

Chapter 4: Linear models with categorical predictors

1. In this exercise, you will analyze a dataset from a classic study¹ on dopamine beta-hydroxylase (DBH) activity in schizophrenia patients. DBH is an enzyme involved in the metabolism of dopamine, a neurotransmitter often implicated in psychiatric disorders. The researchers hypothesized that DBH activity in the cerebrospinal fluid might differ between patients who responded well to treatment (*nonpsychotic*) and those who did not (*psychotic*), and that DBH activity could potentially serve as a *biomarker* for predicting treatment outcomes in schizophrenia.

The dataset contains two variables:

- `group`: whether the patient was judged *nonpsychotic* or *psychotic* following treatment
- `dbh_activity`: DBH activity levels in nmol/(ml)(h)(mg) of protein.

You can load the data using the following command (after ensuring the file `dbh_data.csv`, which you can find in the companion website, is available in your working directory):

```
dbh_data <- read.csv("dbh_data.csv")
```

- Use the `str()` function to examine the structure of the dataset.
 - Identify which variables are numeric and which are factors.
 - Convert the `group` variable to a factor, if it is not already.
 - Use the `levels()` function to list the levels of the `group` factor.
2. Fit a linear model to predict DBH activity using the `group` variable as a predictor.
 - Interpret the coefficients of the model. Which group is the reference level?
 - Reorder the factor levels so that "psychotic" is the reference level, and refit the model.
 3. Create a dummy variable called `is_nonpsychotic` that is 1 if the patient was classified as nonpsychotic, and 0 otherwise.
 - Use the `ifelse()` function to create this variable.
 - Fit a linear model to predict DBH activity using `is_nonpsychotic` as a predictor.
 - Interpret the intercept and slope of the model.
 - Use the `aggregate()` function to compute the mean DBH activity for each group. Compare these means to the coefficients from your models.
 4. Visualize the data using boxplots and histograms.

¹ Data source: Sternberg, D. E., VanKammen, D. P., Lerner, P., & Bunney, W. E. (1982). 'Schizophrenia: dopamine beta-hydroxylase activity and treatment response.' *Science*, 216 (4553), 1423–1425. <https://doi.org/10.1126/science.6124036>.

- Create a boxplot of DBH activity for each group. Use different colours to distinguish between groups, and add appropriate axis labels and a title.
 - Create histograms of DBH activity separately for each group. Compare the distribution shapes.
5. Using the data from the *Tips from the Top* study, reproduce the ‘Histogram of `pdiffmean` by Condition’ plot from Section 4.2.3 (‘Plotting data with a categorical predictor’).
- Add the argument `freq = FALSE` to both calls to the `hist()` function.
 - Explain why the plot looks different when this argument is included. What does setting `freq = FALSE` do?
6. Load the dataset `Early.csv`, which contains data from a study on early childhood intervention and cognitive development in infants. The study followed 103 infants from low-income families, randomly assigning them to either a treatment group (58 infants) or a control group (45 infants). Starting at 0.5 years of age, infants in the treatment group were exposed to an enriched early learning environment. Each child’s *cognitive score* was measured at ages 1, 1.5 and 2 years using an age-specific, normalized scale.

The dataset includes the following variables:

- `id`: an identifier for each infant
 - `cog`: the cognitive score
 - `age`: the age at which the measurement was taken (1, 1.5, or 2)
 - `trt`: a group indicator (‘Y’ for treatment, ‘N’ for control)
 - Use the `read.csv()` function to load the dataset into R.
 - Use the `str()` function to examine the structure of the dataset.
 - Convert the `trt` variable to a factor if it is not already.
 - Use the `levels()` function to check the levels of the treatment factor.
7. Using the `Early` dataset, perform analyses to compare the cognitive scores of children in the treatment and control groups at different ages.
- Use the `aggregate()` function to compute the mean cognitive score (`cog`) for each group (`trt`) at each age.
 - Fit three separate linear models to predict cognitive scores (`cog`) at each age (1, 1.5 and 2), using `trt` as the predictor.
 - Interpret the coefficients of each model. Which group shows higher scores at each age? Are the differences large or small?
 - Visualize the data using boxplots or scatterplots to show the distribution of scores by group at each age. Consider plotting all three age groups side by side for comparison.
 - Based on the model outputs and the visualizations, discuss any trends you observe. Does the treatment group appear to benefit more as age increases?