Multiple regression: Categorical independent variables and interaction effects

Johan A. Elkink
School of Politics & International Relations
University College Dublin

20 November 2017

Outline

Outline

Introduction

So far, we have discussed regressions where both the dependent and the independent variables were **continuous**, or of interval/ratio measurement level.

In particular in the social sciences, variables are often **qualitative** or **categorical** in nature.

When an independent variable is categorical in nature, the estimation remains the same, but the interpretation changes.

Dummy variables

A **dummy variable** is a binary variable that can only have values 0 or 1.

In regression analysis, a dummy variable can be added as an independent variable without any problems. If a categorical variable is coded differently, you cannot add it to the model.

respnr	gender	female
1	Male	0
2	Female	1
3	Male	0
4	Male	0
5	Female	1
6	Female	1
7	Female	1

Dummy variables

A **dummy variable** is a binary variable that can only have values 0 or 1.

In regression analysis, a dummy variable can be added as an independent variable without any problems. If a categorical variable is coded differently, you cannot add it to the model.

respnr	**gender**	**female**
1	Male	0
2	Female	1
3	Male	0
4	Male	0
5	Female	1
6	Female	1
7	Female	1

In SPSS: `RECODE gender ("Male" = 0) ("Female" = 1) INTO female.`

In Stata: `recode gender (1 = 0) (2 = 1), gen(female)`

Regression with dummy variables

Model 1: $y_i = \beta_1$, i.e. a model without any independent variables.

Here you would simply obtain: $\hat{\beta}_1 = \bar{y}$.

(This also shows that regression is close to estimating means and the *t*-test is also the same as for comparing means.)

Regression with dummy variables

Model 2: $y_i = \beta_1 + \beta_2 d_i$, where D is a dummy variable. Here there are two scenarios:

$d_i = 0$:

$$y_i = \beta_1 + \beta_2 \cdot 0 = \beta_1$$

and we just estimate the mean of Y for the group where $D = 0$.

$d_i = 1$:

$$y_i = \beta_1 + \beta_2 \cdot 1 = \beta_1 + \beta_2$$

and that sum is the estimated mean of Y for the group where $D = 1$.

The estimate $\hat{\beta}_2$ is therefore the **difference in means** for the two groups.

Regression with dummy variables

Model 3: $y_i = \beta_1 + \beta_2 d_i + \beta_3 x_i$, where D is a dummy variable and X is continuous. Here there are two scenarios:

$d_i = 0$:

$$y_i = \beta_1 + \beta_2 \cdot 0 + \beta_3 x_i = \beta_1 + \beta_3 x_i$$

and we have an **intercept** $\hat{\beta}_1$ and a **slope coefficient** $\hat{\beta}_3$ for the group where $D = 0$.

$d_i = 1$:

$$y_i = \beta_1 + \beta_2 \cdot 1 + \beta_3 x_i = (\beta_1 + \beta_2) + \beta_3 x_i$$

and we have an **intercept** $\hat{\beta}_1 + \hat{\beta}_2$ and a **slope coefficient** $\hat{\beta}_3$ for the group where $D = 1$.

Dummy variables and interpretation

So, dummy variables test whether the intercept (means) differ—do *not* interpret the respective coefficient as "if X increases by 1 unit, Y increases by ..."

Dummy variables and t-tests

$$y_i = \beta_1 + \beta_2 d_i + \beta_3 x_i$$

In a regression, the t-test for a coefficient tests whether, given the other variables in the model, the slope of a line is different from zero, with zero being no effect of X on Y.

$H_0 : \beta_3 = 0$, so under the null, the slope of the line is zero.

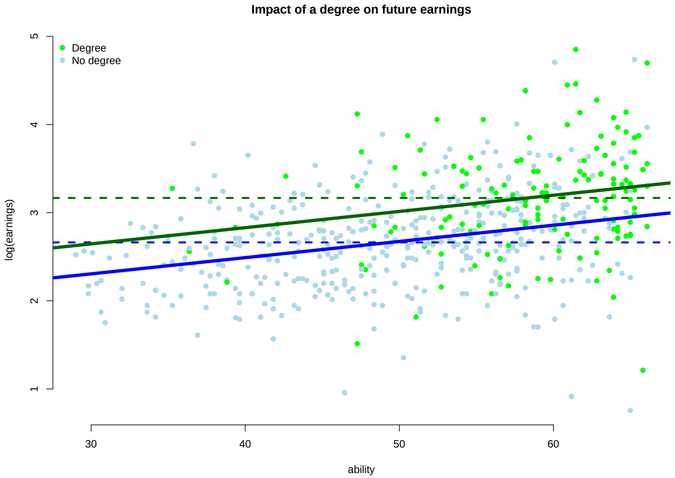
In a regression with a dummy variable, the t-test for that coefficient tests whether, given the other variables in the model, the mean of the two groups differ.

$H_0 : \beta_2 = 0$, so under the null, the two groups have the same intercept.

Example: degree and earnings

degree	0.504***	0.340***
	(0.054)	(0.058)
ability		0.018***
		(0.003)
intercept	2.662***	1.754***
	(0.028)	(0.140)
N	540	540
R^2	0.139	0.204
Adjusted R^2	0.138	0.201
Residual Std. Error	0.552	0.531
F-Statistic	87.020***	68.882***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Example: degree and earnings



$$\log(earnings_i) = \beta_1 + \beta_2 degree_i + \beta_3 ability_i$$

Outline

Multiple categories

Instead of just two categories, a categorical variables can have multiple categories, such as party preference or religious denomination. To add these to the regression, we split them up in **multiple dummy variables**.

respnr	party	ff	fg	lab	sf
1	Fianna Fáil	1	0	0	0
2	Sinn Féin	0	0	0	1
3	Labour	0	0	1	0
4	Sinn Féin	0	0	0	1
5	Fianna Fáil	1	0	0	0
6	Fianna Fáil	1	0	0	0
7	Fine Gael	0	1	0	0
8	Fine Gael	0	1	0	0
9	Labour	0	0	1	0

Multiple categories

respnr	party	ff	fg	lab	sf
1	Fianna Fáil	1	0	0	0
2	Sinn Féin	0	0	0	1
3	Labour	0	0	1	0
4	Sinn Féin	0	0	0	1
5	Fianna Fáil	1	0	0	0
6	Fianna Fáil	1	0	0	0
7	Fine Gael	0	1	0	0
8	Fine Gael	0	1	0	0
9	Labour	0	0	1	0

Note that in a regression always one category has to be left out, and all the other results are relative to this **reference category**, e.g.:

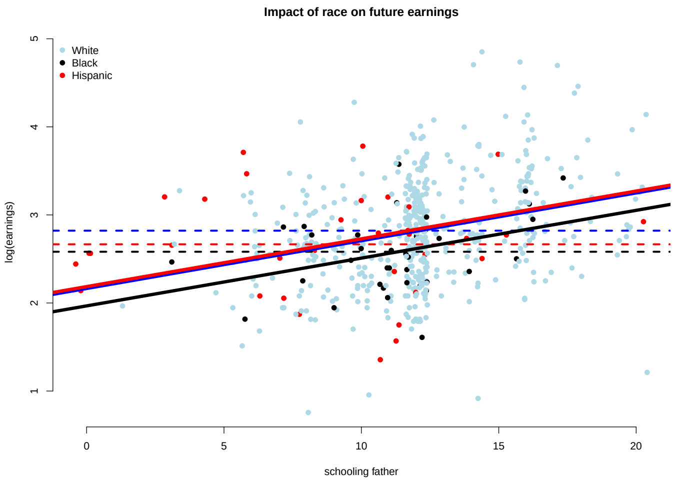
$$Y_i = \beta_1 + \beta_2 fg_i + \beta_3 lab_i + \beta_4 sf_i,$$

such that all coefficients show the difference relative to Fianna Fáil voters.

Example: race and earnings

ethblack	-0.239^{**}	-0.198^{**}
	(0.098)	(0.094)
ethhisp	-0.155	0.022
	(0.105)	(0.103)
schoolingFather		0.054^{***}
		(0.008)
intercept	2.821^{***}	2.164^{***}
	(0.027)	(0.095)
N	540	540
R^2	0.014	0.101
Adjusted R^2	0.010	0.096
Residual Std. Error	0.591	0.565
F-Statistic	3.800^{**}	20.011^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

Example: race and earnings



$$\log(earnings_i) = \beta_1 + \beta_2 ethblack_i + \beta_3 ethhisp_i + \beta_4 schoolingFather_i$$

Example: race and earnings

ethwhite		2.821***
		(0.027)
ethblack	−0.239**	2.582***
	(0.098)	(0.095)
ethhisp	−0.155	2.666***
	(0.105)	(0.101)
intercept	2.821***	
	(0.027)	
N	540	540
R^2	0.014	0.957
Adjusted R^2	0.010	0.957
Residual Std. Error	0.591	0.591
F-Statistic	3.800**	4,026.080***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Outline

Outline

Interactions

So far, we have only been adding variables in an **additive model**.

Imagine, however, that the relation between X and Y would depend on the group—e.g. the effect of ability on income is greater for those with a degree than those without a degree.

We call this an interaction effect, we have to **interact** the variable X with D, for example:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 d_i + \beta_4 x_i d_i.$$

Interaction with dummy variables

Model 4: $y_i = \beta_1 + \beta_2 x_i + \beta_3 d_i + \beta_4 x_i d_i$, where D is a dummy variable and X is continuous. Here there are two scenarios:

$d_i = 0$:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 \cdot 0 + \beta_4 x_i \cdot 0 = \beta_1 + \beta_2 x_i$$

and we have an intercept $\hat{\beta}_1$ and a slope coefficient $\hat{\beta}_2$ for the group where $D = 0$.

$d_i = 1$:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 \cdot 1 + \beta_4 x_i \cdot 1 = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) x_i$$

and we have an intercept $\hat{\beta}_1 + \hat{\beta}_3$ and a slope coefficient $\hat{\beta}_2 + \hat{\beta}_4$ for the group where $D = 1$.

Including component variables

Note that this also shows the importance of including the component variables that make up the interaction. E.g.:

$$y_i = \beta_1 + \beta_2 d_i + \beta_3 x_i d_i,$$

where we exclude the variable X by itself, we would have:

$d_i = 0$:

$$y_i = \beta_1 + \beta_2 \cdot 0 + \beta_3 x_i \cdot 0 = \beta_1$$

and we have an intercept $\hat{\beta}_1$ and a slope coefficient 0 (!) for the group where $D = 0$.

$d_i = 1$:

$$y_i = \beta_1 + \beta_2 \cdot 1 + \beta_3 x_i \cdot 1 = (\beta_1 + \beta_2) + \beta_3 x_i$$

and we have an intercept $\hat{\beta}_1 + \hat{\beta}_2$ and a slope coefficient $\hat{\beta}_3$ for the group where $D = 1$.

So we **arbitrarily fix one slope** to zero.

Including component variables

Or similarly:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i d_i,$$

where we exclude the dummy variable D by itself:

$d_i = 0$:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i \cdot 0 = \beta_1 + \beta_2 x_i$$

and we have an intercept $\hat{\beta}_1$ and a slope coefficient $\hat{\beta}_2$ for the group where $D = 0$.

$d_i = 1$:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i \cdot 1 = \beta_1 + (\beta_2 + \beta_3) x_i$$

and we have an intercept $\hat{\beta}_1$ and a slope coefficient $\hat{\beta}_2 + \hat{\beta}_3$ for the group where $D = 1$.

So we **fix the value of Y to be identical** for the two groups at the **arbitrary point of $X = 0$**.

Interaction models and t-tests

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 d_i + \beta_4 x_i d_i$$

So we can think of the following t-tests:

$H_0 : \beta_2 = 0$, so under the null, the slope of the line is zero, *for the group where* $D = 0$.

$H_0 : \beta_3 = 0$, so under the null, the two groups have the same intercept.

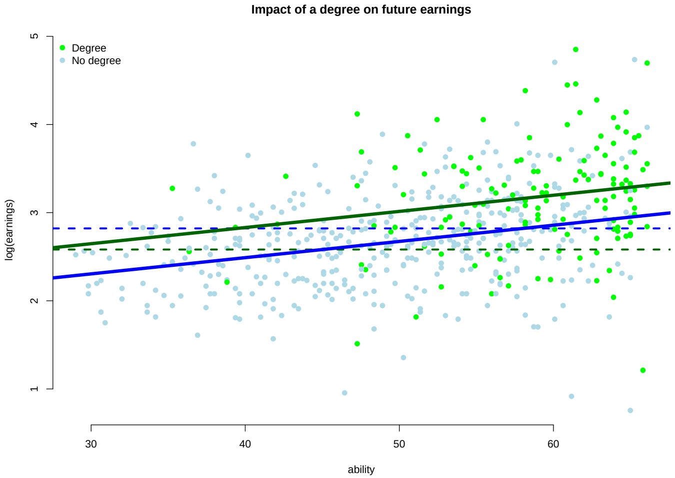
In a regression with an interaction with a dummy variable, the t-test for that coefficient tests whether, given the other variables in the model, the slope for the two groups differ.

$H_0 : \beta_4 = 0$, so under the null, the two groups have the same slope between X and Y.

Example: degree and earnings

degree	0.340***	0.345
	(0.058)	(0.428)
ability	0.018***	0.018***
	(0.003)	(0.003)
degree × ability		−0.0001
		(0.007)
intercept	1.754***	1.753***
	(0.140)	(0.153)
N	540	540
R^2	0.204	0.204
Adjusted R^2	0.201	0.200
Residual Std. Error	0.531	0.531
F-Statistic	68.882***	45.836***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Example: degree and earnings

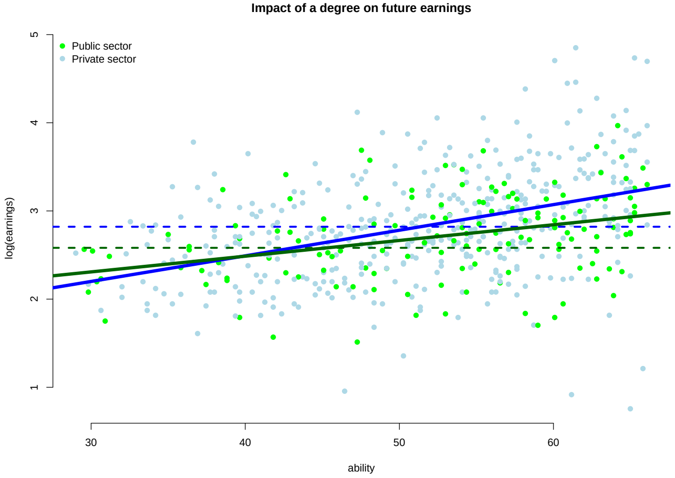


$$\log(earnings_i) = \beta_1 + \beta_2 degree_i + \beta_3 ability_i + \beta_4 degree_i \cdot ability_i$$

Example: public sector and earnings

publicSector	-0.141^{***}	0.445
	(0.053)	(0.300)
ability	0.026^{***}	0.029^{***}
	(0.003)	(0.003)
publicSector × ability		-0.011^{**}
		(0.006)
intercept	1.496^{***}	1.329^{***}
	(0.135)	(0.159)
N	540	540
R^2	0.163	0.169
Adjusted R^2	0.160	0.165
Residual Std. Error	0.544	0.543
F-Statistic	52.418^{***}	36.444^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

Example: public sector and earnings



$$\log(earnings_i) = \beta_1 + \beta_2 publicSector_i + \beta_3 ability_i + \beta_4 publicSector_i \cdot ability_i$$

Outline

Example: race and earnings

$$\begin{aligned}\log(earnings_i) =& \beta_1 + \beta_2 black_i + \beta_3 hisp_i + \beta_4 ability_i \\ &+ \beta_5 black_i \cdot ability_i + \beta_6 hisp_i \cdot ability_i,\end{aligned}$$

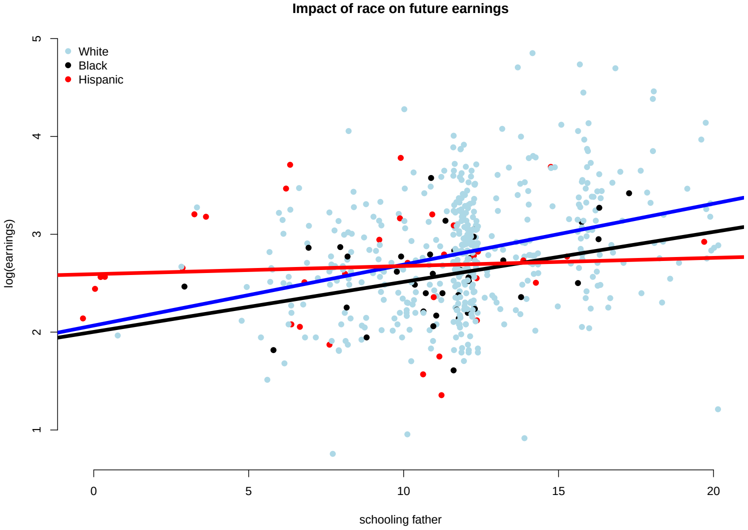
Whites: $\log(earnings_i) = \beta_1 + \beta_4 ability_i$
Blacks: $\log(earnings_i) = (\beta_1 + \beta_2) + (\beta_4 + \beta_5) ability_i$
Hispanics: $\log(earnings_i) = (\beta_1 + \beta_3) + (\beta_4 + \beta_6) ability_i$

So β_2 and β_3 are differences in intercepts, relative to whites; β_5 and β_6 are differences in slopes, relative to whites and t-tests test whether intercepts or slopes, respectively, differ.

Example: race and earnings

ethblack	−0.198**	−0.065
	(0.094)	(0.395)
ethhisp	0.022	0.525**
	(0.103)	(0.229)
schoolingFather	0.054***	0.062***
	(0.008)	(0.008)
ethblack × schoolingFather		−0.011
		(0.034)
ethhisp × schoolingFather		−0.054**
		(0.022)
intercept	2.164***	2.067***
	(0.095)	(0.104)
N	540	540
R^2	0.101	0.111
Adjusted R^2	0.096	0.103
Residual Std. Error	0.565	0.563
F-Statistic	20.011***	13.312***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Example: race and earnings



Outline

Interactions between continuous variables

It is possible to interact two continuous variables. Here you expect the effects of X on Y to gradually change as some third variable Z changes.

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 x_i z_i,$$

so when we take X as the key independent variable, we have:

Intercept:	$\beta_1 + \beta_3 z_i$
Slope:	$\beta_2 + \beta_4 z_i$

Both intercept and slope change with Z. These types of models are typically somewhat difficult to interpret and there is no statistical difference between whether the slope between X and Y varies for different values of Z, or the slope between Z and Y varies for different values of X. It requires a strong theory on causal relations to be able to make sense of the results.