be at a minimum when the matrix $(Y - X\widehat{\beta})$ is identically zero, i.e. when

$$\begin{aligned} X\widehat{\beta} &= Y \\ (X^T X)\widehat{\beta} &= X^T Y \\ \widehat{\beta} &= (X^T X)^{-1} X^T Y \end{aligned}$$

In the case where the ε_i's are *i.i.d.* $N(0, \phi)$ random errors, the least squares solution $\widehat{\beta}$ is also the maximum likelihood solution (MLE). Here ϕ is used instead of the usual σ^2 to be consistent with the notation of the section on Generalized Linear Models (GLMs).

Now, we generally want to test some hypotheses about these coefficients, i.e. if there is an effect with dosage, or if different groups perform a task differently. These hypotheses can be expressed as testing whether a linear combination of the coefficients is equal to a given matrix, usually the 0 matrix. This brings up what is sometimes referred to as the General Linear Hypothesis. Suppose we want to test:

$$H_0 \ : C\beta = h$$

where C is a $r \times p$ matrix specifying the linear combinations of the coefficients we wish to test and h is a known $r \times 1$ matrix. To test this, we use our estimates $\widehat{\beta}$ which, being linear combinations of Gaussian random variables, are also Gaussian, with:

$$\begin{aligned} E(\widehat{\beta}) &= \beta \\ \mathrm{Var}(\widehat{\beta}) &= (X^T X)^{-1} \phi \end{aligned}$$

these two equalities can be gotten easily from the following:

$$\begin{aligned} E(MZ + K) &= ME(Z) + K \\ \mathrm{Var}(MZ + K) &= M\mathrm{Var}(Z)M^T \end{aligned}$$

where M and K are known matrices and Z is any random variable.

The above relations give us, under H_0:

$$\begin{aligned} E(C\widehat{\beta} - h) &= 0 \\ Var(C\widehat{\beta} - h) &= C(X^T X)^{-1} C^T \phi \end{aligned}$$

This leads to the following test statistics:

$$F = \frac{n-p}{r\widehat{\phi}}(C\beta - h)^T (C(X^T X)^{-1} C^T)^{-1} (C\beta - h)$$

where $\widehat{\phi} = \sum_{i=1}^{n} (Y_i - x_i \cdot \widehat{\beta})^2$

under H_0, F has an $F_{r,n-p}$ distribution. In the case ϕ is known, if we replace $\widehat{\phi}$ by ϕ, then F is distributed as χ^2_r. Also, in the case where $r = 1$, the statistic is equivalent to the two sided t-test or Z-test (depending on ϕ known or not). To get a one sided test for the case $r = 1$ we just have to look at:

$$T = \text{sign}(C\widehat{\beta} - h)\sqrt{F}$$

Again, T is distributed as t_{n-p} if ϕ is unknown and as $N(0,1)$ if ϕ is known.

2 Setting up a Linear Model

The next question we have to answer is how to get the design matrix from a given experimental setup. The linear model covers both linear regression analysis as well as analysis of variance or a mixture of the two, analysis of covariance. The main problem is figuring out how to setup the design matrix X. Here we should note, that since we used $(X^T X)^{-1}$ in the previous section, we were implicitly assuming that the matrix X is of full rank. That means, basically, that no columns can be linear combinations of the other columns. This can be a little confusing at first when trying to write out the design matrix X.

Example 2.1 Suppose we had the following experimental set-up: 12 patients were split up into two groups: one receiving a new treatment, the other receiving a placebo. We are interested in seeing if there was any effect of the drug as well as if the subject's age had an effect. What would the

design matrix look like? It's often "standard" practice to fit an intercept term by default to the data, to normalize the data. So, our observations Y_i can be thought of as:

$$\begin{aligned} Y_i &= \beta_0 + \text{age}_i \times \beta_{\text{age}} + \beta_{\text{placebo}} \times (\text{subject } \ i \text{ received } \ \text{placebo}) + \\ &\quad \beta_{\text{treat}} \times (\text{subject } \ i \text{ received } \ \text{treatment}) + \varepsilon_i \end{aligned}$$

So it would be tempting to create the following design matrix:

$$X = \begin{pmatrix} 1 & \text{age}_1 & 1 & 0 \\ 1 & \text{age}_2 & 1 & 0 \\ 1 & \text{age}_3 & 1 & 0 \\ 1 & \text{age}_4 & 1 & 0 \\ 1 & \text{age}_5 & 1 & 0 \\ 1 & \text{age}_6 & 1 & 0 \\ 1 & \text{age}_7 & 0 & 1 \\ 1 & \text{age}_8 & 0 & 1 \\ 1 & \text{age}_9 & 0 & 1 \\ 1 & \text{age}_1 0 & 0 & 1 \\ 1 & \text{age}_1 1 & 0 & 1 \\ 1 & \text{age}_1 2 & 0 & 1 \end{pmatrix}$$

However, this matrix is not of full rank, because $x_{i4} = x_{i1} - x_{i3}$, so the 4^{th} column is a linear combination of the 1^{st} and 3^{rd} columns. So a proper design matrix would be :

$$X = \begin{pmatrix} 1 & \text{age}_1 & 1 \\ 1 & \text{age}_2 & 1 \\ 1 & \text{age}_3 & 1 \\ 1 & \text{age}_4 & 1 \\ 1 & \text{age}_5 & 1 \\ 1 & \text{age}_6 & 1 \\ 1 & \text{age}_7 & 0 \\ 1 & \text{age}_8 & 0 \\ 1 & \text{age}_9 & 0 \\ 1 & \text{age}_{10} & 0 \\ 1 & \text{age}_{11} & 0 \\ 1 & \text{age}_{12} & 0 \end{pmatrix}$$

The β's for this model have a slightly different interpretation than those written above, this is a matrix for the following model:

$$Y_i = \beta_0 + \text{age}_i \times \beta_{\text{age}} + \beta_{\text{placebo}} \times (\text{subject} \quad i \text{ received} \quad \text{placebo}) + \varepsilon_i$$

So β_0 is no longer the grand mean anymore, it is the mean for the treatment group, after the age effect is removed.

This sort of problem will always happen in any sort of ANOVA design, where we have more than one group. The solution is to introduce a constraint for each different "grouping". For instance, we could have a two-way ANOVA, where we have 4 different treatment types, and 5 different patient groups. We would have to introduce a constraint on the parameters concerning treatment types, as well as one on the parameters concerning patient groups. With these constraints we would add 3 columns for the effect of treatment and 4 columns for the different patient groups. The question then is what sort of constraint. In fact, we can use any linear constraint we want on the β's in question, but then what is the easiest one? Primarily, though it should lead to easy interpretation about the parameters of the model. For instance, what if we wanted to test if the effect for patient group #5 is the same as patient group #4 above. How would we do that?

There are a few "standard" ways that this is done, I find the easiest to interpret is constraining the sum of the effects to be zero.

Example 2.2 Let's take the 4 treatment $\times$ 5 patient group scenario as above. There will be 4 different treatment effects (say $\beta_1, \ldots, \beta_4$) and 5 different group effects ($\beta_5, \ldots, \beta_9$). But we know we can only add 3 columns for the treatments and 4 for the groups. Setting the sums of the effects to

be zero leads to the following constraints:

$$\begin{aligned} \sum_{j=1}^{4} \beta_j &= 0 \\ \sum_{j=5}^{9} \beta_j &= 0 \\ \text{or, rearranging } \beta_4 &= -\sum_{j=1}^{3} \beta_j \\ \beta_9 &= -\sum_{j=5}^{8} \beta_j \end{aligned}$$

Our design matrix would than be made up of the columns for β_0 (column of 1's), $\beta_1, \ldots, \beta_3$ and $\beta_5, \ldots, \beta_8$. Whenever we have a subject that is in the 4^{th} treatment group we set $\beta_1 = \beta_2 = \beta_3 = -1$, similarly, when a subject is in the 5^{th} patient group, we set $\beta_5 = \beta_6 = \beta_7 = \beta_8 = -1$, otherwise we set the corresponding β to 1.

Example 2.3 Let's write out the matrix for a design where we have 3 treatment groups and 3 patient groups and 2 patients in each of the 9 different categories, along with CBF also having an effect. Using the same convention

as above, our matrix would be:

$$X = \begin{pmatrix} 1 & \mathrm{CBF}_1 & 1 & 0 & 1 & 0 \\ 1 & \mathrm{CBF}_2 & 1 & 0 & 1 & 0 \\ 1 & \mathrm{CBF}_3 & 0 & 1 & 1 & 0 \\ 1 & \mathrm{CBF}_4 & 0 & 1 & 1 & 0 \\ 1 & \mathrm{CBF}_5 & -1 & -1 & 1 & 0 \\ 1 & \mathrm{CBF}_6 & -1 & -1 & 1 & 0 \\ 1 & \mathrm{CBF}_7 & 1 & 0 & 0 & 1 \\ 1 & \mathrm{CBF}_8 & 1 & 0 & 0 & 1 \\ 1 & \mathrm{CBF}_9 & 0 & 1 & 0 & 1 \\ 1 & \mathrm{CBF}_{10} & 0 & 1 & 0 & 1 \\ 1 & \mathrm{CBF}_{11} & -1 & -1 & 0 & 1 \\ 1 & \mathrm{CBF}_{12} & -1 & -1 & 0 & 1 \\ 1 & \mathrm{CBF}_{13} & 1 & 0 & -1 & -1 \\ 1 & \mathrm{CBF}_{14} & 1 & 0 & -1 & -1 \\ 1 & \mathrm{CBF}_{15} & 0 & 1 & -1 & -1 \\ 1 & \mathrm{CBF}_{16} & 0 & 1 & -1 & -1 \\ 1 & \mathrm{CBF}_{17} & -1 & -1 & -1 & -1 \\ 1 & \mathrm{CBF}_{18} & -1 & -1 & -1 & -1 \end{pmatrix}$$

where columns 3 & 4 correspond to treatment group and columns 5 & 6 correspond to patient group. So patient #15 is in treatment group #2 and patient group #3.