Andrew Gard - equitable.equations@gmail.com

Including Categorical Variables in a Linear Regression Model

The simplest way to introduce categorical variables into a linear model is by using *dummy variables*.

The simplest way to introduce categorical variables into a linear model is by using *dummy variables*.

A *dummy variable* encodes whether or not an observation falls into a specific category or not. In the *crickets* data set, for instance we can set

$$x_2 = \begin{cases} 0 & \text{for O. exclamationis} \\ 1 & \text{for O. niveus} \end{cases}$$

Once a variable like *species* is coded this way, we can include it in a linear model as if it were quantitative.

Once a variable like *species* is coded this way, we can include it in a linear model as if it were quantitative. For instance, if x_1 represents *temp* then we can write

$$y \sim b_0 + b_1 x_1 + b_2 x_2.$$

Once a variable like *species* is coded this way, we can include it in a linear model as if it were quantitative. For instance, if x_1 represents *temp* then we can write

$$y \sim b_0 + b_1 x_1 + b_2 x_2.$$

This is sometimes called an **additive** or **parallel slopes model**. The variables x_1 (temp) and x_2 (species) both affect the response variable y (rate), but they do not interact with one another.

Once a variable like *species* is coded this way, we can include it in a linear model as if it were quantitative. For instance, if x_1 represents *temp* then we can write

$$y \sim b_0 + b_1 x_1 + b_2 x_2.$$

This is sometimes called an **additive** or **parallel slopes model**. The variables x_1 (temp) and x_2 (species) both affect the response variable y (rate), but they do not interact with one another.

This model could also be written,

$$y \sim \begin{cases} b_0 + b_1 x_1 & \text{for O. exclamationis} \\ (b_0 + b_2) + b_1 x_1 & \text{for O. niveus} \end{cases}$$

Here the intercepts differ, but the slopes do not.

```
##
## Call:
## lm(formula = rate ~ temp + species, data = crickets)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0128 -1.1296 -0.3912  0.9650  3.7800
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.21091    2.55094  -2.827  0.00858 **
## temp               3.60275    0.09729  37.032  < 2e-16 ***
## speciesO. niveus -10.06529    0.73526 -13.689 6.27e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This *R* output corresponds to the model

$$y \sim \begin{cases} -7.21091 + 3.60275x_1 & \text{for O. exclamationis} \\ (-7.21091 - 10.06529) + 3.60275x_1 & \text{for O. niveus} \end{cases}$$

To allow for different slopes, we add an **interaction term** to our linear model.

$$y \sim b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2.$$

To allow for different slopes, we add an **interaction term** to our linear model.

$$y \sim b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2.$$

This can also be written,

$$y \sim \begin{cases} b_0 + b_1 x_1 & \text{for O. exclamationis} \\ (b_0 + b_2) + (b_1 + b_3) x_1 & \text{for O. niveus} \end{cases}$$

where, as before, x_1 is *temp* and x_2 is a dummy variable for *species* given by

$$x_2 = \begin{cases} 0 & \text{for O. exclamationis} \\ 1 & \text{for O. niveus} \end{cases}$$

```
## Call:
## lm(formula = rate ~ temp * species, data = crickets)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7031 -1.3417 -0.1235  0.8100  3.6330
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -11.0408     4.1515  -2.659    0.013 *
## temp                    3.7514     0.1601  23.429   <2e-16 ***
## speciesO. niveus       -4.3484     4.9617  -0.876    0.389
## temp:speciesO. niveus  -0.2340     0.2009  -1.165    0.254
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This *R* output corresponds to the model

$$y \sim \begin{cases} -11.0408 + 3.7514x_1 & \text{for O. exclamationis} \\ (-11.0408 - 4.3484) + (3.7514 - 0.2340)x_1 & \text{for O. niveus} \end{cases}$$