Dummy variables in linear regression

Johan A. Elkink

December 2, 2014

This handout provides a number of examples of regression models, primarily to demonstrate the use of dummy variables and interaction models. The example SPSS syntax is based on the `bes.dta` data set available at the teaching data web page.[1]

1 Simple linear regression

The most basic model is a model where both the dependent variable and the independent variable are continuous, and there is only one independent variable. For example, we could explain trust in politicians in general (a 0 to 10 scale) by someone's self-placement on a left-right scale (also a 0 to 10 scale). The equation of this model would be:

$$trustpol_i = \beta_1 + \beta_2 lr_i + \varepsilon_i,$$

whereby β_1 would be the intercept (the predicted value of the trust variable when the left-right self-placement is 0) and β_2 the slope, the increase in trust in politics for every increase of one point on the left-right scale. The results for this simple regression[2] is in Table 1 and the code would be:

```
REGRESSION /DEPENDENT = trustpol /METHOD = ENTER lr.
```

The estimated intercept is thus 4.15 and the estimated slope 0.15, so that if one goes from the left extreme of the self-placement scale to the right extreme, the increase in trust is $0.15 \times 11 = 1.6$ and the predicted value on trust in politicians changes from 4.15 to 5.75. Both t-tests are significant, so we can reject the null hypothesis that $\beta_1 = 0$ and we can also reject the null hypothesis that $\beta_2 = 0$, so this really is evidence of a correlation between left-right self-placement and trust in politicians. The F-test will provide the same result as the t-test, because there is only one independent variable, and $R^2 = 0.03$, so 3% of the variation in trust in politicians is explained by the left-right position of a respondent.

[1] `http://www.joselkink.net/teaching/teaching-data/`

[2] Note that all tables are in a format typical for users of R (instead of SPSS), but not according to the normal requirements of a proper regression table. For the latter, see the combined Table 7.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.145	0.420	9.867	0.000
Left-right placement	0.147	0.073	2.016	0.046

Table 1: Linear regression explaining trust in politicians in general.

2 Simple regression with a dummy variable

For the remainder of this note, we will continue to use the same dependent variable, but explore different variations of the model. We will look at the dummy variable whether a respondent is a voter or not. We can create this using:

```
RECODE turnout (1 = 1) (MISSING = SYSMIS) (ELSE = 0) INTO voter.
REGRESSION /DEPENDENT = trustpol /METHOD = ENTER voter.
```

This way we are estimating the following model:

$$trustpol_i = \beta_1 + \beta_2 voter_i + \varepsilon_i,$$

with `voter` a binary variable that is 1 for voters and 0 for non-voters. Results are in Table 2. We now have an intercept of 4.31, so for non-voters, which is the "reference category" for the `voter` variable, the predicted average score on trust in politicians in general is 4.31 on the 0–10 scale we're using. The coefficient for the `voter` variable is 0.77, so the predicted average score on trust in politicians for voters is $4.31 + 0.77 = 5.08$. The t-tests are significant, so we can reject the null hypotheses that $\beta_1 = 0$ and $\beta_2 = 0$, and therefore voters have a statistically significantly different level of trust in politicians from non-voters.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.310	0.349	12.343	0.000
Voter	0.772	0.389	1.986	0.049

Table 2: Linear regression explaining trust in politicians in general.

3 Regression with a dummy and a continous variable

Let's combine these two models as follows:

$$trustpol_i = \beta_1 + \beta_2 voter_i + \beta_3 lr_i + \varepsilon_i,$$

such that we have a multiple regression with two independent variables. The SPSS code is then:

```
REGRESSION /DEPENDENT = trustpol /METHOD = ENTER voter lr.
```

The regression results are in Table 3. The estimate of $\hat{\beta}_1 = 3.45$ implies that for a non-voter who scores on the left extreme of the left-right scale, the predicted level of trust in politicians is 3.45. Voters, on the other hand, at the same point on the left-right scale, have a predicted level of

$3.45 + 0.81 = 4.26$. Voters are thus more trusting in politicians than non-voters. The coefficient on the left-right variable is positive, so the more right-wing a voter is, the more trust in politicians. From one extreme of the scale to the other, this implies a difference of $0.15 \times 11 = 1.6$ points on the trust in politicians scale – so the effect is not very large. All t-tests are significant, so voters are different from non-voters and left-right placement matters; we can reject the three null hypotheses (similar to above). Furthermore, the F-test is significant with a p-value of 0.015, thus rejecting the null hypothesis that all betas, jointly, are 0 ($H_0 : \beta_2 = \beta_3 = 0$). The R^2 is only 0.055, so just under 6% of the variation in trust in politicians is explained joinly by the left-right self-placement and whether someone is a voter or not.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.454	0.529	6.525	0.000
Voter	0.810	0.385	2.107	0.037
Left-right placement	0.154	0.072	2.135	0.034

Table 3: Linear regression explaining trust in politicians in general.

4 Regression with a nominal (or ordinal) categorical variable

In the next example, we will be using a nominal independent variable, but the exact same strategy holds for ordinal variables.[3] The strategy for a categorical variable is to generate dummy variables representing the various categories and, excluding one category, include all these dummy variables to the regression. It is not necessary to have exactly one dummy variable for each category, as one can create different groupings (e.g. just "Labour", "Conservatives" and "Other" for party, instead of more parties), as long as the categorisation makes sense. For example, it would not make sense to talk about "Labour and Conservatives" versus all other voters in the British context, since these two are very different groups, so one would not combine them like that. The grouping has to be of interest. For an ordinal scale, it is common to group multiple categories together, for example to create one dummy variable for "(strongly) agree" and one for "(strongly) disagree", thus grouping an original five point scale ("strongly agree", "agree", "neither agree nor disagree", "disagree", and "strong disagree") into just two dummy variables, with the neutral answer as reference category.

By way of example, let's add the `country` variable. We can use the following code:

```
RECODE country (2 = 1) (MISSING = SYSMIS) (ELSE = 0) INTO scotland.
RECODE country (3 = 1) (MISSING = SYSMIS) (ELSE = 0) INTO wales.
REGRESSION /DEPENDENT = trustpol /METHOD = ENTER scotland wales.
```

The results of this model, which follows equation

$$trustpol_i = \beta_1 + \beta_2 scotland_i + \beta_3 wales_i + \varepsilon_i,$$

[3] Except, with ordinal variables one would expect a consistent pattern in the coefficients, either consistently increasing or consistently decreasing as one moves along the ordinal scale; if this does not happen, one might cast doubt on the measurement of the ordinal variable.

are available in Table 4. In this model there are three possible scenarios, three possible types of respondents. A respondent might be from England, in which case the equation simplifies to $trustpol_i = \beta_1 + \beta_2 \cdot 0 + \beta_3 \cdot 0 = \beta_1$. In other words, the predicted value for trust in politicians for an Englishman is just $\beta_1 = 4.82$. For a Welsh person on the other hand, this would be $trustpol_i = \beta_1 + \beta_2 \cdot 0 + \beta_3 \cdot 1 = \beta_1 + \beta_3$, which in the case of the estimates of Table 4 implies a predicted value of $4.82 + 0.08 = 4.90$. For the Scottish respondent, we have instead $\beta_1 + \beta_2$ as the intercept. Note that because there are only dummy variables in the model, there is no such thing as a slope, there are only intercepts, which can be seen as estimates group means. Note that the t-tests for both dummies are insignificant, which implies that we cannot reject the null hypothesis that the mean trust score for English is the same as that of the Welsh, nor can we reject the null hypothesis that the mean trust score is the same for the Scottish and the English. There is no test in this model output about the difference between Scottish and Welsh respondents.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.816	0.204	23.607	0.000
Scotland	0.457	0.389	1.174	0.242
Wales	0.084	0.403	0.208	0.835

Table 4: Linear regression explaining trust in politicians in general.

5 Regression with different types of independent variables

We can now extend our original model by adding the country dummies. We might be interested in the following model:

$$trustpol_i = \beta_1 + \beta_2 voter_i + \beta_3 lr_i + \beta_4 scotland_i + \beta_5 wales_i + \varepsilon_i,$$

which includes both dummy and continuous variables. The SPSS code would straightforwardly be:

```
REGRESSION /DEPENDENT = trustpol /METHOD = ENTER voter lr scotland wales.
```

The output is presented in Table 5. What we can see here is that controlling for country, we find very similar effects to those in Table 3 and the interpretation is also very similar. For an English non-voter, who is at the extreme left end of the left-right scale, the mean trust level would be 3.41. Voters a statistically significantly more trusting in politicians than non-voters (a difference of 0.80) and more right-wing voters are more trusting, for each increase on the 11-point left-right scale by one point, an increase on the 11-point trust scale of 0.15 can be expected. A statistically significant, but very modest effect. With all those variables together, the model only explains 6% of the variation in trust ($R^2 = 0.06$) and indeed the adjusted R^2, which penalizes for adding a lot of variables, is now clearly lower ($Adj.R^2 = 0.04$). There are now many different scenarios to read the equation: voters in England, non-voters in Scotland, etc. To take one example, let's look at a voter in Wales – the equation in this case would simplify

to: $trustpol_i = \beta_1 + \beta_2 \cdot 1 + \beta_3 lr_i + \beta_4 \cdot 0 + \beta_5 \cdot 1 = (\beta_1 + \beta_2 + \beta_5) + \beta_3 lr_i$, thus for this category of respondent, we would have an intercept of $\beta_1 + \beta_2 + \beta_5$ and a slope of β_3, which in this estimation are $3.41 + 0.80 + 0.01 = 4.22$ and 0.15, respectively. Note how the slope is the same for all categories of respondents – it is always $\beta_3 = 0.15$.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.412	0.538	6.340	0.000
Voter	0.802	0.389	2.060	0.041
Left-right placement	0.148	0.073	2.034	0.044
Scotland	0.361	0.383	0.942	0.348
Wales	0.011	0.398	0.028	0.978

Table 5: Linear regression explaining trust in politicians in general.

6 Interaction between a dummy and a continuous variable

In the previous regression, while voters had a different intercept from non-voters, and the Welsh a different intercept from the English (so we have 6 different intercepts), the slope is the same for all respondents. The slope is the coefficient on the continuous variable, which is here the left-right self-placement. In some cases, however, one might expect the slope to be different for different groups. One might expect that for voters, left-right self-placement has a different impact on trust in politicians than for non-voters. We could estimate the following model:

$$trustpol_i = \beta_1 + \beta_2 voter_i + \beta_3 lr_i + \beta_4 voter_i \cdot lr_i + \varepsilon_i.$$

In this case, in SPSS one would need to compute the multiplicative term first:

```
COMPUTE voterLR = voter * lr.
REGRESSION /DEPENDENT = trustpol /METHOD = ENTER voter lr voterLR.
```

The results are presented in Table 6. The first thing to note, of course, is that all the t-tests are now insignificant. We cannot reject the null hypothesis that $\beta_2 = 0$, or that $\beta_3 = 0$, or that $\beta_4 = 0$. So we do not have evidence in this model that voters have a different intercept from non-voters (β_2), that left-right self-placement for non-voters (β_3) has an inpact on trust in politicians, or that this impact differs for voters from non-voters (β_4). Strikingly, the F-test (not visible in the table, but in the SPSS output) is in fact statistically significant with a p-value of 0.02. This implies that we can in fact reject the null hypothesis that all betas are zero ($H_0 : \beta_2 = \beta_3 = \beta_4 = 0$). In other words, these variables jointly do explain to some extend the level of trust in politicians (with an R^2 of 0.07), but there is insufficient support to properly separate the effect of being a voter or not or being left or right in politics.

In this example there are just two scenarios, we either have a voter or a non-voter. For non-voters the equation simplifies to: $trustpol_i = \beta_1 + \beta_2 \cdot 0 + \beta_3 lr_i + \beta_4 \cdot 0 \cdot lr_i = \beta_1 + \beta_3 lr_i$, so we have an intercept that is β_1 and a slope coefficient β_3. For voters, however, this would be: $trustpol_i = \beta_1 + \beta_2 \cdot 1 + \beta_3 lr_i + \beta_4 \cdot 1 \cdot lr_i = (\beta_1 + \beta_2) + (\beta_3 + \beta_4) lr_i$ and we have an intercept of

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.345	0.896	4.851	0.000
Voter	-0.342	1.011	-0.339	0.735
Left-right placement	-0.006	0.149	-0.042	0.967
Voter × left-right placement	0.210	0.170	1.232	0.220

Table 6: Linear regression explaining trust in politicians in general.

$\beta_1 + \beta_2$ with a slope of $\beta_3 + \beta_4$. So in this model both the intercept and the slope differ between voters and non-voters (albeit statistically insignificantly so). For voters the estimated intercept is therefore $4.35 - 0.34 = 4.01$ and the estimated slope $-0.01 + 0.21 = 0.20$, while for non-voters these are 4.35 and -0.01, respectively.

Table 7 provides a table summarizing all models described in this note and presents them more appropriately, with the relevant statistics added and the stars for ease or readability.

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
Left-Right	0.150**		0.150**		0.150**	−0.006
	(0.073)		(0.072)		(0.073)	(0.150)
Voter × L-R						0.210
						(0.170)
Voter		0.770**	0.810**		0.800**	−0.340
		(0.390)	(0.390)		(0.390)	(1.000)
Scotland				0.460	0.360	
				(0.390)	(0.380)	
Wales				0.084	0.011	
				(0.400)	(0.400)	
Constant	4.100***	4.300***	3.500***	4.800***	3.400***	4.300***
	(0.420)	(0.350)	(0.530)	(0.200)	(0.540)	(0.900)
N	150	150	150	150	150	150
R^2	0.027	0.026	0.055	0.009	0.061	0.065
Adj. R^2	0.020	0.019	0.042	−0.004	0.035	0.046
F-test	4.100**	3.900**	4.300**	0.690	2.400*	3.400**

***p $<$ .01; **p $<$ .05; *p $<$ .1

Table 7: Linear regression explaining trust in politicians in general.