Three qualitative independent variables

The idea of using dummy variables to model independent qualitative variables can be applied to situations where there are more than two factors.

Suppose that in the example of the performance related to fuels we consider an additional factor corresponding to season. We then have a four levels new factor and three corresponding dummy variables

\[ x_4 = \begin{cases} 
1 & \text{if Spring} \\
0 & \text{otherwise}
\end{cases}, \quad x_5 = \begin{cases} 
1 & \text{if Summer} \\
0 & \text{otherwise}
\end{cases}, \quad x_6 = \begin{cases} 
1 & \text{if Fall} \\
0 & \text{otherwise}
\end{cases} \]

note that Winter is used as baseline.

We now have three more terms describing the main effect of season, say

\[ \beta_4x_4 + \beta_5x_5 + \beta_6x_6 \]

as well as six terms describing the interactions between fuel type and season and three terms for the interactions between brand and season.

Given that we now have three factors we can also consider third order interactions. We obtain these by multiplying the dummy variables of each of the factors, say

\[ \beta_{16}x_1x_3x_4 + \beta_{17}x_2x_3x_4 + \ldots \]

Thus, in the fuel example we have six three way interaction terms.

Note that the number of parameters increases rapidly with the number of levels in each factor. So it is good practice to perform an initial experiment with only two levels per factor to screen possible significant main and interaction effects. These are known as $2^k$ designs, where $k$ is the number of factors.

An alternative parametrisation

An alternative way of parametrising a model for independent qualitative variables is obtained by considering a baseline given by an average level of the response.

Back to the fuel example with two factors we can write the model for the main effects of fuel type and brand as

\[ y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \]

\(i = 1, 2, 3; \ j = 1, 2; \ k = 1, \ldots, n_{ij}.\)

- \(y_{ijk}\) corresponds to the performance of fuel type \(i\), brand \(j\), \(k\)th replicate.
- \(\alpha_i\) is the main effect of fuel type \(i\)
- \(\beta_j\) is the main effect of fuel brand \(j\)

This is an overparametrised model, as can be seen from the fact that its design matrix has columns which are not linearly independent.

To obtain a unique estimate of the model parameters the restrictions

\[ \sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0 \]

are introduced.

Interactions are introduced in the model with additional parameters that depend on both indexes

\[ y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \]

with the restriction that

\[ \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0 \]

These models are usually known as two-way linear models.
Quantitative and qualitative covariates

The advantage of the dummy variable approach is that it is very natural to mix qualitative and quantitative independent variables in the same model.

Suppose that for the fuel example we now consider the engine speed as an explanatory variable of performance. A description of the data suggests that, in absence of difference between fuel type, brand or season, the mean response is

\[ E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 \]

where \( x_1 \) corresponds to engine speed.

Suppose that fuel type is now considered as a factor that changes the performance. If the different types of fuel affect the performance homogeneously for all values of the engine speed, then we will observe three parallel mean response curves corresponding to

\[ E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 \]

where \( x_2 \) and \( x_3 \) are the dummy variables corresponding the types of fuel.

Notice that this model is only changing the intercept from \( \beta_0 \) for fuel type B to \( \beta_0 + \beta_3 \) for fuel type P and \( \beta_0 + \beta_4 \) for fuel type C.

If we believe that the performance is affected by the fuel type in a way that depends on the engine speed then the three curves will not be parallel and we augment the model with interaction terms

\[ \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1^2 x_2 + \beta_8 x_1^2 x_3 \]

so that the mean response could be differently shifted and curved depending on fuel type.