[STAT-12] 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 we have not yet 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, $Y \sim \mathcal{N}(\beta_0 + \beta_1, \sigma^2)$. With $X = 0$, 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. 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

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

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, $\hat{\beta}_0 = \bar{y}_0$ is the intercept, and $\hat{\beta}_1 = \bar{y}_1 - \bar{y}_0$ is the slope.

It can proved without pain that the estimated standard error for the slope is

$$\hat{\text{se}}[\hat{\beta}_1] = s\sqrt{\frac{1}{n_0} + \frac{1}{n_1}},$$

Then, the corresponding t statistic is

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

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 $X_1, \ldots, X_k$. 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, $\beta_1, \ldots, \beta_k$ 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

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

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:

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

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

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

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 already been used in this course 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. Here, $R^2=0.021$ ($F=4.82$, $P=0.029$). 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. We see below how to deal with them.

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.

Product terms

In general, when the effect of X_1 on Y depends on the value of X_2, we say that there is an interaction effect of X_1 and X_2 on Y. An interaction effect is included in a regression equation through a product term, as in

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon.$$

We identify the interaction effect with the coefficient β_{12}. In mathematical terms, the interpretation of the coefficient is easy, taking X_2 as a moderator of the effect of X_1: holding constant X_2, the effect of X_1 on Y is $\beta_1 + \beta_{12} X_2$. Nevertheless, in practice, one has to be careful with product terms. Let me point out some practical issues:

- There is no way to separately interpret the three terms in which X_1 and X_2 are involved. If your research question involves an interaction effect of X_1 and X_2, you should better leave the three terms in the equation, either significant or not.
- The best way to understand an interaction effect is to assign to one of the two factors the role of a moderator of the effect of the other, as suggested above.
- Product terms can be affected by a multicollinearity problem, showing less significance than expected. This is usually fixed by centering the moderator (subtracting the mean).

Interpretation of a moderation effect

The simplest case is that of a binary moderator. Suppose an equation

$$Y = \alpha + \beta X + \gamma Z + \delta X Z + \epsilon,$$

in which Z is a dummy variable which codes two groups (e.g. male/female). To interpret the coefficient of the product term, we look at the two groups separately. In the group $Z = 0$, the model becomes

$$Y = \alpha + \beta X + \epsilon,$$

while in the group $Z = 1$ we have

$$Y = \big(\alpha + \gamma\big) + \big(\beta + \delta\big) X + \epsilon.$$

So, the effects of X in the two groups are β and $\beta + \delta$, respectively. Therefore, δ accounts for the change in the effect of X across groups.

In the case of a continuous moderator, the simplest way to address the moderation issue is to reduce the discussion to a three level scenario. For instance, you can standardize Z and consider the cases $Z = -1$, $Z = 0$ and $Z = 1$. Alternatively, you can pick the Z values so that they allow for an interesting discussion.

Standardized coefficients

Linear transformations can be applied to the variables involved in a regression equation. This does not affect the R-squared and t statistics, but could make the interpretation of the coefficients more appealing. First, it is easy to see that, if we replace the dependent variable Y by $Y^* = (Y - a)/b$, we have to divide all the coefficients on the right side of the equation by b and subtract a/b from the constant term. Second, if we replace an independent variable X by $X^* = (X - a)/b$, this only affects the coefficient of X and the constant term.

Centering all the variables, that is, subtracting the means, is sometimes applied, to get an equation without constant term. Standardizing all the variables ($Z = (X - \bar{x})/s$), we go a bit further. The coefficients of the resulting equation are called **standardized coefficients**, or beta coefficients.

In the regression line, the standardized slope coefficient is the same as the correlation. This is no longer true in multiple regression, although it may seem so. Indeed, the absolute value of a standardized coefficient can exceed 1. But most people look at standardized coefficients as if they were correlations. The correct interpretation of a standardized coefficient β would be: increasing X one standard deviation while holding the other dependent variables constant, the expected value of Y is increased by β standard deviations.

Standardization is usually recommended when each variable in the equation is a perception measure derived from a **Likert scale**, that is, from a set of questions with a small number of response options (typically 1–5 or 1–7). Standardized coefficients can be compared and, with a bit of practice, analysts get used to interpret them at first sight. For instance, a trained analyst may tell you that 0.1 is a weak effect, but 0.75 is a strong one, irrespective of the specific perceptions involved in the model.

But, in many cases, you do not standardize all the variables. Typical examples are:

- Some typical control variables, such as age, are rarely standardized.
- Dummies are never standardized.
- If you include a product term in the equation to account for an interaction effect of two perception measures, you will probably standardize the factors but not the product.

Example 2. The `supermarket` data set was obtained in a large chain of retail supermarkets offering two alternative check-out systems, a self-service option and a traditional, employee-assisted, checkout process. The managers were interested in the customer's perception of service quality and in the extent to which the difference between the two check-out systems made a real difference in the customer's perception. They also wished to examine another point about which there was no consensus: whether the connection between the quality perception by the customers and their fidelity, that quality specialists take for granted, is real.

The sample ($n = 210$) consisted of two groups: one group of customers using the self-service option and another group using the employee-assisted system. The survey was based on short interviews conducted at the store. The questionnaire contained two items related to the overall customer's satisfaction and intention to reuse, and three items for each of the three potential drivers of these outcomes, process performance (speed and effectiveness of the scanning and payment operations), process convenience (service layout and waiting time), employee performance (interaction with employees).

The variables are:

- `type`, type of service, type = 1 for self-service.

- `perfo`, process performance, 1-7 scale (average).
- `conv`, process convenience, 1-7 scale (average).
- `employee`, employee's performance, 1-7 scale (average).
- `sat`, overall satisfaction, 1-7 scale.
- `reuse`, intention to reuse, 1-7 scale.

I fit a linear regression equation to these data (Table 3), with the intention to reuse as the dependent variable. Except for the dummy, which I leave for the end, the coefficients of the other four regressors can be compared, because the scale is common. The results give a certain support for the role of customer satisfaction as a mediator between the process attributes and the intention to reuse, but we see that, given the satisfaction level, the process performance still contributes. The other two attributes do not contribute much, in comparative terms. The coefficient of the dummy can be interpreted as a mean difference between the two groups (given the other regressors), and is not relevant in a 1–7 scale.

TABLE 3. Linear regression results (Example 2)

Coefficient	Estimate	Std. error	t value	p-value
Intercept	4.34	0.33	13.2	0.000
type	–0.074	0.078	-0.95	0.341
perfo	0.220	0.060	3.66	0.000
conv	–0.013	0.037	–0.36	0.721
employee	0.002	0.054	0.04	0.964
sat	0.169	0.058	2.93	0.004

Since the units of the metric variables used in the analysis are artificial, most analysts would standardize them. I denote with a prefix `z` the standardized version. The regression results obtained after standardization are shown in Table 4. Note that the t statistics and the P-values of the coefficients do not change. The interpretation, for instance of the coefficient of `perfo`, would be: increasing `perfo` one standard deviation while holding the other independent variables constant, `reuse` increases by 0.292 standard deviations (on average). The interpretation of the coefficient of `type` is a bit different: switching from employee-assisted to self-service checkout while holding the other regressors constant, `reuse` decreases by 0.074 standard deviations.

TABLE 4. Linear regression results (standardized variables)

Coefficient	Estimate	Std. error	t value	p-value
Intercept	0.066	0.093	0.71	0.479
type	–0.132	0.138	–0.95	0.341
z.perfo	0.292	0.080	3.66	0.000
z.conv	–0.028	0.078	–0.36	0.721
z.employee	0.004	0.081	0.04	0.964
z.sat	0.246	0.084	2.93	0.004

But, do these four drivers of the intention to reuse have the same effect in both groups? To address this question, I introduce a product term for each one, using the prefix `t` for the product of `type` with any of the other variables. In Table 5, the first line associated to the performance (the coefficient of `perfo`) is related to the effect of this variable in the group type = 0, that is, with the employee-assisted checkout, while the second term (the coefficient of `tperfo`) is related to the difference between the effects in the two groups. This shows that the performance matters only in the self-service group. For `conv`, we find effects similar in absolute value, but with opposite sign,

so we do not have evidence of this effect. For `employee`, nothing is gained with the addition of the product term. For `sat`, I do not find difference between the two groups. Therefore, three of the product terms may be dropped.

TABLE 5. Linear regression results (with moderation effects)

Coefficient	Estimate	Std. error	t value	p-value
Intercept	–0.029	0.098	–0.30	0.766
type	–0.112	0.137	–0.81	0.417
z.perfo	0.074	.111	0.66	0.508
tperfo	0.414	0.161	2.57	0.011
z.conv	0.159	0.129	1.23	0.219
tconv	–0.301	0.161	–1.87	0.063
z.employee	–0.058	0.115	–0.51	0.613
temployee	0.061	0.165	0.37	0.712
z.sat	0.253	0.124	2.03	0.044
tsat	0.021	0.168	0.13	0.899

So, in my final equation (Table 6), I get a strong effect of the perfomance in the self-service group. The effect is weaker in the employee-assisted group.

TABLE 6. Final results

Coefficient	Estimate	Std. error	t value	p-value
Intercept	.031	0.093	0.33	0.739
type	–0.138	.137	–1.01	0.314
z.perfo	0.133	0.101	1.31	0.190
tperfo	0.321	0.128	2.50	0.013
z.conv	–0.036	0.077	–0.46	0.647
z.employee	–0.019	0.081	–0.23	0.817
z.sat	0.260	0.083	3.14	0.002

¶ Source: MP Castro-Amorim, PhD Dissertation.

Homework

A. In the `jobsat2` data set, test the gender and the country effects using a regression analysis. This would be equivalent to two-way ANOVA testing.

B. The `wage1` data set, used in the example of the preceding lecture, contains also two dummies, for being female and being married, respectively. Test the effect of education on wages taking into account that this effect may be moderated by gender and/or marital status.