[STAT-21] Regression with dummy variables

Miguel-Angel Canela
Associate Professor, IESE Business School

Regression with dummy variables

Dummy variables have already appeared in some examples, but I never discussed the interpretation of the coefficients of these variables. Let me start with the simplest case, the equation $Y = \beta_0 + \beta_1 X + \epsilon$, in which X is a dummy used to code two groups, which I call group 0 and group 1.

By replacing $X = 1$ in the equation, we see that, in group 1, $Y \sim \mathcal{N}(\beta_0 + \beta_1, \sigma^2)$. Putting $X = 0$, we get $Y \sim \mathcal{N}(\beta_0, \sigma^2)$ in group 0. So, the null $\beta_1 =$ is the same as the equality of means of the two-sample t test. Indeed, the t statistic associated to β_1 is the same as that used to test the equality of means assuming equal variances. The details are given below.

What if the equation includes other variables $X_2, \ldots, X_k$? Then, β_1 is interpreted as the average change in Y when switching from group 0 to group 1, holding $X_2, \ldots, X_k$ constant.

The two-sample t test as a linear regression analysis

Let me take a sample with two groups, of sizes n_0 and n_1, coded with a dummy X. I denote by $y_{1,i}$ the values of Y in group 1, by $y_{0,i}$ those in group 0, and by $\bar{y}_1$ and $\bar{y}_0$ the group means. In the identity

$$\begin{bmatrix} y_{0,1} \\ \vdots \\ y_{0,n_0} \\ y_{1,1} \\ \vdots \\ y_{1,n_1} \end{bmatrix} = \bar{y}_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \big(\bar{y}_1 - \bar{y}_0\big) \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} y_{0,1} - \bar{y}_0 \\ \vdots \\ y_{0,n_0} - \bar{y}_0 \\ y_{1,1} - \bar{y}_1 \\ \vdots \\ y_{1,n_1} - \bar{y}_1 \end{bmatrix},$$

the third vector on the right side is orthogonal to the other two. So, it can be read in terms of the regression of Y on X. The third vector is the residual vector, $\bar{y}_0$ is the intercept, and $\bar{y}_1 - \bar{y}_0$ is the slope.

Let me see how to calculate the standard error for the slope. For this particular case, the formula of the error variance estimator of lecture STAT-17 gives

$$\hat{\sigma}^2 = \frac{\displaystyle\sum_{i=1}^{n_0} \big(y_{0,i} - \bar{y}_0\big)^2 + \sum_{i=n_1+1}^{n_0+n_1} \big(y_{1,i} - \bar{y}_1\big)^2}{n_0 + n_1 - 2} = \frac{\big(n_0 - 1\big)s_0^2 + \big(n_1 - 1\big)s_1^2}{n_0 + n_1 - 2},$$

which is the same as the pooled variance of the 2-sample t test (with equal variances). The second element in the calculation of the standard error of the slope is the total sum of squares of X. Since the mean of X is

$$\bar{x} = \frac{n_1}{n_0 + n_1},$$

we get

$$\sum_{i=0}^{n_0+n_1} \big(x_i - \bar{x}\big)^2 = n_0 \left(\frac{-n_1}{n_0 + n_1}\right)^2 + n_1 \left(\frac{n_0}{n_0 + n_1}\right)^2 = \frac{n_0\, n_1}{n_0 + n_1} = \left(\frac{1}{n_0} + \frac{1}{n_1}\right)^{-1}.$$

The t statistic associated to the slope is, then,

$$t = \frac{\hat{\beta}_1}{\hat{\mathrm{se}}[\hat{\beta}_1]} = \frac{\big(\bar{y}_1 - \bar{y}_0\big)}{s\,\sqrt{(1/n_0) + (1/n_1)}}\,,$$

which is the same as in the two-sample t-test (equal variances).

Coding groups with dummies

Consider now a sample with more than two groups. For instance, a sample of executives can be classified according to marital status, as single, married or divorced. These three groups can be coded with a dummy X_1 for being married, and a dummy X_2 for being divorced. The single group is coded as $X_1 = X_2 = 0$, the married group as $X_1 = 1$, $X_2 = 0$, and the divorced group as $X_1 = 0$, $X_2 = 1$.

In general, a factor with $k + 1$ groups is coded with k dummies X_1, ..., X_k. In a regression equation

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon,$$

the coefficients are interpreted as:

- In group 0, the mean is β_0.
- In group 1, the mean is $\beta_0 + \beta_1$.
- An so on, until group k, where the mean is $\beta_0 + \beta_k$.

So, for $i = 1, \ldots, k$, the null $\beta_i = 0$ is equivalent to the equality of the means of group 0 and group i. The F statistic for the null $\beta_1 = \ldots = \beta_k = 0$ is the same as in one-way ANOVA (see next section). When there are additional independent variables, β_1, ..., β_k are still interpreted as mean differences between groups, but holding the additional variables constant.

The one-way ANOVA test as a linear regression analysis

In the general case, with k dummies, we have an orthogonal decomposition

$$\begin{bmatrix} y_{0,1} \\ \vdots \\ y_{0,n_0} \\ y_{1,1} \\ \vdots \\ y_{1,n_1} \\ \vdots \\ \vdots \\ y_{k,1} \\ \vdots \\ y_{k,n_k} \end{bmatrix} = \bar{y}_0 \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \big(\bar{y}_1 - \bar{y}_0\big) \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \cdots + \big(\bar{y}_k - \bar{y}_0\big) \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} y_{0,1} - \bar{y}_0 \\ \vdots \\ y_{0,n_0} - \bar{y}_0 \\ y_{1,1} - \bar{y}_1 \\ \vdots \\ y_{1,n_1} - \bar{y}_1 \\ \vdots \\ \vdots \\ y_{k,1} - \bar{y}_k \\ \vdots \\ y_{k,n_k} - \bar{y}_k \end{bmatrix},$$

which shows that testing the regression coefficients in this context is the same as testing the differences $\bar{y}_j - \bar{y}$. Note that the residuals are the same as the one-way ANOVA residuals.

Instead of testing the coefficients separately, which always involves a choice of the zero group, we may want to test if there is *any* significant difference between group means, i.e. among the

k coefficients. This is one-way ANOVA testing. The one-way ANOVA decomposition can be obtained with a slight manipulation of the equation above:

$$\begin{bmatrix} y_{0,1}-\bar{y} \\ \vdots \\ y_{0,n_0}-\bar{y} \\ y_{1,1}-\bar{y} \\ \vdots \\ y_{1,n_1}-\bar{y} \\ \vdots \\ \vdots \\ y_{k,1}-\bar{y} \\ \vdots \\ y_{k,n_k}-\bar{y} \end{bmatrix} = (\bar{y}_0-\bar{y}) \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} + (\bar{y}_1-\bar{y}) \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \\ \vdots \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \cdots + (\bar{y}_k-\bar{y}) \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \begin{bmatrix} y_{0,1}-\bar{y}_0 \\ \vdots \\ y_{0,n_0}-\bar{y}_0 \\ y_{1,1}-\bar{y}_1 \\ \vdots \\ y_{1,n_1}-\bar{y}_1 \\ \vdots \\ \vdots \\ y_{k,1}-\bar{y}_k \\ \vdots \\ y_{k,n_k}-\bar{y}_k \end{bmatrix}.$$

The $k+1$ terms on the right side are orthogonal, and Pythagoras theorem gives

$$\sum_{j=0}^{k}\sum_{i=1}^{n_j}\big(y_{j,i}-\bar{y}\big)^2 = \sum_{j=0}^{k} n_j\big(\bar{y}_j-\bar{y}\big)^2 + \sum_{j=0}^{k}\sum_{i=1}^{n_j}\big(y_{j,i}-\bar{y}_j\big)^2.$$

The term on the left side is the total sum of squares of Y (SST). The second term on the right is the residual sum of squares of the regression, which coincides with the one-way ANOVA within-groups sum of squares (SSW). The sum of the other terms is the sum of squares explained by the regression, which coincides with the between-groups sum of squares (SSB). So, the one-way ANOVA decomposition is a particular case of the ANOVA decomposition of the linear regression analysis. The same for the F test.

Example 1. The `jobsat1` data set has been used in lecture 16 to illustrate the one-way ANOVA test. I come back to this example to present it in regression style. I define two dummies, for Mexico and Spain, respectively. So. I take Chile as zero group (your stat package will do this, based on alphabetic order).

TABLE 1. Linear regression results (Example 1)

Coefficient	Estimate	Std. error	t value	p-value
Intercept	4.158	0.080	51.7	0.000
countryME	0.255	0.116	2.20	0.029

I drop first from the analysis the Spanish group, using only the Mexico dummy. The table of coefficients is Table 1. The R-squared statistic is 0.021 ($F = 4.82$, $P = 0.029$). Note that the slope coefficient is exactly the same as the mean difference Mexico minus Chile, and the t statistic is the same as in the two-sample t test.

TABLE 2. Linear regression results (3 groups)

Coefficient	Estimate	Std. error	t value	p-value
Intercept	4.158	0.078	53.6	0.000
countryME	0.255	0.112	2.28	0.023
countrySP	0.005	0.099	0.05	0.963

I use next the complete sample and the two dummies. The table of coefficients is Table 2. Now, $R^2 = 0.017$ ($F = 3.58$, $P = 0.029$). The associated F statistic is the same as the one-way ANOVA F statistic. □

Analysis of variance

The analysis of variance (ANOVA) is one of the classical methods of Statistics. In ANOVA, there is a dependent variable Y, usually called **response** and a set of categorical independent variables, called **factors**. The values of the factors are called **levels**. ANOVA techniques are based on different ways of decomposing the sum of total squares of the response. In the decomposition, there is one term for each factor and, eventually additional terms for **interaction effects**, which occur when the effect of one factor depends on the level of another factor.

The decomposition is presented in an ANOVA table, and used for testing the mean differences among the groups defined by the combinations of the levels of the factors. In the ANOVA table, every factor has as many degrees as the number of levels minus one.

We have already seen the simplest case, the one-way ANOVA test, in which there is only one factor. With more than two factors, ANOVA becomes embroiled, and the formulas of the sums of squares get nasty, crowded with subscripts (such as y_{ijk}, or $\bar{y}_{ij+}$). An additional source of complexity is the interpretation of the interaction effects for factors with more than two levels.

When a set of continuous independent variables, called **covariates**, are included, the expression **analysis of covariance** (ANCOVA) is used. ANCOVA is based on an ANOVA table in which every covariate has one degree of freedom (corresponding to one term in the regression equation).

Although ANOVA is a classic, still included in many Statistics textbooks, it is less used nowadays, because, as we have seen in this lecture, it can be replaced by the adequate regression analysis, which is simpler to manage. Econometrics textbooks do not contain ANOVA chapters, since a regression approach, in which the factors are handled through dummy variables, is preferred. Nevertheless, ANOVA still survives in the analysis of experimental data.