8 Factors and Dummy Variables in Regression

8.1 The Simplest Possible Case

Suppose that for each unit in a sample we measure \( Y \) (say weight) and a factor \( x \) indicating membership in one of two groups (say male or female). If we code \( x \) as

\[
x = \begin{cases} 
0, & \text{if in group 1} \\
1, & \text{if in group 2}
\end{cases}
\]

and then regress \( Y \) on \( x \), i.e., fit the model

\[ Y = \beta_0 + \beta_1 x + \epsilon, \]

then it turns out that

\[ \hat{\beta}_0 = \bar{y}_1 \]  
(group 1 sample mean)

\[ \hat{\beta}_1 = \bar{y}_2 - \bar{y}_1 \]  
(difference in group means)

Thus our estimate for the mean of \( y \) is

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times 0 = \bar{y}_1 \]

for those in group 1 (\( x = 0 \))

and

\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times 1 = \bar{y}_1 + (\bar{y}_2 - \bar{y}_1) = \bar{y}_2 \]

for those in group 2 (\( x = 1 \))

which is comforting.

Note also that

\[
\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2 = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2
\]

so that

\[
\hat{\sigma}^2 = \frac{\text{SSE}}{n - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = s_p^2
\]

The \( t \) test of \( H_0 : \beta_1 = 0 \) is a test for equality of means for the two groups. Recall that the estimated standard error of \( \hat{\beta}_1 \) is

\[ \hat{\sigma} \sqrt{\frac{T}{S_{xx}}}. \]
In the present case,\[
\bar{x} = \frac{n_2}{n_1 + n_2}
\]
and
\[
S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = n_1 \left(0 - \frac{n_2}{n_1 + n_2}\right)^2 + n_2 \left(1 - \frac{n_2}{n_1 + n_2}\right)^2 = \frac{n_1 n_2}{n_1 + n_2}
\]
Thus
\[
\frac{1}{S_{xx}} = \frac{n_1 + n_2}{n_1 n_2} = \frac{1}{n_1} + \frac{1}{n_2}
\]
and the test statistic is
\[
t = \frac{\hat{\beta}_1}{\hat{\sigma} \sqrt{\frac{1}{S_{xx}}}} = \frac{\bar{y}_2 - \bar{y}_1}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]
the traditional two-sample $t$ statistic.

See R transcript for an example.

8.2 A Single Factor: One Way ANOVA

A factor is a categorical variable used as a predictor. Generally even a single factor will have more than two levels. Suppose for example that the factor $A$ has three levels $A_1$, $A_2$, and $A_3$. To represent the factor in a regression model we define “dummy variables” (or “indicator variables”)

\[
d_1 = \begin{cases} 
1, & \text{if observation has level } A_2 \text{ of factor } A \\
0, & \text{otherwise}
\end{cases}
\]

and

\[
d_2 = \begin{cases} 
1, & \text{if observation has level } A_3 \text{ of factor } A \\
0, & \text{otherwise}
\end{cases}
\]

If the factor $A$ is our only predictor, then the appropriate linear regression model is

\[
Y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \varepsilon.
\]

The following table gives the values of $d_1$ and $d_2$ and the mean of $Y$ in this model for each level of the factor $A$: 

\[
\begin{array}{c|c|c}
\hline
\text{Level of } A & d_1 & d_2 & \bar{Y} \\
\hline
A_1 & 0 & 0 & \bar{Y}_1 \\
A_2 & 1 & 0 & \bar{Y}_2 \\
A_3 & 0 & 1 & \bar{Y}_3 \\
\hline
\end{array}
\]
<table>
<thead>
<tr>
<th>Level of Factor A</th>
<th>d₁</th>
<th>d₂</th>
<th>E(Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A₁</td>
<td>0</td>
<td>0</td>
<td>β₀</td>
</tr>
<tr>
<td>A₂</td>
<td>1</td>
<td>0</td>
<td>β₀ + β₁</td>
</tr>
<tr>
<td>A₃</td>
<td>0</td>
<td>1</td>
<td>β₀ + β₂</td>
</tr>
</tbody>
</table>

To reiterate,
- β₀ is the mean response when A has value A₁.
- β₁ is the difference in mean response between level A₂ and level A₁ of A.
- β₂ is the difference in mean response between levels A₃ and A₁.

Thus in this coding of the factor, level A₁ is treated as a baseline, and the other levels are contrasted with this one.

If μ₁, μ₂, and μ₃ represents the mean of Y for each of the three factor levels, then so-called one-way analysis of variance (one-way ANOVA) tests:

\[ H_0 : \mu_1 = \mu_2 = \mu_3 \text{ vs } H_a : \text{at least two of the } \mu_k \text{ s differ} \]

This is equivalent to testing:

\[ H_0 : \beta_1 = \beta_2 = 0 \text{ vs } H_a : \text{at least one of } \beta_1 \text{ or } \beta_2 \text{ is nonzero} \]

in our regression model:

\[ Y = \beta_0 + \beta_1 d_1 + \beta_2 d_2 + \varepsilon. \]

In other words, we are comparing this full model to the reduced model:

\[ Y = \beta_0 + \varepsilon, \]

and we use the usual F-test.

Note:
- In general, the number of dummy variables required to code a factor into a regression model is one fewer than the number of levels of the factor.
- The default in R is to code factors in the way we have above, so that the “first” level of a factor is treated as a baseline, and the other levels are contrasted with it. R calls this kind of coding “treatment contrasts”.
- There are many other ways to code factors, and this can lead to much confusion. On the whole, I think the “treatment contrasts” used by R are generally the simplest to deal with.

See R transcript.