Dummy Variables
Vartanian

In regression analysis, the independent variables must be either interval scale variables or dummy variables. Dummy variables are variables that have values of either 0 or 1. What dummy variables allow us to do is compare the mean value for the dependent variable between members of two groups. For example, we may be interested in whether single adult households receive more concrete services than non-single adult households. We can use a dummy variable approach within an ordinary least squares regression analysis to determine the difference in the mean number of children between members of the two groups. In this example, whether the family is a single adult household is the independent variable and the number of children is the dependent variable. Another example is examining whether or not females have a higher level of income than do males. In this second example, gender would be the independent variable and income would be the dependent variable.

How is this done?

Let's say that we're examining the relationship between gender and income. We hypothesize that females make more money than males.

What we would do is the following:

\[ Y = \text{income} \]

Our independent variable is going to be gender:

**Gender has 2 values:**
When gender=1 then the person is a male.
When gender=2 then the person is a female.

We will make a dichotomous variable called female.

If (gender=2) female=1.
If (gender=1) female=0.

Our regression equation looks like the following:

\[ Y_p = a + b_1 \text{female} \]

What we are doing in this analysis is comparing the income of females with the excluded category, males. What we do when using dummy variables is exclude one category from our analysis, and the regression model automatically compares what is in the model to what is left out of the model. (Generally, the sample coefficient is expressed as b. However, SPSS indicates the B coefficient. I will
use both $b$ and $B$ to indicate the sample coefficient estimate. This $b$ coefficient is also referred to as the parameter estimate.)

We might then get the following results if we ran this model:

<table>
<thead>
<tr>
<th>Variable</th>
<th>B coefficient or Parameter estimate</th>
<th>SE B</th>
<th>t value</th>
<th>Signif</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>400</td>
<td>200</td>
<td>400/200=2.0</td>
<td>.05</td>
</tr>
</tbody>
</table>

What these results indicate is that females make $400 more than males. We see that this is a significant (at the .05 level) difference in income. The constant indicates the value for the intercept or the $a$ coefficient. The $t$ value is determined by dividing the $B$ coefficient or the parameter estimate by the standard error for $B$.

Now, let's say that we're examining the effects of marital status on income:

There are 3 different types of marital statuses:

- MS=1 if you've never been married
- MS=2 if you're married.
- MS=3 if you're divorced or separate (I'll refer to this category as divorced)

Since we have 3 categories for this nominal scale variable, you will use two dummy variables in the regression model. These two included variables in the model will be compared to the third and excluded variable. To make dummy variables in SPSS, do the following:

```
compute single=0.
compute married=0.
compute divorce=0.
if (ms=1) single=1.
if (ms=2) married=1.
if (ms=3) divorce=1.
```

The regression model looks like the following:

$$Y_p = a + b_1 \text{single} + b_2 \text{married}.$$
Let's assume that Y, the dependent variable, is income.

Here, we will get 2 B coefficients. Both of these coefficients are in relation to those who are divorced. That is, the B coefficients for married and single will indicate the difference in income between married people and divorced people (B₂) and between single people and divorced people (B₁).

These B coefficients for single and married will not tell us what the slope is between marital status and income, but will instead tell us what the income difference is between say married people and divorced people.

For example:

Let's say we run this model and come up with the following coefficients:

\[ N=500 \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>B coefficient</th>
<th>SE B</th>
<th>t value</th>
<th>Signif</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single</td>
<td>-400</td>
<td>150</td>
<td>-2.667</td>
<td>.01</td>
</tr>
<tr>
<td>Married</td>
<td>600</td>
<td>200</td>
<td>3</td>
<td>.01</td>
</tr>
</tbody>
</table>

In other words we have the following equation:

\[ Y_p = 100+(-400)\times\text{single}+(600)\times\text{Married} \]

This is telling us that single people have an income level that is $400 less than divorced people. The model is also telling us that married people have an income level that is $600 greater than divorced people. Each of these two coefficients is significantly different from 0 (both are significant at the .01 level). The constant indicates the value of the intercept or the a coefficient.

Now let's say we run the following model, examining the difference in income levels between males and females, between those people who are married versus those who are divorce or single versus those who are divorced.

\[ Y_p=a+b_1\times\text{single}+b_2\times\text{married}+b_3\times\text{female}+b_4\times\text{experience} \]

Three of the variables in our model are nominal scale variables while the other independent variable is a ratio scale variable. How do we interpret this?

The b₁ coefficient will tell us the difference in income between
never married people and divorced people, controlling for the other variables in the model.

$b_2$ will tell us the difference between married people and divorced people, controlling for the other variables within the model.

$b_3$ will tell us the difference in income between females and males, controlling for the other variables within the model.

$b_4$ will be a slope coefficient, which will indicate how much income rises and falls when there is a change in experience.

Example:

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SE b</th>
<th>t</th>
<th>Signif t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>-500</td>
<td>200</td>
<td>-2.5</td>
<td>.01</td>
</tr>
<tr>
<td>Married</td>
<td>300</td>
<td>200</td>
<td>1.5</td>
<td>.15</td>
</tr>
<tr>
<td>Female</td>
<td>300</td>
<td>100</td>
<td>3.0</td>
<td>.01</td>
</tr>
<tr>
<td>Experience</td>
<td>350</td>
<td>100</td>
<td>3.5</td>
<td>.001</td>
</tr>
<tr>
<td>Constant</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This model is telling us that the income level of single people is $500 less than for divorce people. The income level of married people is $300 more than for divorced people. The income of females is $300 more than for males. The $b$ coefficient for experience indicates that for every year of experience, income increases by $350. The constant indicates the value of the intercept or a coefficient.

If we were to use an equation to examine this, we would have

$$Y_p = 100 + (-500) * \text{single} + (300) * \text{Married} + (300) * \text{female} + (350) * \text{experience}$$

We could now predict the level of income for say some individual who was single, female and had 10 years of experience. Since the person is single, the value for single=1. If the person is single, it means that the person is not married, so the value for married is 0. Since the person is female the value for female=1. And since the person has 10 years of experience, the value for experience=10. Our prediction for the level of income for this individual is therefore

$$Y_p = 100 + (-500) * 1 + (300) * 0 + (300) * 1 + (350) * 10 =$$

$$Y_p = 100 - 500 + 0 + 300 + 3500 =$$
\[ Y_p = 3,400 \]