2

Two or more categorical predictors

Here we extend the ANOVA methods to handle multiple categorical predictors. The statistician has to watch carefully to see whether the effects being considered are properly treated as fixed or random. Just turning the variables into binary indicators and doing regressions is like treating them all as fixed and this can give wrong answers.

2.1 Two fixed effects

The simplest setting has two categorical variables interpreted as fixed effects. The first variable has levels $i = 1, \ldots, I$ and the second has levels $j = 1, \ldots, J$. The cell mean model for the data is

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \tag{2.1}$$

for $k = 1, \ldots, n_{ij}$, where ε_{ijk} are independent random variables with mean zero. The standard model for them is $\varepsilon_{ijk} \sim N(0, \sigma^2)$. Because we need a third index, we no longer run i from 1 to k.

To focus on the essentials, we will consider only the balanced case where $n_{ij} = n$ for all i and j. For the more general case one can turn to standard books on experimental design and anova. In industrial statistics, there is the book by Box, Hunter and Hunter as well as one by Montgomery. In pscychology there is a book by Winer.

Usually we want to write the cell means as a sum of a global mean, plus effects for the two categorical variables, plus an interaction term. We define the

global mean to be

$$\mu = \bar{\mu}_{\cdot\cdot} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij}.$$

We ordinarily write μ instead of $\bar{\mu}_{\cdot\cdot}$ for this average.

The average effect for level i of the first treatment is defined to be

$$\alpha_i = \frac{1}{J}\sum_{j=1}^{J}(\mu_{ij} - \mu) = \bar{\mu}_{i\cdot} - \mu.$$

Notice that α_i is defined as the average effect, up or down, on μ_{ij} compared to the baseline μ. The average is over all values of j. Similarly

$$\beta_j = \frac{1}{I}\sum_{i=1}^{I}(\mu_{ij} - \mu)$$

is the average effect for level j of the second treatment.

The interaction between levels i and j of these two treatments is

$$(\alpha\beta)_{ij} = \mu_{ij} - \mu - \alpha_i - \beta_j = \mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \mu.$$

The new quantity $(\alpha\beta)_{ij}$ is the interaction or synergy at level i of the first variable and level j of the second. It is not a product of any α and β. Instead $(\alpha\beta)$ is just a combined symbol for their interaction.

The two factor anova model for fixed effects is then

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}. \tag{2.2}$$

It is overparameterized but we compensate by imposing constraints

$$0 = \sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij}. \tag{2.3}$$

The main effects α and β as well as the interaction $(\alpha\beta)$ defined above obey these constraints.

As before there is an ANOVA decomposition for the observations. Now it is

$$\mathrm{SST} = \mathrm{SSA} + \mathrm{SSB} + \mathrm{SSAB} + \mathrm{SSE},$$

where

$$\begin{aligned}
\mathrm{SST} &= \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{...})^2 \\
\mathrm{SSA} &= \sum_i \sum_j \sum_k (\bar{Y}_{i..} - \bar{Y}_{...})^2 \\
\mathrm{SSB} &= \sum_i \sum_j \sum_k (\bar{Y}_{.j.} - \bar{Y}_{...})^2 \\
\mathrm{SSAB} &= \sum_i \sum_j \sum_k (\bar{Y}_{..k} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 \\
\mathrm{SSE} &= \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{..k})^2.
\end{aligned}$$

Several of those sum of squares formulas simplify because they ignore one or more of their indices. Conceptually it is simpler to run them all over all IJn data values, though when we study them or compute them individually we take advantage of the simplification.

The ensemble of data values forms a vector in $\mathbb{R}^{IJn}$. The sums of squares are the squared lengths of that vector after projecting it onto spaces of dimensions $IJn-1$, $I-1$, $J-1$, $(I-1)(J-1)$, and $IJ(n-1)$ respectively. These are the degrees of freedom in the sums of squares.

An intuitive reason why the interaction has $(I-1)(J-1)$ degrees of freedom is that we are free to specify any values we like for $(\alpha\beta)_{ij}$ for $i = 1, \ldots, I-1$ and $j = 1, \ldots, J-1$. But having set those $(I-1)(J-1)$ values, the rest are determined by equation (2.3). We get $(\alpha\beta)_{Ij} = -\sum_{i=1}^{I-1}(\alpha\beta)_{ij}$ for $j = 1, \ldots, J-1$ and then $(\alpha\beta)_{iJ} = -\sum_{j=1}^{J-1}(\alpha\beta)_{ij}$ for $i = 1, \ldots, I$.

Using methods we've seen before, the distribution theory gives us

$$\begin{aligned}
\mathrm{SSA} &\sim \sigma^2 \chi'^2_{(I-1)}\left(\frac{nJ\sum_i \alpha_i^2}{\sigma^2}\right) \\
\mathrm{SSB} &\sim \sigma^2 \chi'^2_{(J-1)}\left(\frac{nI\sum_j \beta_j^2}{\sigma^2}\right) \\
\mathrm{SSAB} &\sim \sigma^2 \chi'^2_{(I-1)(J-1)}\left(\frac{n\sum_i\sum_j (\alpha\beta)_{ij}^2}{\sigma^2}\right), \quad \text{and,} \\
\mathrm{SSE} &\sim \sigma^2 \chi^2_{(IJ(n-1))}.
\end{aligned}$$

The noncentrality parameters look a bit simpler if we replace factors I, J, and n by redundant sums over i, j, and k respectively.

The expected value of $\chi'^2_\nu(\lambda)$ is $\nu + \lambda$. Dividing each sum of squares by its

degrees of freedom and taking expectations leads us to

$$\begin{aligned} E(\text{MSA}) &= \sigma^2 + \frac{nJ\sum_i \alpha_i^2}{(I-1)\sigma^2} \\ E(\text{MSB}) &= \sigma^2 + \frac{nI\sum_j \beta_j^2}{(J-1)\sigma^2} \\ E(\text{MSAB}) &= \sigma^2 + \frac{n\sum_i\sum_j (\alpha\beta)_{ij}^2}{(I-1)(J-1)\sigma^2}, \quad \text{and,} \\ E(\text{MSE}) &= \sigma^2. \end{aligned}$$

What we see in each of these is that a large or at least nonzero value for an effect makes the corresponding mean square larger than σ^2 on average. As a result we use F tests based on MSA/MSE, MSB/MSE, and MSAB/MSE to test for the main effects of the first factor, second factor, and interaction respectively. Those F ratios have F distributions when their corresponding effects are zero and non-central F distributions otherwise.

In class we looked at an example from Box, Hunter and Hunter. There were 3 toxins, 4 antidotes and 4 replicates at each of the 12 combinations. The response was survival time. The F tests for toxin and antidote were statistically significant. For the interaction they had $F = 1.9$ with 6 numerator and 36 denominator degrees of freedom. Because $\Pr(F_{6,36} \geq 1.9) \doteq 0.108$ this result is not statistically significant. There is no apparent interaction.

2.1.1 What if $n = 1$?

If $n = 1$ then SSE $= 0$ and MSE becomes $0/0$. In this case we cannot use the usual F tests. A common practice is to use MSA/MSAB to test for the first factor and MSB/MSAB for the second. What we get is a doubly non-central F distribution for the ratio. The denominator is inflated by the amount of the interaction among the variables. This reduces the power of the F test. If that F test still rejects an hypothesis like $\alpha_1 = \cdots = \alpha_i = 0$, then we can be confident that the effect is real. If it fails to reject it might just be that the interaction was large enough to hide the main effect.

2.1.2 Pooling

Suppose that MSAB/MSE is not very large. Then we might be tempted to decide that there is no interaction and form a pooled error estimate

$$\frac{\text{SSE} + \text{SSAB}}{IJ(n-1) + (I-1)(J-1)} = \frac{\text{SSE} + \text{SSAB}}{IJ - I - J + 1}.$$

If we knew for certain that there was no interaction then this would be a better error estimate. It would have more degrees of freedom.

But if we are not certain that there is interaction, then pooling the error terms is problematic, especially if our rule is to pool if and only if MSAB is

not large. That rule would bring a downward bias to the variance estimate and make tests unreliable.

In principle we can do a Monte Carlo simulation to find out the effects of pooling. In practice that would be awkward to use.

2.2 Two random effects

Now suppose that there are two random effects. Perhaps we are studying a set of I doctors treating a chronic condition in each of J patients. Then we could look at the short term impact of the treatment in a study where every doctor sees every patient. For this setting to be appropriate we have to rule out the possibility that one of the doctors completely cures the patient. More generally there may be follow on effects, but suppose for sake of illustration that follow on effects are small.

The two factor anova for random effects takes the form

$$Y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + \varepsilon_{ijk}$$

and we make the assumptions that

$$\begin{aligned} a_i &\sim N(0, \sigma_A^2), \\ b_j &\sim N(0, \sigma_B^2), \\ (ab)_{ij} &\sim N(0, \sigma_{AB}^2), \quad \text{and} \\ \varepsilon_{ij} &\sim N(0, \sigma_E^2) \end{aligned}$$

are all independent.

The ANOVA identity still holds when we replace a fixed effect model by a random effects model, just as it held up for the one way layout. The distributions of the sums of squares change though. We get

$$\begin{aligned} \mathrm{SSA} &\sim \Big(\sigma_E^2 + n\sigma_{AB}^2 + Jn\sigma_A^2\Big)\chi^2_{(I-1)} \\ \mathrm{SSB} &\sim \Big(\sigma_E^2 + n\sigma_{AB}^2 + In\sigma_B^2\Big)\chi^2_{(J-1)} \\ \mathrm{SSAB} &\sim \Big(\sigma_E^2 + n\sigma_{AB}^2\Big)\chi^2_{(I-1)(J-1)} \\ \mathrm{SSE} &\sim \sigma_E^2\chi^2_{(IJ(n-1))}. \end{aligned}$$

For each of these, the expected value of the mean square is just the factor to the left of the chi-squared random variable.

As before we test whether there is an AB interaction by comparing MSAB/MSE to a quantile of the $F_{(I-1)(J-1),IJ(n-1)}$ distribution. But it no longer makes sense to test whether σ_A^2 equals zero via MSA/MSE. The numerator could be large if σ_{AB}^2 is large enough, even if $\sigma_A^2 = 0$. The appropriate test is instead based on

$$\frac{\mathrm{MSA}}{\mathrm{MSAB}} = \frac{\frac{1}{I-1}\mathrm{SSA}}{\frac{1}{(I-1)(J-1)}\mathrm{SSAB}} \sim \left(1 + \frac{Jn\sigma_A^2}{\sigma_E^2 + n\sigma_{AB}^2}\right) F_{I-1,(I-1)(J-1)}.$$

The MSE does not capture the real uncertainty in the data if we're interested in factors A or B. Suppose in the extreme that doctor i always brought the exact same benefit to patient j for any ij combination. Then $\sigma_E^2 = 0$ and MSE $= 0$ too. Then all our factors would look significant at any level, even with only $I = J = 2$.

2.3 Mixed effect models

Things get more complicated when there is one fixed and one random effect in the data. For example one factor might be a small number of drugs that were tried on a larger number of subjects. The subjects are properly considered a random effect because we may want to know how the drugs will work in the population as a whole. The drugs are fixed effects, assuming that we want comparisons among that exact group of drugs.

The two factor mixed effects model has

$$Y_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + \varepsilon_{ijk},$$

where α_i are fixed numbers that sum to 0, $b_j \sim N(0, \sigma_B^2)$ and $\varepsilon_{ijk} \sim N(0, \sigma_E^2)$. The exact way to handle $(\alpha b)_{ij}$ was controversial for a while, but was settled by Cornfield and Tukey in 1956. They found that a proposal of Anderson and Bancroft in 1952 gave the right expected mean squares.

Anderson and Bancroft's model has $(\alpha b)_{ij} \sim N(0, \sigma_{AB}^2)$ individually but with negative correlations among $(\alpha b)_{1j}, \ldots, (\alpha b)_{Ij}$ so that $\sum_{i=1}^{I} (\alpha b)_{ij} = 0$ for each $j = 1, \ldots, J$.

Cornfield and Tukey settled matters using an ingenious 'pigeonhole' model. They supposed that there was a large finite table with $\mathcal{I}$ rows and $\mathcal{J}$ columns. At each of the $\mathcal{I}\mathcal{J}$ cells of this table was a pigeonhole stuffed with N numbers. The data arise by randomly selecting I of the $\mathcal{I}$ rows and J of the $\mathcal{J}$ columns. Then from the selected pigeonholes they sample n of the N numbers.

To sample from distributions at each combination, they take the limit $N \to \infty$. Fixed effects correspond to $I = \mathcal{I}$ (or $J = \mathcal{J}$) while random effects arise in the limits $\mathcal{I} \to \infty$ and $\mathcal{J} \to \infty$ respectively. Their model even accounts for an intermediate situation with $0 < I < \mathcal{I} < \infty$. If for example we sample 15 states out of 50, then we are in this intermediate setting.

The sums of squares for the mixed effects regression are

$$\begin{aligned}
\text{SSA} &\sim \Big(\sigma_E^2 + n\sigma_{AB}^2\Big)\chi'^2_{(I-1)}\Big(\frac{nJ\sum_{i=1}^{I}\alpha_i^2}{\sigma_E^2 + n\sigma_{AB}^2}\Big)\\
\text{SSB} &\sim \Big(\sigma_E^2 + n\sigma_B^2\Big)\chi^2_{(J-1)}\\
\text{SSAB} &\sim \Big(\sigma_E^2 + n\sigma_{AB}^2\Big)\chi^2_{(I-1)(J-1)}\\
\text{SSE} &\sim \sigma_E^2\chi^2_{(IJ(n-1))}.
\end{aligned}$$

The F test for A is MSA/MSAB. The F test for B is MSB/MSE, and the F test for AB is MSAB/MSE. The interaction test has not changed. The test

for the fixed effect divides by the interaction as we did for two random effects. Similarly the test for the random effect divides by the error term as we did for two fixed effects.

We have an 'opposite effect rule' wherein the test for one factor depends on what the other factor is. If the other factor is fixed, then use MSE while if the other factor is random, then use MSAB. Intuitively, this is reasonable. The opposite effect is the one that drives the quality of the data set for the effect being tested.

Treating everything as a fixed effect can give very misleading answers. Suppose that we have 10 subjects each trying 3 medications and we repeat the process 5 times. Of course we have 150 numbers. If instead we did 50 repeats we would have 1500 numbers, but could not really claim to have 10 times the information. We still just have 10 subjects. Even if we measured them infinitely often, we would end up with a 10×3 matrix of mean values from which to learn how patients respond to the various treatments. Raising n gives us 30 better values but not $30n$ values, for the purposes of studying drug differences.

2.4 Higher way tables

In class we looked at expected mean squares for a 3 factor experiment. Sometimes there is a factor in there with no clearly appropriate denominator for an F test. This happens in the 3 way setting with all effects random. There is no suitable mean square to test the main effects.

Suppose that we have factors A, B, and C and that we have observed p of P levels for A, q of Q levels for B, and r of R levels for C. At each level we have a sample of n values out of N possible. Our usual models sample from a hypothetical distribution with $N = \infty$.

Here are the expected mean squares:

$$
\begin{aligned}
E(\mathrm{MSA}) &= (1-\frac{n}{N})\sigma^2_E + n(1-\frac{q}{Q})(1-\frac{r}{R})\sigma^2_{ABC} + nq(1-\frac{r}{R})\sigma^2_{AC} + nr(1-\frac{q}{Q})\sigma^2_{AB} + nqr\sigma^2_A\\
E(\mathrm{MSB}) &= (1-\frac{n}{N})\sigma^2_E + n(1-\frac{p}{P})(1-\frac{r}{R})\sigma^2_{ABC} + np(1-\frac{r}{R})\sigma^2_{BC} + nr(1-\frac{p}{P})\sigma^2_{AB} + npr\sigma^2_B\\
E(\mathrm{MSC}) &= (1-\frac{n}{N})\sigma^2_E + n(1-\frac{p}{P})(1-\frac{q}{Q})\sigma^2_{ABC} + np(1-\frac{q}{Q})\sigma^2_{BC} + nq(1-\frac{p}{P})\sigma^2_{AC} + npq\sigma^2_C\\
E(\mathrm{MSAB}) &= (1-\frac{n}{N})\sigma^2_E + n(1-\frac{r}{R})\sigma^2_{ABC} + nr\sigma^2_{AB}\\
E(\mathrm{MSAC}) &= (1-\frac{n}{N})\sigma^2_E + n(1-\frac{q}{Q})\sigma^2_{ABC} + nq\sigma^2_{AC}\\
E(\mathrm{MSBC}) &= (1-\frac{n}{N})\sigma^2_E + n(1-\frac{p}{P})\sigma^2_{ABC} + np\sigma^2_{BC}\\
E(\mathrm{MSABC}) &= (1-\frac{n}{N})\sigma^2_E + n\sigma^2_{ABC}\\
E(\mathrm{MSE}) &= (1-\frac{n}{N})\sigma^2_E.
\end{aligned}
$$

The quantity σ^2_A is a usual variance when A is a random effect. It is $(I -$

$1)^{-1}\sum_{i=1}^{I}\alpha_i^2$ for a fixed effect. See Cornfield and Tukey for intermediate cases.

2.5 Other

Here are some further topics, adjacent to, but outside this course.

2.5.1 High level factorials

Industrial experiments often feature many factors each run at 2 levels. For L factors we might need 2^L experimental units, and more if we replicate. There are strategies for cutting down the costs by running only half or fewer of the units. The tradeoff is that the smaller experiments do not allow you to estimate the high order interactions. This is acceptable if most of the interest is in the low order interactions. The text "Statistics for Experimenters" by Box, Hunter and Hunter, is a good place to look for solutions to this problem.

2.5.2 Nested and crossed effects

The above discussion has assumed crossed effects. Every level of each factor is observed in conjunction with every level of every other factor.

Nested factors are different. As an example, suppose that we sample 10 apples at random from each of 5 trees. There is no connection between apple 1 from one tree and apple 1 from another tree. It would not then make sense to test for systematic differences between apple 1 and apple 2 from multiple trees. The factor for apples is said to be nested inside that for trees.

When A is nested within B we describe the variable via $A(B)$. We define a sum of squares for it via SSA(B) = SSA+SSAB on $I-1+(I-1)(J-1) = (I-1)J$ degrees of freedom.

Factors can be nested to many levels: A within B within C and so on. Nesting and crossing can both be present, in the same problem. Factors A and B could be crossed within the levels of C which is nested inside D. One or more factors can be nested within the interactions of two or more other factors.

There are automated formulas for getting the expected mean squares in these settings. It is however beyond the scope of this course. The text "Design and Analysis of Experiments" by D. C. Montgomery is one place to look if you face this problem.