Consider the categorical variable \textit{Class} representing the class in high school with categories of freshman, sophomore, junior, or senior.

I’ve mentioned that a categorical variable with 4 categories would require 3 dummy variables.

Why can’t I just use the following coding system with only 2 dummy variables?

<table>
<thead>
<tr>
<th>Category</th>
<th>$D_1$</th>
<th>$D_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshmen</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Sophomore</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Junior</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Senior</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
\( D_1 \) and \( D_2 \) do define 4 distinct categories.

Let’s look at the fitted models for each group if we also have 1 quantitative covariate \( X \):

\[
Y_i = \beta_0 + \beta_1 x_i + \beta_2 D_{1i} + \beta_3 D_{2i} + \epsilon_i
\]

Freshmen: \( Y_i = (\beta_0 + \beta_2 + \beta_3) + \beta_1 x_i + \epsilon_i \)

Sophomore: \( Y_i = (\beta_0 + \beta_2) + \beta_1 x_i + \epsilon_i \)

Junior: \( Y_i = (\beta_0 + \beta_3) + \beta_1 x_i + \epsilon_i \)

Senior: \( Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \)

Let’s assume all the parameters are not 0.

There are 4 distinct intercepts here, but they are not all ‘free’ to be any value....

There is a restriction such that the Freshmen intercept is dependent on the Sophomore intercept and the Junior intercept.
This would not fit the ‘best fitting’ intercept for each respective group, which is usually what we want.

One possible correct coding:

<table>
<thead>
<tr>
<th>Category</th>
<th>$D_1$</th>
<th>$D_2$</th>
<th>$D_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshmen</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sophomore</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Junior</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Senior</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The overall model:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 D_{1i} + \beta_3 D_{2i} + \beta_4 D_{3i} + \epsilon_i$$

The fitted model for each group:

Freshmen:  $$Y_i = (\beta_0 + \beta_2) + \beta_1 x_i + \epsilon_i$$

Sophomore: $$Y_i = (\beta_0 + \beta_3) + \beta_1 x_i + \epsilon_i$$

Junior:    $$Y_i = (\beta_0 + \beta_4) + \beta_1 x_i + \epsilon_i$$

Senior:    $$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
This coding would allow for a separately fit intercept for each group.

Use $m - 1$ dummy variables for a categorical variable with $m$ categories.

Recall that this model would fit 4 separate parallel lines because there is no interaction between Class and $X$ in the model.

A model that includes interaction would be:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 D_{1i} + \beta_3 D_{2i} + \beta_4 D_{3i} + \beta_5 D_{1i} x_i + \beta_6 D_{2i} x_i + \beta_7 D_{3i} x_i + \epsilon_i$$
The fitted model for each group (includes interaction):

Freshmen: \[ Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_5)x_i + \epsilon_i \]

Sophomore: \[ Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_6)x_i + \epsilon_i \]

Junior: \[ Y_i = (\beta_0 + \beta_4) + (\beta_1 + \beta_7)x_i + \epsilon_i \]

Senior: \[ Y_i = \beta_0 + \beta_1x_i + \epsilon_i \]

The model fits separate intercepts AND separate slopes.

To test if any of the interaction variables are significant...

\[ H_0 : \beta_5 = \beta_6 = \beta_7 = 0 \]
\[ H_A : \text{not } H_0 \]

this can be accomplished with a partial F-test.
Notes
Notes