NRES 798 — Lab 10

Categorical predictors, ANOVA

1 Foundations

Following Section 1 from the Chapter 10 notes.

Create the data set, one way is:

```
> ants <- data.frame(nests = c(9, 12, 9, 6, 4, 10),
+ habitat = c(’field’, ’field’, ’forest’, ’forest’,
+ ’forest’, ’scrub’))
```

Inspect it: `print` (or `ants <Enter>`), `summary`, `str`.

Fitting a linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \ ,$$

where the predictors are dummy variables for the levels of `habitat`. Generate the dummy variables, e.g., `ants$field = as.numeric(ants$habitat == ’field’)`, etc. Think about how this works. Display `ants`.

Look at the dummy variables, and convince yourself that if observation j is in habitat i, then the regression equation for that observation becomes

$$Y_j = \beta_0 + \beta_i + \varepsilon_j \ .$$

Try fitting the regression model on the 3 dummy variables: `summary(lm(nests ∼ field + ..., ants))` (remember that the intercept is included by default). What happens?

There is a redundancy in the parameters that needs to be resolved by imposing some linear constraint (contrast). Simple constraints are fixing one of the β's at 0. For instance, $\beta_3 = 0$. See from the regression equation that this is the same as omitting the third level, `scrub`. Try it, fit `nests` $\sim$ `field + forest`. Compare to the previous result.

Instead of $\beta_3 = 0$, suppress the intercept ($\beta_0 = 0$). Remember that that is done by including `- 1` in the formula. How are these β's related to those in the previous regression? Hint: you can get three equations relating the β_i from the previous model and the β_i' for the new model, by noting that the expected Y's that were $\beta_0 + \beta_i$ become simply β_i' (and $\beta_3 = \beta_0' = 0$). Confirm with the parameter estimates that you obtained.

Fit the dummy variable regression omitting the first level. Now, use the R notation for categorical linear models: `nests` $\sim$ `habitat`. Compare.

Add another factor: `ants$site <- as.factor(c('east', 'west', 'east', 'west', 'east', 'west'))`. Generate the corresponding dummy variables. Inspect the data.

Fit a dummy variable regression omitting the first level from each factor. Then, fit `nests` $\sim$ `habitat + site`. Compare.

The standard notation for the one-factor model uses $\mu + \alpha_i$ instead of $\beta_0 + \beta_i$, with the constraint (contrast) $\sum \alpha_i = 0$. The default contrast for unordered factors in R is $\beta_1 = 0$, that is, omit the first level. The μ/α parametrization for `nests` $\sim$ `habitat` can be obtained by adding the `lm` argument `contrast = list(habitat = 'contr.sum')` (the default is `'contr.treatment'`). Try it. The intercept is $\hat{\mu}$. The last $\hat{\alpha}_i$ is omitted, but can be obtained as minus the sum of the others (all of them add up to 0). Compare to the default contrast. If you feel like it, try to figure out the relationship between the two sets of parameters.

Run the two-factor model with the `sum` contrast. Instead of giving a list of contrasts for each factor as an argument, the type of contrast to be used can be set as a global option: `oldopt <- options(contrasts = c('contr.sum', 'contr.poly'))` (the second entry `'contr.poly'` is the default for ordered factors). The old defaults can be reset later with `options(oldopt)`. Compare the fit statistics with those for the default contrasts. It is the same model, just written differently.

2 Hypothesis testing, ANOVA

Section 3.2 in the Chapter 10 notes.

The usual hypothesis test is for some factor having an effect on the response or not. For instance, in a one-factor model the null and the alternative hypotheses are

$$H_0 : Y_{ij} = \mu + \varepsilon_{ij} \ , \qquad H_a : Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \ .$$

The tests use an F-ratio computed from the residual sums of squares and degrees of freedom from both models:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_a)/(\mathrm{df}_0 - \mathrm{df}_a)}{\mathrm{RSS}_a/\mathrm{df}_a} \ .$$

With the `ants` data, test for the `habitat` effect. Fit the null model `H0 <- lm(nests ~ 1, ants)`, and the model `Ha` with `habitat` as the predictor. Compute the F-ratio. Remember that the RSE given by `summary` is $\sqrt{\text{RSS / df}}$.

Under H_0, this F has an F distribution with $\mathrm{df}_0 - \mathrm{df}_a$ and df_a degrees of freedom. Compute the p-value. Is the effect of habitat significant?

Get an ANOVA table for the difference between the two models: `anova(H0, Ha)`. Compare to your results.

A conventional ANOVA table for the full model H_a can be obtained as `anova(Ha)` or as `summary(aov(nests ~ habitat, ants))`. "Usually" the table gives the correct F and p-values for testing the significance of the predictors. Check.

Repeat for

$$H_0 : Y_{ijk} = \mu + \beta_j + \varepsilon_{ijk} \quad \textit{vs.} \quad H_a : Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \ ,$$

where α is habitat and β is site.

What happened with the ANOVA table? Try swapping the order of the factors (`site + habitat`). This data is unbalanced, and the ANOVA calculations do not work.

Try with a balanced data set, back to the last part of the previous lab:

Load the CO2 data set: `data(CO2)`. Inspect it: `summary, head, str, ?CO2`.

Test for the effect of `Treatment` in `uptake` $\sim$ `Type + Treatment`.