Andrew Gard - equitable.equations@gmail.com

Including Variable Interaction in a Linear Regression Model

Suppose we have two explanatory variables, x_1 and x_2, the first quantitative and the second categorical with two levels. For instance, in the present example,

$$x_2 = \begin{cases} 0 & \text{for Chinstrap} \\ 1 & \text{for Gentoo} \end{cases}$$

Suppose we have two explanatory variables, x_1 and x_2, the first quantitative and the second categorical with two levels. For instance, in the present example,

$$x_2 = \begin{cases} 0 & \text{for Chinstrap} \\ 1 & \text{for Gentoo} \end{cases}$$

To include interaction between the variables in a linear regression, we add a new term to our parallel slopes model.

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

Suppose we have two explanatory variables, x_1 and x_2, the first quantitative and the second categorical with two levels. For instance, in the present example,

$$x_2 = \begin{cases} 0 & \text{for Chinstrap} \\ 1 & \text{for Gentoo} \end{cases}$$

To include interaction between the variables in a linear regression, we add a new term to our parallel slopes model.

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

The word *linear* refers to the parameters, not the variables themselves

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

This can also be written,

$$y \sim \begin{cases} b_0 + b_1 x_1 & \text{for Chinstrap} \\ (b_0 + b_2) + (b_1 + b_3) x_1 & \text{for Gentoo} \end{cases}$$

Remember that x_2 is a dummy variable for *species* given by

$$x_2 = \begin{cases} 0 & \text{for Chinstrap} \\ 1 & \text{for Gentoo} \end{cases}$$

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

This can also be written,

$$y \sim \begin{cases} b_0 + b_1 x_1 & \text{for Chinstrap} \\ (b_0 + b_2) + (b_1 + b_3) x_1 & \text{for Gentoo} \end{cases}$$

Remember that x_2 is a dummy variable for *species* given by

$$x_2 = \begin{cases} 0 & \text{for Chinstrap} \\ 1 & \text{for Gentoo} \end{cases}$$

So b_2 represents the difference in the intercepts for the two groups and b_3 represents the difference in slopes. The category for which $x_2 = 0$ is known as the *reference level*.

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm * species, data = penguins_sm)

Residuals:
    Min      1Q  Median      3Q     Max
-911.18 -215.38  -42.69  162.67 1015.71

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -3037.196   1138.679  -2.667  0.00832 **
flipper_length_mm                  34.573      5.811   5.950  1.3e-08 ***
speciesGentoo                   -3750.085   1534.769  -2.443  0.01548 *
flipper_length_mm:speciesGentoo    20.049      7.496   2.674  0.00815 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 339.2 on 187 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.8205,    Adjusted R-squared:  0.8176
F-statistic: 284.9 on 3 and 187 DF,  p-value: < 2.2e-16
```

This R output corresponds to the model

$$y \sim \begin{cases} -3037.196 + 34.573x_1 & \text{for Chinstrap} \\ (-3037.196 - 3750.085) + (34.573 + 20.049)x_1 & \text{for Gentoo} \end{cases}$$