Multiple regression with qualitative predictors

A qualitative is one that does not naturally correspond to a numerical scale.

Example:
- Gender: female or male
- Brand: Kellogg, Post, General Mills.
- Location: top shelf, display aisle, ends.
- Category: home-schooled, not home-schooled.

When qualitative predictors are included in regression models, they are sometimes assigned numerical values. For example:

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-6 yrs of education</td>
<td>1</td>
</tr>
<tr>
<td>7-12 yrs</td>
<td>2</td>
</tr>
<tr>
<td>13-17 yrs</td>
<td>3</td>
</tr>
<tr>
<td>18+ yrs</td>
<td>4</td>
</tr>
</tbody>
</table>

This is wrong. Why? Because we impose an arbitrary scale on a variable that has no scale. In example above, we implicitly establish an equal "distance" between the age categories. Why not assign values 0 to 5, 1, 2, 3 and 4?

Instead of using an arbitrary numerical scale, we use dummy variables to represent qualitative predictors.
For a qualitative predictor with \( m \) possible different categories, we define \( m-1 \) dummy variables.

**Example:** we wish to test whether there is an association between sales of a product and the location of the product in the supermarket and its price.

- \( y \) = \# of units sold in a month.
- \( x_1 \) = price of the product, \( \text{in } \$ \)
- \( x_2 \) = location of product in supermarket
  - special display
  - aisle end
  - middle of aisle, top shelf
  - middle of aisle, middle shelf

How do we formulate a model?

- \( x_1 \) is a quantitative predictor
- \( x_2 \) is qualitative, with 4 categories.

Define \( 4-1 = 3 \) dummies:

\[
\begin{align*}
   d_1 &= 1 \text{ if } x_2 = \text{ special display} \\
        &= 0 \text{ otherwise} \\
   d_2 &= 1 \text{ if } x_2 = \text{ aisle end} \\
        &= 0 \text{ otherwise} \\
   d_3 &= 1 \text{ if } x_2 = \text{ middle of aisle, middle shelf} \\
        &= 0 \text{ otherwise} \\
\end{align*}
\]

If \( x_2 = \text{ middle of aisle, middle shelf} \), then \( d_1 = d_2 = d_3 = 0 \).
Consider the model

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 d_1 + \beta_3 d_2 + \beta_4 d_3 + \varepsilon. \]

Note that:

- If \( x_2 = \) special display, then
  \[ E(y) = \beta_0 + \beta_1 x_1 + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x_1 \]

- If \( x_2 = \) aisle end, then
  \[ E(y) = \beta_0 + \beta_1 x_1 + \beta_3 = (\beta_0 + \beta_3) + \beta_1 x_1 \]

- If \( x_2 = \) middle aisle, top shelf, then
  \[ E(y) = \beta_0 + \beta_1 x_1 + \beta_4 = (\beta_0 + \beta_4) + \beta_1 x_1 \]

- If \( x_2 = \) middle aisle, middle shelf, then
  \[ E(y) = \beta_0 + \beta_1 x_1 \]

Suppose first \( \beta_1 < 0 \) and that \( \beta_2 > \beta_3 > \beta_4 > 0 \). Then the prediction equation looks like:
If the model has no interaction between the dummy and the quantitative variable, we say that the association between $x_1$ and $y$ does not depend on the level of the qualitative predictor.

Interpreting the regression coefficients: we test hypotheses on $\beta_2$, $\beta_3$ and $\beta_4$. If the test $H_0: \beta_j = 0$

$H_0: \beta_j \neq 0$

for $j = 2, 3, 4$

leads to rejection of $H_0$, we conclude that for any value of $x_1$, $E(y)$ is higher (lower) when $x_2$ is category $j$ than when $x_2$ is equal to the reference or base category (the one for which all the dummies are equal to 0).

Example: suppose that in example, we conclude that $\beta_2 \neq 0$, $\beta_3 \neq 0$, $\beta_4 = 0$. Then we say that for any price, sales of the product are higher when product is placed in a special display or aisle end than when
the product is placed in a middle aisle top or middle shelf.

When we include a qualitative predictor, we always define a "reference" category and then draw inference relative to that reference category.

Interaction between quantitative and qualitative predictors.

Consider now the model with 2 predictors:

\[ x_1 = \text{quantitative} \]
\[ x_2 = \text{qualitative with 4 categories} \]

We define 3 dummies as before. Including interactions:

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 d_1 + \beta_3 d_2 + \beta_4 d_3 + \beta_5 d_1 x_1 + \beta_6 d_2 x_1 + \beta_7 d_3 x_1 + \epsilon. \]

When \( x_2 = \text{special display} \)

\[ E(y) = \beta_0 + \beta_1 x_1 + \beta_2 + \beta_5 x_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_5) x_1 \]

When \( x_2 = \text{aisle end} \)

\[ E(y) = \beta_0 + \beta_1 x_1 + \beta_3 + \beta_6 x_1 = (\beta_0 + \beta_3) + (\beta_1 + \beta_6) x_1 \]
When $x_2 =$ middle aisle, middle shelf

$E(y) = \beta_0 + \beta_1 x_1$.

$\rightarrow$ Intercept and slope are allowed to change across the different categories of $x_2$.

From graph we know that:

$\beta_2 > \beta_4 > \beta_3 > 0 \quad \rightarrow \text{order intercepts}$

$\beta_7 < \beta_5 < \beta_6 < 0 \quad \rightarrow \text{order slopes}$.