GV101 Intro to Polsci

Professor Simon Hix

Regression Revision Document

Cedric Tan

May 2019

Abstract

This is a review of regression tables and how to interpret them from Simon Hix's lectures in the academic year of 2018-2019. The notes are mine fully and may not be authentic to the lecturer's as they have been modified.

The format of this material is taken from lectures where necessary and classes with Dr Anastasia Ershova.

Contents

1 Overview of Regression

Regressions can be run for several purposes including:

1. To give a **descriptive summary** og how the outcome varies with the explanatory variables
2. To **predict the outcome** given a set of values for the explanatory variables
3. To **estimate the parameters of a model** describing a process that generates the outcome
4. To **study causal relationships** that could be explained by the model

1.1 Descriptive Summary

Through regressions such as Ordinary Least Squares (OLS), we achieve the best-fitting linear relationship, where best is defined as minimizing the sum of squares of the residuals i.e. difference between predicted and real values.

OLS estimates the best-fitting line in the population. However, the summary provided by OLS may miss important features of the data, such as outliers or non-linear relationships.

1.2 Prediction

OLS regression gives the best linear predictor in the sample. If the sample is drawn randomly from a larger population, OLS is a consistent estimator of the population's best linear predictor.

1.3 Estimation and Causality

Estimating the parameters of a model is the purpose that receives the most discussion in traditional textbooks. However, causality is the real motivation for regression. This is causal inference.

2 Basic Terms for Regression

Below are some basic terms for regression that one should memorise to use at all times:

- **Direction:** this is the way the effect is going, either positive or negative which corresponds to an increase or decrease
- **Neutral Language:** ensure to use this language if you are not confident in the units that are associated with the change and the direction of the change. For example: *there is a unit increase in Y due to X*
- **Controls:** these are effects that have influence on other variables which can ultimately affect the magnitude and direction of the dependent variables
- R^2**:** this is a measurement of how much variance is explained by the dependent variables in the model

- **Variance:** helps understand the differences in the level of dependent variables that occur depending on other variables
- **Confidence:** this explains how much faith you have in the fact that something occurred by chance or is truly explained by the model constructed
- **Omission:** this is when a model has missed out crucial variables which might make insights biased one way or another

3 Language use for Regressions

When explaining certain parts of a regression table, there needs to be a critical understanding of the language used to answer how the table works. Below are these critical phrases to keep in mind:

- When testing for a single variable: remember that the answer given in the regression output is while **holding all other variables constant.**
- When comparing models, recognise that **bias might be present due to omission** which is a key factor when seeing the difference in the amount of variables present in one model over another
- Underspecification which means it suffers from omitted variable bias
- Absorption means taking in the effect of another variable which is highly correlated. When adding one variable, the other may become insignificant as a result of this absorption

4 Full Examples

4.1 Example 1: Unemployment and Authoritarianism

Example 1:

Variables	Model 1	Model 2
Authoritarianism		0.46
		(0.23)**
Crisis	0.87*	0.88
	(0.51)	(0.54)
Inflation		-0.50**
		(0.02)
GDP Per Capita	-0.001**	-0.001***
	(0.00)	(0.00)
Constant	9.12	9.10
R Squared	0.12	0.20

*p>0.90; **p>0.95; ***p>0.99

Dependent Variable: Level of Unemployment
Independent Variable: Authoritarian Govt. (1 = Authoritarian)
Controls:

- Crisis (1 = year of crisis starting)
- Inflation (%)
- GDP per Capita (USD)

1. What is the effect of GDP on the level of unemployment in Model 1?
2. How does the effect of Crisis changes from Model 1 to Model 2? Why?
3. What does the change in R squared indicate?
4. How do the changes in the models affect our understanding of the results?

Question 1:
An increase in GDP by one US Dollar would lead to a 0.001 unit decrease, i.e. have a negative effect, in unemployment holding all other control variables constant. This result is statistically significant to the 95th percentile.

Question 2:
Crisis has a positive effect on the level of unemployment with a magnitude of positive 0.87 units per unit of crisis increasing. It is statistically significant to the 90th percentile. However, in model 2, the crisis effect becomes not statistically significant at all despite it increasing in magnitude by 0.01. This may be due to the omitted variable bias where the effect of inflation may have absorbed the effect of crisis due to their high correlation with inflation being a more influential factor to unemployment.

Question 3:
R squared measures the amount of variance which is captured by the dependent variables in the model. The change in R squared shows that the second model, which

captures 0.08 units more (8%), accounts for more of the variance associated with unemployment meaning it has more explanatory power than the first model. This might be due to the introduction of the new variables in the second model

Question 4:
The addition of authoritarianism

4.2 Example 2: Political Activity and Factors

Example 2:

Variables	Model 1	Model 2
Constant	-1.14	-0.93
	(0.21)	(0.24)
Education	0.37**	0.24**
	(0.02)	(0.03)
Job Level	0.07**	0.00
	(0.02)	(0.02)
Family Income		0.07**
		(0.01)
Job Skills		0.12**
		(0.03)
Church Skills		0.19**
		(0.03)
N	2489	2415
R Squared	0.29	0.33

*p>0.95; **p>0.99

Dependent Variable:

- Political Activity

Independent Variables:

- Education: level of qualification
- Job Level: level of training
- Family Income: $15,000 intervals
- Job Skills: number of specific skills
- Church Skills: specific skills for voluntary work

1. What is the estimated effect of an additional level of educational qualification on the number of political activities undertaken by respondents? Show your calculation based on Model 1.
2. What does the direction of the coefficient for family income indicate?
3. Does the main result from Model 1 change with the addition of family income, job skills and church skills in Model 2? What, if anything, does this suggest?
4. How does the simultaneous addition of family income and job skills in Model 2 affect our interpretation of the change in the coefficient for job level between Model 1 and Model 2?
5. What does the change in the size of the R-squared between Model 1 and Model 2 indicate? What could be causing this change?

Question 1:
Knowing that the regression equation is:

$$\begin{gathered} \textit{Political Activity} = \beta + \beta_1(\textit{Education}) + \beta_2(\textit{Job Level}) + \epsilon \\ \textit{Political Activity} = -1.14 + .37(\textit{Education}) + 0.07(\textit{Job Level}) \end{gathered}$$

We can set education to 0 and 1 and then see the difference to see what effect a unit increase in education will have on the output such that:

$$\begin{gathered} [-1.14 + .37 \times 0 + 0.07(\textit{Job Level})] - [-1.14 + .37 \times 1 + 0.07(\textit{Job Level})] \\ [-1.14 - (-0.77)] = .37 \end{gathered}$$

Therefore, based on Model 1, the estimated effect of an additional level of education qualification, when job level is held constant, is to raise the number of political activities undertaken by the respondent by .37 units.

Alternatively, plug in the right coefficients to show that it is a 0.37 unit increase.

Question 2:
From the results in Model 2, the direction of the coefficient for Family Income is positive. This indicates that, when all other independent variables in Model 2 are held constant, an increase in family income is associated with an increase in the number of political activities undertaken by respondents.

Question 3:
The main result from Model 1 is the estimated effect of Education and Job level on the number of political activities undertaken by respondents.

The results suggest that with the addition of Family Income, Job Skills and Church Skills in Model 2, the estimated effect of Job Level has been changed. The size of the coefficient for Job Level decreases from 0.07 to 0.00 whilst standard errors remain the same. Therefore, the level of significance of the coefficient for Job Level is greatly reduced. This, in turn, implies that we are no longer certain that variation in Job Level has a positive effect on the level of political activities as suggested by Model 1.

There are also minor changes to the estimated effect for Education with the addition of new variables in Model 2. The size of the coefficient is reduced, which means the magnitude of the estimated effect of Education on political participation is slightly smaller. Nevertheless, the sign of the coefficient remains the same and the significance level still shows it is statistically significant to the 99th percentile. Hence, even with the addition of new variables, we remain confident about the positive association between Education and political participation observed in Model 1.

The fact that the coefficient associated with Job Skills ceases to be statistically significant in model 2 suggests that the addition of Family Income, Job Skills and Church Skills captures the important elements of the effect of Job Level. In other words, it seems that the positive association between Job Level and Political Activity is, in fact, explained by the income and skills that people hold, which are also likely to be associated with their Job Level i.e. it is absorbed by these other factors.

Question 4:
The change in estimated effect of Job Level leads us to question the validity of the relationship between Job Level and political participation identified in Model 1. This

is because Model 2 shows that once Family Income and Job Skills are added, the coefficient for Job Level loses its significance. A possible explanation for this would be that Job Level is likely to be correlated with both Family Income and Job Skills, and Model 2 shows that variations in the latter two variables are better able to account for respondents' political activities. This also shows that Model 1 probably suffers from **omitted variable bias.**

The simultaneous addition of both Family Income and Job Skills, along with Church Skills, in Model 2 means that we cannot be sure which of their relationships with Political Activities removes the significance of the relationship between Job Level and Political Activities. In other words, we are not sure whether the apparent relationship between Job Level and Political Activity in Model 1 is actually to do with respondents' incomes, job skills or even church skills, all of which may plausibly be correlated with Job Level. In order to determine which of the new variables or combinations of variables in Model 2 captures the apparent relationship between Job Level and Political Activity in Model 1, we would need to add each of them in turn.

Question 5:
The size of R^2 increased from 0.29 in Model 1 to 0.33 in Model 2. This change indicates that Model 2 has higher explanatory power as it is able to account for 4% more of the variation in the dependent variable. This change is likely to be caused by the increase of information into the model with the addition of three new variables.