Indicator Variables (or, dummy variables; these typically code for group membership)

We begin with a distinction between quantitative and qualitative variables:

Quantitative – the numbers are assumed to represent magnitudes of some quantity

Qualitative – the numbers are assumed to be labels, i.e., the categorical or nominal level of measurement

The question: how can we incorporate categorical variables into multiple regression

Suppose I have a categorical variable $X$ that I would like to use in explaining some quantitative variable $Y$: 
\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \] where \( X_i = 1 \) when \( i \) is from class 1; and \( X_i = 0 \) when \( i \) is from class 2

\( X \) is the dummy variable indicating group (class) membership

Thus, if \( X_i = 1 \), then \( Y_i = \beta_0 + \beta_1 + \epsilon_i \);

if \( X_i = 0 \), then \( Y_i = \beta_0 + \epsilon_i \)

We can carry out the least-squares fit and get \( b_0 \) and \( b_1 \)

Now, what do you think these estimates turn out to be?
Thus, a test of $H_0 : \beta_1 = 0$ is the same as as a test of $H_0 : \mu_1 = \mu_2$

Do we have a procedure?
Remember the $t$-test for two independent samples:

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\hat{\sigma}^2 \left( \frac{n_1 + n_2}{n_1 n_2} \right)}} \sim t_{n_1 + n_2 - 2}$$

where $\hat{\sigma}^2$ is the pooled error for two groups.

The test ratio for $H_0 : \beta_1 = 0$ has the form

$$\frac{b_1}{\sqrt{s^2(b_1)}} \sim t_{n-2}$$

where $n = n_1 + n_2$ and

$$s^2(b_1) = \frac{MSE}{\sum(X_i - \bar{X})^2} = MSE \left( \frac{n_1 + n_2}{n_1 n_2} \right)$$
Now, suppose I have 3 groups:

Let $X_{i1} = 1$ if $i$ is in group 1; let $X_{i2} = 1$ if $i$ is in group 2.

Then for the model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$, we have the following chart:

<table>
<thead>
<tr>
<th>$X_{i1}$</th>
<th>$X_{i2}$</th>
<th>$X_{i1}$</th>
<th>$X_{i2}$</th>
<th>$X_{i1}$</th>
<th>$X_{i2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_i$</td>
<td>$\beta_0 + \beta_1 + \epsilon_i$</td>
<td>$\beta_0 + \beta_2 + \epsilon_i$</td>
<td>$\beta_0 + \epsilon_i$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{Y}_i$</td>
<td>$b_0 + b_1$</td>
<td>$b_0 + b_2$</td>
<td>$b_0$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Thus, $b_0 = \bar{Y}_3$; $b_1 = \bar{Y}_1 - \bar{Y}_3$; $b_2 = \bar{Y}_2 - \bar{Y}_3$

and

$\beta_0 = \mu_3$; $\beta_1 = \mu_1 - \mu_3$; $\beta_2 = \mu_2 - \mu_3$
So, \( H_o : \beta_1 = 0, \beta_2 = 0 \) can be tested in the usual way with

\[
\frac{MSR}{MSE} \sim F_{2,n-3};
\]
here \( p-1 = 2 \) and is the number of groups minus one; \( n-p = n-3 \) and is \( n \) minus the number of groups.

This is the same as a one-way analysis-of-variance with 3 groups since \( H_o : \beta_1 = 0, \beta_2 = 0 \) implies

\[
H_o : \mu_1 - \mu_2 = 0, \mu_2 - \mu_3 = 0, \text{ and in turn,}
\]

\[
H_o : \mu_1 = \mu_2 = \mu_3
\]

This can be extended to any number of groups.
Suppose I have two factors (factor 1 and factor 2); factor 1 has two levels of a and b (e.g., male and female); factor 2 has two levels of c and d (two difficulties of a test)

let $X_{i1} = 1$ if $i$ is in the a level on factor 1 and 0 otherwise;

let $X_{i2} = 1$ if $i$ is in the c level on factor 2 and 0 otherwise;

thus, $X_{i1}X_{i2} (\equiv X_{i3}) = 1$ is $i$ is in the a level on factor 1 and the c level on factor 2

Consider the model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$
The following table gives $E(Y_i)$ under various combinations of the two factors:

<table>
<thead>
<tr>
<th>Factor 1</th>
<th>Factor 2</th>
<th>$E(Y_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>c</td>
<td>$\beta_0 + \beta_1 + \beta_2 + \beta_3$</td>
</tr>
<tr>
<td>a</td>
<td>d</td>
<td>$\beta_0 + \beta_1$</td>
</tr>
<tr>
<td>b</td>
<td>c</td>
<td>$\beta_0 + \beta_2$</td>
</tr>
<tr>
<td>b</td>
<td>d</td>
<td>$\beta_0$</td>
</tr>
</tbody>
</table>

$H_o : \beta_1 = 0$ is the main effect test for Factor 1

$H_o : \beta_2 = 0$ is the main effect test for Factor 2

$H_o : \beta_3 = 0$ is the test for interaction between Factors 1 and 2

This can all be extended to more than two levels on each factor, and to more than 2 factors – also, a quantitative variable could be incorporated as well

If the cell sizes are equal, the independent dummy variables are uncorrelated and the design is said to be “orthogonal”
Now, suppose I have one quantitative independent variable, $X_1$, and a dummy variable $X_2$, where $X_{i2} = 1$ if $i$ is in class 1 and equal to 0 if in class 2

Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$

So, for group 1: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 + \epsilon_i = Y_i = (\beta_0 + \beta_2) + \beta_1 X_{i1} + \epsilon_i$

for group 2: $Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$

Assuming the slopes within groups are the same, a test of $H_o : \beta_2 = 0$ is an attempt to test whether the intercepts are also the same.
Or, is there a group difference if I include variable $X_1$

The is called “analysis-of-covariance”; it can extended to more than two groups by test-ing the regression coefficients that are on the dummy variables as a group.

What if the slopes within groups are not the same:

$$Y_i = \beta_0 + \beta_1(X_{i1}X_{i2}) + \beta_2X_{i1} + \beta_3X_{i2} + \epsilon_i$$

For group 1:

$$Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_2)X_{i1} + \epsilon_i$$

For group 2:

$$Y_i = \beta_0 + \beta_2X_{i1} + \epsilon_i$$
Thus, to test the hypothesis of “same slopes”, test $H_0 : \beta_1 = 0$

To test the hypothesis of “same intercepts”, test $H_0 : \beta_3 = 0$

To test the hypothesis of “same regressions”, test $H_0 : \beta_1 = 0, \beta_3 = 0$
What to do when the dependent variable is binary –

First, the usual assumptions “go to hell”: \( Y \) can’t be normal but must be, say, Bernoulli; also, the variance of \( Y \) will depend on \( X \)

We could approach this with Logistic Regression or through the use of weighted least-squares; there is another way to view this that we will follow — through the use of Fisher’s Linear Discriminant analysis

This is developed in great detail in any Multivariate Analysis course; it is also the cornerstone of some statistical approaches to “Big Data”

We begin by assuming that \( Y \) is binary and defines two groups: \( Y \) is 0 if the observation is in Group I; \( Y \) is 1 if the observations is in Group II
Suppose I get \( \hat{Y} = b_0 + b_1 X_{i1} + \cdots + b_{p-1} X_{i(p-1)} \)

If I put in the means on the independent variables for group I and II, I get \( \hat{Y}_I \) and \( \hat{Y}_{II} \) (assume without loss of generality that \( \hat{Y}_I \leq \hat{Y}_{II} \), or we could interchange the group designations)

I will view the independent variables as random; I’m interested in classifying a new observation into I or II as follows:

Obtain \( \hat{Y}_{new} \) and classify into II if \( \hat{Y}_{new} \) is greater than \( C \) (yet to be found) and into I if \( \hat{Y}_{new} \) is less than or equal to \( C \)
If the a priori probabilities of group membership are equal, then $C = (\hat{Y}_I + \hat{Y}_{II})/2$ gives the minimum for the probability of misclassification.

This assumes multivariate normality and the population.

To evaluate the actual rule, we can look at the misclassification table:

<table>
<thead>
<tr>
<th>Group Membership</th>
<th>I</th>
<th>II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>II</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

where $n = a + b + c + d$
$\frac{a+d}{n}$ is the percentage of correct classifications

We can also get a similar table using a sample reuse method since $\frac{a+d}{n}$ is inflated (i.e., we need cross-validation)

This is called Fisher’s Linear Discriminant Function

It has the property of maximizing the $t^2$ value over all linear combinations of the independent variables