Some notes on presenting and interpreting regressions

Johan A. Elkink

February 3, 2013

Comments or suggestions welcome

Tabular presentation

Good examples of regression tables can easily be found in current political science journals. An example is also provided in Table 1, which is based on two multiple linear regressions. This is my preferred style, but variations are possible and common. Regression tables are generally better formatted and presented than how they come out of a statistical package and it is therefore not advisable to simply copy-and-paste from R, SPSS, Stata or any other package. A number of issues arise when presenting regression tables like this.

1. Make sure the variable names are clear descriptions of the variables, not the abbreviated names used in the software package used for the estimation. For example, in most packages spaces are not allowed in variable names, but in printed regression tables, this is perfectly normal.

2. In addition to the regression coefficients, always include either standard errors or t-values (z-values in the case of output from a Maximum Likelihood estimation).

3. Typically, add some indication of statistical significance, either by adding stars or by including p-values.

4. Always include R^2 or some alternative measure of fit and for Maximum Likelihood estimates, include the log-likelihood.

5. Always include the sample size after missing cases are removed or imputed.

6. When there are categorical variables included, it helps to name the variables after the relevant category - e.g. "female" is more informative than "sex", if it is coded such that females are 1 and males 0.

7. Round numbers to some reasonable number of digits. Many digits suggests a level of precision that is utterly unrealistic given the lack of precision and accuracy of most measures we use to generate the data in the first place. In the natural sciences people adhere to strict rules about "significant figures", with their specific rules for arithmetic,[1] but even without applying such rules, it would be reasonable to be more reluctant to suggest high precision. As a very rough rule of thumb, two digits after the decimal point is plenty.

8. Sometimes, this kind of rounding means that all interesting digits disappear. For example, if $\beta = 0.00032$ with $\sigma_\beta =$

[1] See `http://en.wikipedia.org/wiki/Significance_arithmetic`.

	Model 1		Model 2	
Movie rating	0.75	*	1.10	**
(scale 1-4)	(0.39)		(0.35)	
Cast size	0.53	**	0.42	**
(scale 3-13)	(0.14)		(0.12)	
Year	–		0.07	**
(scale '24-'95)			(0.01)	
Intercept	4.66	**	0.29	
	(1.12)		(1.31)	
Adjusted R^2	0.19		0.36	
N	100		100	

Standard errors in parentheses. * $\alpha = 0.10$; ** $\alpha = 0.05$.

Table 1: Linear regression models explaining the length of movie reviews – measured as lines of text ranging from 5 to 21 lines – in the Leonard Malvin's Movie and Video Guide (1996).

0.00011, you don't want to be writing $0.00(0.00)$ in your table. Your statistical software might suggest some scientific notation, such as `3.2e-04`, but this looks awkward in a regression table. If you want, you could include this as 3.2×10^{-4}, but it is still cumbersome to read. Better would be to divide your independent variable by, say, 1000, such as to rescale the variable, and get coefficients that are much more interpretable. You would end up with $0.32(0.11)$, which looks fine.

9. For reasons of comparison, it might be a good idea to standardize your variables first before performing the regression. You can do this by subtracting the mean and dividing by the standard deviation, or, as Gelman (2007) suggests, dividing by two standard deviations. Note that this changes what "one unit change in x" means, so if affects the substantive interpretation of the regression coefficients.

10. In the example in Table 1, I included brief descriptions of the measurement of the independent variables. I think this is very useful in regression output to allow the reader to be able to interpret the size of the coefficients, but it is not actually common practice. Journal reviewers might be less enthousiast about this than I am.

Graphical presentation

An alternative to presenting a regression table is to present the regression coefficients in graphical form (Gelman, Pasarica and Dodhia, 2002; Kastellec and Leoni, 2007).[2] This is a much easier communication device, but some readers will appreciate if the detailed regression table is still available in the appendix.

In particular when including interactions of quadratic forms, a graphical presentation of the main effect might also be very useful. Ideally, this would be accompanied by the appropriate confidence intervals (King, 1998).

Interpretation

In a linear model, the β-coefficients represent the derivative of $\mathbf{y}$ towards $\mathbf{X}$. Hence the formulation "β represents the value by which y

[2]See also `http://www.stat.columbia.edu/~gelman/research/published/tables4.pdf`.

changes as x changes by one unit, everything else constant", and the $\hat{\beta}$'s printed in the regression table are the estimates of these β's. This is correct, but often not the most useful formulation:

- when x is a dummy variable, taking on the values 0 or 1, then the coefficient represents the difference between those two groups. E.g. if we have a variable `female` which is 1 for females and 0 for males, then $\beta_{\texttt{female}}$ represents how much higher females score on y than males;
- when x represents a proportion, a "one unit change" means a change from 0% to 100%. Hence, it makes more sense to say explicitly that you are referring to this all-or-nothing change. In many cases, it is more informative to talk about the effect of a small change, for example from 10% to 20%, thus to divide the β by 10 and report that as an substantive finding;
- when x represents a scale from a survey, the unit is often rather arbitrary and it is doubtful whether anyone would at all be interested in the effect on y of a one unit change. Instead, you could think about a change from the mean score to one standard deviation higher;
- when x and/or y are on a logarithmic scale, this affects the interpretation. E.g. if x is on a log scale, then β reflects the effect of a 1% increase in x on y.

More in general it holds that you should always think about what exactly the units on x and y represent and formulate the effect accordingly. Furthermore, it is generally useful to look at the distribution of x first, such that the reported effect is within a reasonable range of x. If all observed xs are, say, between 0.3 and 0.4, then a "one unit change" is not an interesting quantity.

For example, when we look at Model 2 in Table 1, we could say that when the cast size of a movie increases by 10, we estimate the length of the movie review to be 10.6 lines longer. Any rating point higher implies a movie review of about 1.1 extra lines.

Sometimes it might be the case that particular scenarios are of interest (e.g. the voting behaviour of an older, loyal Fianna Fail voter versus a younger Green voter) and might be worth presenting. The idea to look at considered "quantities of interest" is defended extensively by King (1998) and King, Tomz and Wittenberg (2000). While they refer primarily to categorical dependent variable models, the point holds in general.

References

Gelman, Andrew. 2007. "Scaling regression inputs by dividing by two standard deviations." *Statistics in Medicine* 27(15):2865–2873.

Gelman, Andrew, C. Pasarica and R. Dodhia. 2002. "Let's practice what we preach: Turning tables into graphs." *American Statistician* 56:121–130.

Kastellec, Jonathan P. and Eduardo L. Leoni. 2007. "Using graphs instead of tables in political science." *Perspectives on Politics* 5(4).
URL: *http://www.tables2graphs.com*

King, Gary. 1998. *Unifying political methodology. The likelihood theory of statistical inference*. University of Michigan Press.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the most of statistical analyses: improving interpretation and presentation." *American Journal of Political Science* 44(2):341–355.