

# Completely Random Designs

BIOE 498/598

2/10/2020

# Three Types of Variables

- ▶ **Numerical** (or **continuous**) variables are modeled by real numbers using a single coefficient.
- ▶ **Ordinal** variables have discrete but *ordered* levels. If the levels are evenly spaced, we model them using integers.
- ▶ **Nominal** (or **categorical**) variables are unordered with no numeric relationship between levels.

# One-hot encoding

- ▶ In one-hot encoding, a nominal variable with  $k$  levels is modeled with  $k$  binary dummy variables.
- ▶ Only one dummy variable is nonzero ("hot") at a time.
- ▶ Example:  $\text{DNA} \in \{\text{A}, \text{C}, \text{G}, \text{T}\}$ .

$$\beta_A x_A + \beta_C x_C + \beta_G x_G + \beta_T x_T$$

## Fitting models with one-hot encoded variables

Consider a model with  $x \in \{\text{low, medium, high}\}$ :

$$y = \beta_0 + \beta_{\text{low}}x_{\text{low}} + \beta_{\text{med}}x_{\text{med}} + \beta_{\text{high}}x_{\text{high}}$$

which, after fitting is

$$y = 60 + 12x_{\text{low}} - 20x_{\text{med}} + 30x_{\text{high}}$$

where

$$y(x_{\text{low}} = 1) = 72, \quad y(x_{\text{med}} = 1) = 40, \quad y(x_{\text{high}} = 1) = 90$$

## Fitting models with one-hot encoded variables

Consider a model with  $x \in \{\text{low, medium, high}\}$ :

$$y = \beta_0 + \beta_{\text{low}}x_{\text{low}} + \beta_{\text{med}}x_{\text{med}} + \beta_{\text{high}}x_{\text{high}}$$

which, after fitting is

$$y = 60 + 12x_{\text{low}} - 20x_{\text{med}} + 30x_{\text{high}}$$

where

$$y(x_{\text{low}} = 1) = 72, \quad y(x_{\text{med}} = 1) = 40, \quad y(x_{\text{high}} = 1) = 90$$

We could define another model with equivalent predictions:

$$y = 50 + 22x_{\text{low}} - 10x_{\text{med}} + 40x_{\text{high}}$$

# Degeneracy

There are infinitely many models with coefficients

$$\beta_0 - \Delta, \quad \beta_{\text{low}} + \Delta, \quad \beta_{\text{med}} + \Delta, \quad \beta_{\text{high}} + \Delta$$

all with the same predictions, residuals, etc.

To avoid the degeneracy, R will not estimate the first (or *base*) level of a factor variable if the model has an intercept. This ensures a unique solution.

## Degeneracy in Matrix Form

Consider a design matrix with an intercept, a three-level categorical variable, and two replicates:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

This matrix is not full rank since the columns are not linearly independent. ( $\mathbf{X}(:, 1) = \mathbf{X}(:, 2) + \mathbf{X}(:, 3) + \mathbf{X}(:, 4)$ ). If we drop any column the matrix will be full rank; R's choice to drop the second column is arbitrary.

# Contrasts in the Rothamsted Experiment

The sugar beet experiment is modeled as a single variable with four treatments:

- ▶ (A) no fertilizer
- ▶ (B) plowed fertilizer in January
- ▶ (C) broadcast fertilizer in January
- ▶ (D) broadcast fertilizer in April

By default the first treatment (A) will be absorbed into the intercept. The remaining effect sizes are relative to the no fertilizer treatment:

$$\text{yield} = \beta_A + \beta_B x_B + \beta_C x_C + \beta_D x_D$$



## Contrasts in the Rothamsted Experiment

- ▶ (A) no fertilizer
- ▶ (B) plowed fertilizer in January
- ▶ (C) broadcast fertilizer in January
- ▶ (D) broadcast fertilizer in April

What if we wanted to make other comparisons?

- ▶ Effect of broadcast vs. plowed:  $(C \ \& \ D) = (B)$
- ▶ Effect of early vs. late application:  $(B \ \& \ C) = (D)$
- ▶ Effect of any fertilizer:  $(A) = (B, C, \ \& \ D)$

## Contrasts in the Rothamsted Experiment

- ▶ (A) no fertilizer
- ▶ (B) plowed fertilizer in January
- ▶ (C) broadcast fertilizer in January
- ▶ (D) broadcast fertilizer in April

What if we wanted to make other comparisons?

- ▶ Effect of broadcast vs. plowed:  $(C \ \& \ D) = (B)$
- ▶ Effect of early vs. late application:  $(B \ \& \ C) = (D)$
- ▶ Effect of any fertilizer:  $(A) = (B, C, \ \& \ D)$

There are all *contrasts*, or comparisons between effect sizes. The null hypotheses for each contrast can be written as a linear combination of the model's coefficients:

$$\frac{1}{2}\beta_C + \frac{1}{2}\beta_D - \beta_B = 0$$

When specifying contrasts, we require that the coefficient sum to zero (hence the 1/2 factors above).

## How do we test contrasts?

- ▶ Fit a linear model with a categorical variable:

```
model <- lm(y ~ var1 + var2)
```

- ▶ Let's say var1 had three levels and we wanted to test if  $\beta_1 = (\beta_2 + \beta_3)/2$ . First we define the *contrast coefficients* for the null hypothesis.

```
contrast <- c(1, -0.5, -0.5, 0)
```

- ▶ Then we use the fit.contrast function from the gmodels package to test the contrast.

```
gmodels::fit.contrast(model, var1, contrast)
```

We can also test multiple contrasts at the same time using a contrast matrix as shown in the textbook.

## Can we test any contrasts?

No. A contrast must be *estimable* for it to be tested. A contrast is estimable if

- ▶ its coefficients sum to zero
- ▶ it can be expressed as a linear combination of the rows of the design matrix.

## Estimable Example: Measuring only main effects

$$\mathbf{X} = \begin{matrix} & \beta_0 & \beta_2 & \beta_3 & \beta_{12} & \beta_{13} & \beta_{23} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

Any contrast about the interaction terms is not estimable. To test for an  $\beta_{12}$  effect ( $H_0 : \beta_{12} - \beta_0 = 0$ )

$$c = (1 \quad 0 \quad 0 \quad -1 \quad 0 \quad 0)$$

which is not a combination of the rows in  $\mathbf{X}$ . In fact, we cannot fit this model since the interaction terms are confounded!