5.2 Coding Categorical Variables

Suppose we model the response Y using two predictors X and D, where X is a numerical variable and D is categorical with k levels, such as `student status` undergraduate vs. graduate (2 levels), or `political affliation` democrat, republican vs. independent (3 levels) or in the Birthweight example `smoker status` smoker vs. non-smoker (2 levels). To incorporate categorical variable in the regression model, we need to code them using **indicator variables** (or otherwise called *dummy variables*).

Birthweight Example

We consider two predictors to describe the response `Birthweight` : `Head Circumference` and `Smoker` , where the first one is a continuous variable that we denote by X and the second one is a categorical with two levels that we denote by D.

In the case of a categorical variable D with *two levels*, similar to the `smoker` variable in the `Birthweight` example, assume that we use ***two*** indicator variables to describe D as follows:

$$d_2 = \begin{cases} 1, \text{ if in level 1 - e.g. mother is a smoker} \\ 0, \text{ otherwise} \end{cases}, \; d_3 = \begin{cases} 1, \text{ if in level 2 - e.g.mother is a n} \\ 0, \text{ otherwise} \end{cases}$$

In this case, a first order model (i.e. a model with no interaction terms) that includes `Head Circumference` and `Smoker` would look like

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_{i2} + \beta_3 d_{i3} + \varepsilon_i$$

Unfortunately, this intuitive approach for introducing one dummy variable for each level leads to computational challenges. To illustrate this point, consider the *design matrix* with $n = 4$[17]:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & 1 & 0 \\ 1 & x_2 & 1 & 0 \\ 1 & x_3 & 0 & 1 \\ 1 & x_4 & 0 & 1. \end{pmatrix}$$

The first column corresponds to the intercept, the second column corresponds to the continuous variable `Head Circumference` and the third and fourth columns correspond to the indicator variables d_2 and d_3 (respectively) we created above. As we can easily check, the *sum of the last two columns is equal to the column of 1s*. This means that the these columns are **linearly dependent**, which implies that the $\mathbf{X}^T\mathbf{X}$ matrix is singular, cannot be inverted and no unique estimators of the regression coefficients can be found.

One simple way to overcome this difficulty is to *drop* one of the indicator variables, e.g. we drop d_3. So, from now on, even when the number of levels of the categorical predictor is more than two, we follow the principle:

Coding Categorical Variables Principle

A qualitative variable with k levels (classes) will be represented by $k - 1$ indicator (dummy) variables, each taking the value 0 or 1.

Remark: You can code the two levels using *any* two different values, which will not change $\hat{y}$, but only the interpretation of the estimated coefficients.