A guide to interpreting regression tables

Semra Sevi
Université de Montréal
Department of Political Science

email: semra.sevi@umontreal.ca
website: www.semrasevi.com

January 2021

1 Types of variables

- The type of model researchers use will depend on the variables they have. Take the time to get to know the data and the research design. When you read the assigned articles, ask yourself how are the dependent (DV) and independent variables (IV) measured. The independent variable is the variable that influences the value of the dependent variable (also known as the outcome of interest).

- Continuous variables: these are measured along a numerical scale. For example: age, and income.

- Discrete variables are variables that take discrete values (0, 1, 2, 3, 4...). There are two general types:

  - Categorical or nominal variables. For example: religion (Protestant, Catholic, Jewish, Muslim...), parties (Liberal, Conservative, NDP, Green, Bloc...), and occupation. When a variable takes two categories for example gender (male or female) this variable is known as a 'dummy'/dichotomous variable. When they are an independent variable in a regression model, their interpretation is very simple. As we move from one category to the next, the associated change in the dependent variable is...? For example, "women are more likely to vote for Left parties compared to men." When there are multiple-category variables in the models, we can use a similar comparison when we examine their effect. For example, lets say we have three-category variable for religion (Catholic, Protestant and Others). In this case you have to use a reference category. Suppose it is Others. "Compared to Others, Catholics are 10 percent more likely to vote for the Liberal Party of Canada, but Protestants are only 3 per cent more likely than Others to do so."
  - Ordinal-level variables take on values that have a specific order. For example: degree of satisfaction (very, somewhat, a little, not at all), and degree of interest in politics (very high, somewhat high, somewhat low, very low).

2 Linear Regressions

Linear regressions are a method to test whether/to what extent whenever the IV increases in value the DV systematically increases/decreases in value. By convention the DV is y and the IV is x and the relationship between the two is represented by an equation.

Bivariate ordinary least squares (OLS) regression analysis:

- Bivariate means two variables. They can tell us something about the extent to which two unobserved population variables, x and y, are related.

The equation for a bivariate linear regression can be simply expressed as:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Where :

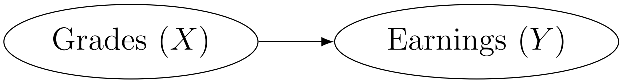
$$\begin{aligned} y &= \text{Dependent variable} \\ \beta_0 &= \text{Intercept/Constant} \\ \beta_1 &= \text{Slope Coefficient} \\ x_1 &= \text{Independent variable} \\ \varepsilon &= \text{Error term/epsilon} \end{aligned}$$

β_0 is the average value of y when the independent variable equals 0. This is rarely meaningful/interpretable so for our purposes it can be ignored.

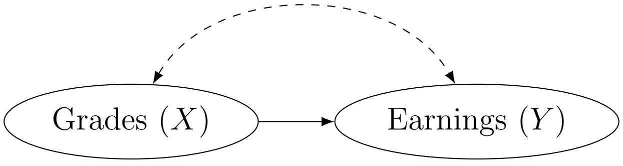
β_1 represents the coefficient for the slope which is the average change in y associated with a unit change in x.

ε is the error term which represents the remaining variation in y that cannot be explained with x.

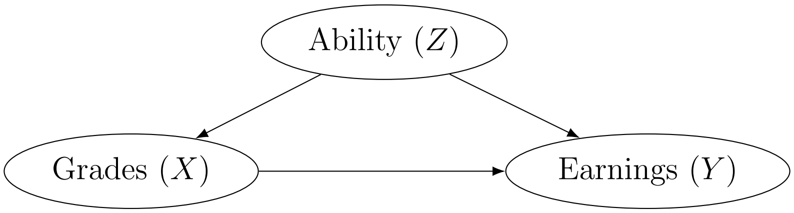
Let's look at this visually. Suppose you were interested in the relationship between student grades (X) and earnings (Y). Your hypothesis is that higher grades cause higher salaries.



But do they really? Is it possible that some other variable is affecting both student grades (X) and earnings (Y)? Could you think of a variable that affects both grades and earnings?



Suppose you said ability. Students who are more competent tend to get better grades and they also have higher salaries later in life. Perhaps grades does not matter. What matters is their intrinsic ability. In regression, we call this an omitted variable bias because in the earlier example we tried to explain a cause and effect without controlling for all potential factors that could affect both student grades and earnings. This is why we have multiple regressions.



Multiple OLS regression analysis:

- Often times we need to include multiple variables to control for confounders or to measure the impact of different IVs and which one has the strongest impact. Multiple regression provides estimates of the impact of each independent variable on the dependent variable, accounting for the impact of the other variables in the model. If there are two variables in the model, the coefficient on x_1 (e.g. explanatory variable) indicates its impact on y after controlling for x_2 (e.g. additional explanatory variable/confounding/control variable).

The equation for a multiple regression with two x's looks like this:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Same interpretation as the bivariate linear regression except that in a multiple regression, β_1 represents the coefficient for the slope which is the average change in y associated with a unit change in x_1 when all the other independent variables (x_2) are held constant.

To get more information about regressions you can consult and play around in this app created by the Centre for Computational and Quantitative Social Science and the Consortium on Electorial Democracy at Western University.

3 Examples

1. DV & IV's continuous

Political Ideology in Sweden. Ordinary least squares (OLS) regression models with political ideology as the dependent variable.

	(Model 1)	(Model 2)
Age	0.011*	0.005
	(0.005)	(0.005)
Income		0.277***
		(0.069)
Political Information		-0.013
		(0.145)
Constant	4.309***	3.696***
	(0.246)	(0.294)
Observations/N	921	921
R^2	0.005	0.023

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Source: Shane P. Singh's Course on Statistics and Data Analysis at ICPSR

The first thing to do when reading a table is to look at the Observations (N) and understand what they are. Here the N are the respondents in a survey.

The DV is left-right ideology (0 means extreme left and 10 means extreme right). The independent variables are: Age (the age of the respondent in years), Income (in categories from lowest income to highest income), Political information (number of political knowledge questions answered correctly). The unit of analysis is respondents in Sweden in this survey. For each respondent in the survey we have: ideology, age, income and political information.

The numbers not in the parentheses is called the coefficient. The coefficient provides the expected change in the dependent variable (party ideology) for one-unit increase in the independent variable (which is age in model 1).

The numbers in parentheses is called the standard error (we know this because of the note below the table), which is the estimate of the standard deviation of the coefficient. A rule of thumb, if a coefficient is twice as large compared to its standard error, then the coefficient is likely significant at the .05 level. For our purposes, you do not need to interpret this. Instead, you can look at the p-values (more on this soon).

First thing you should do is figure out what the N represents, what the dependent variable is and identify the important independent variables that you will base your interpretation on. Then ask yourself if the coefficient for the IV(s) is positive or negative. If it is positive, this means there is a positive relationship between the

IV and DV. As the IV increases, the DV increases. A negative relationship would indicate that as the IV increases, the DV decreases. The size of the coefficient for each independent variable gives you the size of the effect that variable is having on your dependent variable.

Let's try interpreting this table.

Here is the equation for model 1:

$$\text{Ideology} = \beta_0 + \beta_1 \times \text{Age} + \varepsilon$$

From our table above, we get the following values for our coefficients:

$$\text{Ideology} = 4.309 + .011 \times \text{Age} + \varepsilon$$

In model 1 (bivariate), we see that as your age increases by one year, your party ideology goes .011 point more to the right. So we can say that a ten-year increase in age, is associated with, on average, a 0.11 unit rightward shift in ideology (.011*10=0.11). You can tell that this is statistically significant (at the .05 level) by looking at the asterisks.

There are two basic hypotheses: the null and an alternative. The null hypothesis means that x is not linearly associated with y which indicates that the slope is zero. The alternative hypothesis means that x is linearly associated with y so the slope is not zero.

So statistically significant means that the β_1 which is the coefficient for the effect of *Age* on political ideology in this model is statistically different from zero.

The legend tells you that this coefficient is significant at $p < 0.05$ level. What does this mean? The 'p-values' indicate the statistical significance of the coefficient which is based off the principle of random sampling. In this model, there is only a 5 in a 100 chance (or less) that there really is no relationship between political ideology and age. If you want to read more about p-values and significance, you can read about it here. Note that the it is possible to have a very significant p-value (***) for a very small effect so the size of the p-value for a coefficient says nothing about the size of the effect that variable is having on your dependent variable.

Is model 1 **substantively** significant? Recall that ideology ranges from 0-10 so aging ten years only corresponds with an expected shift equal to 1 per cent of the scale.

Technically, the constant indicates the ideology of respondents aged zero which is found to be center left. This obviously does not make any sense which is why we normally do not interpret the constant. Sometimes, however, authors select variables in such a way that the constant is interpretable. For example, if they shifted the age variable to start at 18 instead of 0 then the constant would represent political ideology at age 18.

Now lets look at a multiple regression (model 2). The data tell us that those who are older are more on the right. But you may have reason that this may not be because of their age. It is rather that older people tend to be richer and it is the fact that they are rich that 'causes' them to be on the right, NOT the fact that they are old. So it is important to control for income. This means we need to have multiple variables in our model.

Here is the equation for model 2:

$$\text{Ideology} = \beta_0 + \beta_1 \times \text{Age} + \beta_2 \times \text{Income} + \beta_3 \times \text{Information} + \varepsilon$$

From our table above, we get the following values for the coefficients:

$$\text{Ideology} = 3.696 + .005 \times \text{Age} + .277 \times \text{Income} + (-.013) \times \text{Information} + \varepsilon$$

In model 2, with income and political information controlled, age no longer has a significant impact on ideology. Perhaps the effect of age that we found in model 1 can be explained by its relationship to income. Consider this as a speculation as the model does not test for this relationship. However it is possible that in a bivariate model an independent variable is a significant predictor of a dependent variable but is no longer significant in a multiple regression. In a multiple regression, the significance levels given for a given independent variable indicates whether that independent variable is a significant predictor of the dependent variable, controlling for the other independent variables.

R^2 is low in both models. This is ok. It may mean that there are lots of other things that contribute to political ideology that we did not add to the model. Or that the relationship is not linear. But just so you know, R^2 ranges from 0 to 1 and has a very simple interpretation: It is the proportion of variance in y "explained" by x(s). We can think of the model as defining a line in an n dimensional space where n is the number of variables. R^2 is a measure of how close the data is to this line. If all the points fall exactly on the line this would indicate a perfect relationship between x and y. When you add more independent variables to your model, R^2 will always increase.

2. DV continuous & Main IV dichotomous

The gender gap in Canadian federal elections from 1921-2015. Ordinary least squares (OLS) regression models with party vote share as the dependent variable.

	(Model 1)	(Model 2)
Woman	-9.877***	-0.450***
	(0.366)	(0.135)
Vote share lag		0.276***
		(0.006)
Party performance		0.666***
		(0.004)
Incumbent Party		6.783***
		(0.148)
Distance from contention		-0.015**
		(0.005)
Constant	27.682***	0.281
	(0.139)	(0.246)
Observations/N	23903	23903
R^2	0.030	0.872

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Source: Sevi, Semra, Vincent Arel-Bundock and André Blais. 2018. "Do Women Get Fewer Votes? No." *Canadian Journal of Political Science.* 52(1): 201-210

The equation for model 1 is:

$$\text{Vote share} = \beta_0 + \beta_1 \times \text{Woman} + \varepsilon$$

From our table above, we get the following values for our coefficients:

$$\text{Vote share} = 27.682 + (-9.877) \times \text{Woman} + \varepsilon$$

First thing is the N: they refer to all the candidates. In column 1, we have the percentage of the vote (DV) obtained by each candidate and whether the candidate is a man or a woman (IV).

The woman variable is a dummy variable (1 = Woman 0 = Man).

Note that because gender is a dummy variable, the interpretation is slightly different. It does not make sense to talk about the effect on vote share as woman increases or decreases (woman is not measured as a continuous variable. See above.)

Model 1 (bivariate) shows that compared to men, women candidate vote share is on average 10 percentage points lower. Another way to say this is when the candidate is a woman, the vote share is reduced by 10 percentage points. This is a very big effect and the estimate is statistically significant.

In model 2 (multiple), we have many controls (vote share lag, party performance, incumbent party and distance from contention as we may expect these variables to have an impact on both the dependant and the main independent variable). So the interpretation is: everything else being equal, women candidates get 0.45 percentage points fewer votes than men. But is this effect substantively important?

Here is the equation for model 2:

$$\begin{aligned}\text{Vote share} =& \beta_0 + \beta_1 \times \text{Woman} + \beta_2 \times \text{Vote share lag} \\ &+ \beta_3 \times \text{Party performance} + \beta_4 \times \text{Incumbent Party} \\ &+ \beta_5 \times \text{Distance from contention} + \varepsilon\end{aligned}$$

From our table above, we get the following values for our coefficients:

$$\begin{aligned}\text{Vote share} =& 0.281 + (-450) \times \text{Woman} + .276 \times \text{Vote share lag} \\ &+ .666 \times \text{Party performance} + 6.783 \times \text{Incumbent Party} \\ &+ (-.015) \times \text{Distance from contention} + \varepsilon\end{aligned}$$

3. DV dichotomous

The impact of voting for Trump over Clinton in the 2016 election. Ordinary least squares (OLS) and Logit regression models with voting for Trump as the dependent variable.

	(Model 1)	(Model 2)
	OLS	Logit
Party	0.163***	1.020***
	(0.004)	(0.048)
Female	-0.066***	-0.551***
	(0.018)	(0.165)
Income	-0.007	-0.108
	(0.020)	(0.179)
65 or over	-0.065*	-0.594*
	(0.029)	(0.254)
Latinx	-0.104***	-0.805**
	(0.028)	(0.246)
Black	-0.132***	-1.426***
	(0.028)	(0.315)
Other	-0.078*	-0.569*
	(0.034)	(0.278)
Education	-0.059**	-0.524**
	(0.019)	(0.178)
Constant	-0.047*	-3.181***
	(0.024)	(0.214)
R^2	0.54	
$Pseudo\ R^2$		0.48
Observations/N	1563	1563

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Source: Dean Lacy's Course on Maximum Likelihood Estimation for General Linear Models at ICPSR

The N is the the number of respondents in this survey.

In both the OLS and the logit model, the dependent variable is whether the respondent voted for Trump. Because this DV takes only a 0 or 1 (0 = Vote for Clinton, 1 = Vote for Trump), sometimes researchers use a Logit model instead of OLS (we will come back to this).

The IV's:

Party is the respondent's party identification on a 7-point scale, from 1= strong Democrat to 7=strong Republican.

Female is a binary variable (another way of saying this is a dummy variable) which is 0 if the respondent is male and 1 if the respondent is a female.

65 or over is also a dummy variable where 0 = under 65 and 1 = 65 and over.

Latinx, Black and Other are all dummy variables. The reference category (0) is White people.

Education is a dummy variable. B.A. or higher = 1, everyone else = 0.

Note when the DV is dichotomous and its an OLS model, the interpretation is in terms of probability.

Here is the equation for model 1:

$$\begin{aligned}\text{Vote for Trump} = {} & \beta_0 + \beta_1 \times \text{Party} + \beta_2 \times \text{Female} \\ & + \beta_3 \times \text{Income} + \beta_4 \times \text{65 or over} \\ & + \beta_5 \times \text{Latinx} + \beta_6 \times \text{Black} \\ & + \beta_7 \times \text{Other} + \beta_8 \times \text{Education} + \varepsilon\end{aligned}$$

From our table above, we get the following values for our coefficients:

$$\begin{aligned}\text{Vote for Trump} = {} & -0.047 + 0.163 \times \text{Party} + (-0.066) \times \text{Female} \\ & + (-0.007) \times \text{Income} + (-0.065) \times \text{65 or over} \\ & + (-0.104) \times \text{Latinx} + (-0.132) \times \text{Black} \\ & + (-0.078) \times \text{Other} + (-0.059) \times \text{Education} + \varepsilon\end{aligned}$$

For example, in model 1, the probability of voting for Trump is 7 points lower among women than among men. The probability of voting for Trump increases by 16 points for every point shift towards Republicans...

If it's a Logit model (model 2), we cannot interpret the Logit coefficients directly from the table. The assessment of the relationship between our independent variable and the dependent variable proceeds with a discussion of the significance and direction of the model coefficients or odds ratio (for more information on the latter, consult this site). Sometimes researchers will show marginal effects to get around this limitation.

Let's interpret the logit model with significance and direction of the model coefficients:

Party is positive and significant which means that Republicans are more likely to vote for Trump.

The coefficient for female is negative and significant which means women are less likely to vote for Trump than men.

65 or over is negative and significant which means that voters that are 65 and over are less likely to vote for Trump than voters below 65.

The dummy variables for race (Latinx, Black and Other) are all negative and significant which means that compared to white people, they are less likely to vote for Trump.

This effect seems to be most pronounced among black voters, followed by latin and the other category.

The coefficient on the education variable is negative and significant which means that voters with a BA or higher are less likely to vote for Trump.

Note: While the coefficients in model 1 (OLS) and 2 (Logit) change, the statistical significance for each of the independent variables in both models remain the same. This suggests that we can simply use OLS for dichotomous DV variables to make the interpretation easier.

4. Interactions: DV and IV dichotomous

OLS Interaction Between Partisanship and Political Knowledge on Pro-Choice Attitudes

	(Model 1)	(Model 2)
Partisanship	-0.270***	-0.202***
	(0.016)	(0.021)
Political Knowledge	0.137***	0.200***
	(0.016)	(0.020)
Partisanship × Political Knowledge		-0.152***
		(0.032)
Constant	0.642***	0.618***
	(0.012)	(0.013)
Observations	3715	3715
R^2	0.083	0.089

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Source: ANES, 2012

The N in this dataset are respondents in the ANES in 2012.

DV: In favour of abortion = 1, Opposition of abortion = 0. So higher values = pro-choice.

Partisanship (0 = Democrats, 1 = Republican). Independents are not included.

Political Knowledge (0 = Doesn't know Speaker, 1 = Knows Speaker).

Let's start with model 1.

Here is the equation for model 1:

$$\text{Pro-choice} = \beta_0 + \beta_1 \times \text{Partisanship} + \beta_2 \times \text{Political Knowledge} + \varepsilon$$

From our table above, we get the following values for our coefficients:

$$\text{Pro-choice} = 0.642 + (-0.270) \times \text{Partisanship} + 0.137 \times \text{Political Knowledge} + \varepsilon$$

We see that the coefficient for Partisanship is negative and significant. This means that the probability of Republicans supporting pro-choice is 27 percentage points compared to Democrats. The coefficient for Political Knowledge shows that the probability of respondents with political knowledge being pro-choice is 14 percentage points higher than respondents without political knowledge.

For model 2, note there is an interaction so we cannot interpret the coefficients directly. In general, an interaction effect exists when the impact of one independent variable (x_1) depends on the value of another independent variable (x_2).

Here is the equation for model 2:

$$\begin{aligned}\text{Pro-choice} =& \beta_0 + \beta_1 \times \text{Partisanship} + \beta_2 \times \text{Political Knowledge} \\ &+ \beta_{12} \times \text{Partisanship} \times \text{Political Knowledge} + \varepsilon\end{aligned}$$

From our table above, we get the following values for our coefficients:

$$\begin{aligned}\text{Pro-choice} =& 0.618 + (-0.202) \times \text{Partisanship} + 0.200 \times \text{Political Knowledge} \\ &+ (-0.152) \times \text{Partisanship} \times \text{Political Knowledge} + \varepsilon\end{aligned}$$

The Partisanship coefficient can be interpreted as the effect of Partisanship when Political Knowledge = 0.

The Political Knowledge coefficient can be interpreted as the effect of Political Knowledge when Partisanship variable = 0 (among Democrats).

The intercept (0.618) is the probability of being pro-choice among low knowledge Democrats.

The interaction shows the moderation effect of Political Knowledge on the relationship between Partisanship and pro-choice. But in order to understand the interaction, we need to do some calculations.

In what follows, Party is Partisanship and Knowledge is Political Knowledge. This is simply abbreviated to fit the equation.

Here is the equation for model 2 again:

$$\text{Pro-choice} = \beta_0 + \beta_1 \times \text{Party} + \beta_2 \times \text{Knowledge} + \beta_{12} \times \text{Party} \times \text{Knowledge} + \varepsilon$$

Let's look at this in more detail (these are the models predicted y values):

y (the probability of being pro-choice) for low knowledge (knowledge = 0) Democrats (Party = 0) is:

$$0.618 = 0.618 + (-0.202) \times 0 + 0.200 \times 0 + (-0.152) \times 0 \times 0$$

y (the probability of being pro-choice) for high knowledge (knowledge = 1) Democrats (Party = 0) is:

$$0.818 = 0.618 + (-0.202) \times 0 + 0.200 \times 1 + (-0.152) \times 0 \times 1$$

y (the probability of being pro-choice) for low knowledge (knowledge = 0) Republicans (Party = 1) is:

$$0.416 = 0.618 + (-0.202) \times 1 + 0.200 \times 0 + (-0.152) \times 1 \times 0$$

y (the probability of being pro-choice) for high knowledge (knowledge = 1) Republicans (Party = 1) is:

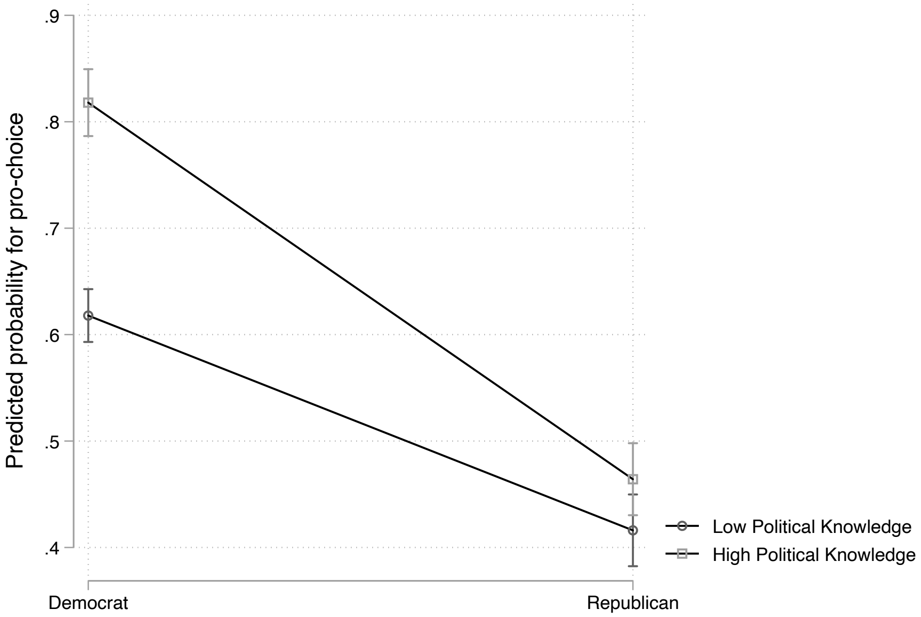
$$0.464 = 0.618 + (-0.202) \times 1 + 0.200 \times 1 + (-0.152) \times 1 \times 1$$

As you can see, there is a big effect for Democrats but not really a difference for Republicans. This means that knowledge has a big effect among Democrats and only a small impact among Republicans. Among Democrats, knowledge increases the probability of being pro-choice by 20 percentage points (from 0.618 to .818) while it increases that probability by only 5 percentage points among Republicans (from 0.416 to 0.464). The impact of knowledge varies across parties. Knowledge really makes a difference only among Democrats.

The calculation we did above could also be presented in a figure.

Here is an example:

Figure 1: Marginal Effect Plot By Partisanship


Predicted probability for pro-choice
Democrat
Republican
Low Political Knowledge
High Political Knowledge