Tutorial on centering

Constantin Manuel Bosancianu[1]
February 16, 2019

This will be a very short tutorial on how centering impacts your regression coefficients, as well as your interpretation of these regression coefficients. It's important to mention that these insights only when you do standard centering on cross-sectional data. There is also a large literature on centering in multilevel models (Kreft, de Leeuw, & Aiken, 1995; Paccagnella, 2006; Enders & Tofighi, 2007), but for these types of data structures my insights here will not fully hold.

The variables I use comes from the Quality of Government data set, January 2018 version. The data comes from and around 2014. If no data exists for 2014 for a particular country, 2015 is used; if 2015 is also missing, 2013 is used. This choice rule is used up to $+/-$ 3 years. As a dependent variable, I use the World Justice Project's *Rule of Law Index*. The variable is named `wjp_overall`, and ranges from 0 (lowest possible extent of rule of law) to 1 (highest possible extent of rule of law). In practice, though, the minimum value is 0.319 (Venezuela), while the maximum is 0.87 (Denmark). As an independent variable, I use the estimate of a country's level of income inequality, obtained from the World Bank's *World Development Indicators* (the name of the variable is `wdi_gini`). This ranges from 0 (perfect equality) to 100 (perfect inequality). Although this is only a demonstration, existing theoretical models have linked the two phenomena, via the increased incentives businessmen have to prevent efficient tax collection through subversion of political institutions and the courts (Glaeser, Scheinkman, & Shleifer, 2003).

An initial examination of a simple regression of rule of law on income inequality finds support for our presumed link (see Table 1). A 1-unit increase in income inequality (measured on a 0–100 scale) is associated with a 0.007 decrease in rule of law. To put it in other terms, a 10-point increase in inequality is associated with a 0.07 decrease in rule of law. Given that rule of law only varies in practice between 0.32 and 0.87, a 0.07 effect seems moderate to high.

What about the intercept? Here things are a bit trickier, since inequality is never observed to be below about 23–24 points. This means that we would predict that the level of rule of law in a country with 0 inequality is 0.817. This is easiest to see in Figure 1, where the regression line intersects the vertical axis at a little bit above 0.8.

The problem is that this interpretation makes little sense. We have never seen a society have a value of Gini of 0, and it is likely that no such society ever existed. Even with Romania's brutal Communism, the level of income inequality in 1990 was only 22.7. This makes our interpretation of the intercept slightly awkward. A solution to this, as I mentioned during the lecture, would be centering the predictor in our regression. We do this by subtracting from each country's value of Gini, the average value for Gini in our sample: 37.41772.

You can see, by comparing the results presented in Table 2 with those of Table 2, that the centering does not change much. The slope value is the same, as is the R^2, adjusted R^2, and RMSE. The only

Table 1: Results from simple regression

	DF: Rule of law
(Intercept)	0.817***
	(0.066)
Income inequality	−0.007***
	(0.002)
R^2	0.157
Adj. R^2	0.146
Num. obs.	79

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

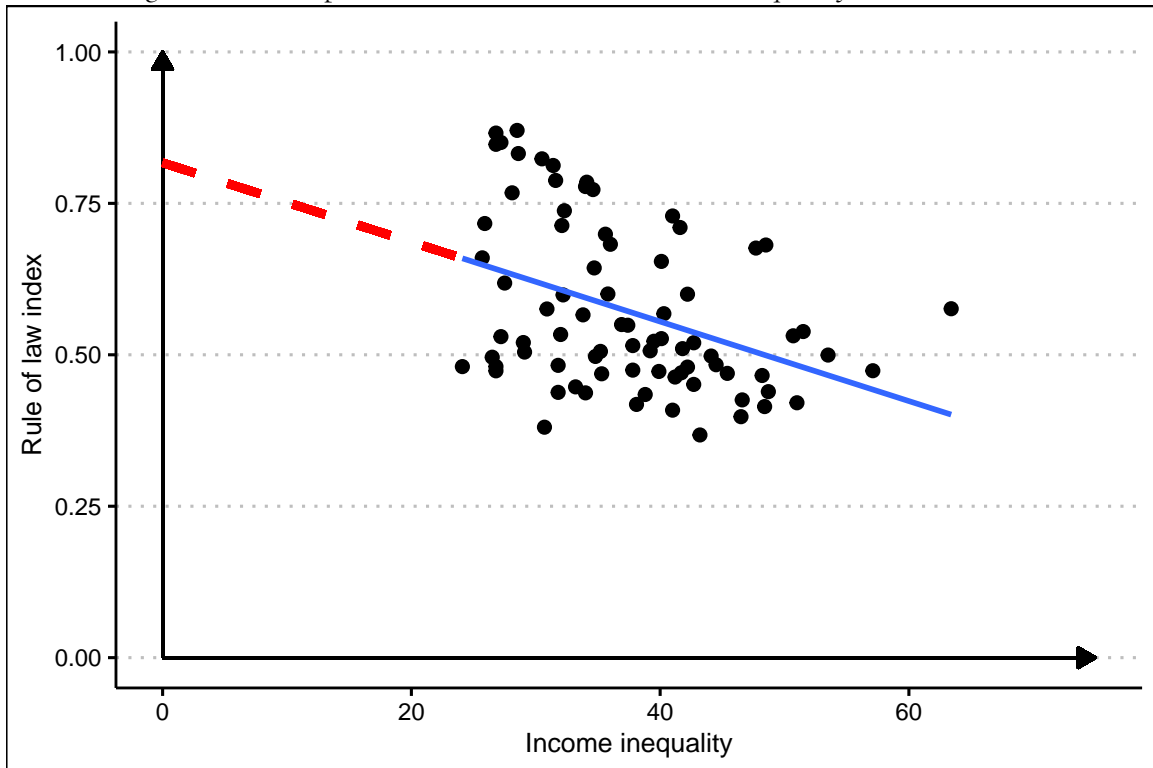Figure 1: Scatterplot of the link between income inequality and rule of law

Table 2: Results from simple regression (centered predictor)

	DF: Rule of law
(Intercept)	0.572***
	(0.014)
Income inequality (centered)	−0.007***
	(0.002)
R^2	0.157
Adj. R^2	0.146
Num. obs.	79

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

thing that changes is the value of the intercept; this is because it now refers to a different quantity. In Figure 1 it was the expected level of rule of law at 0 income inequality. Now it becomes the expected level of rule of law at 37.41772 income inequality (which is 0 on the centered version of the income inequality indicator). This can be seen in Figure 2, where the entire horizontal axis was shifted to the left by 37.41772 points now, bringing the scatterplot directly over the vertical axis. In this version, the regression line intersects the vertical axis at exactly 0.572, which is what our intercept is telling us.

In a sense, centering a predictor in this cross-sectional data structure has only benefits: it keeps identical what we would like to preserve the same (slope coefficients, model fit statistics) and changes what we would prefer to have changed (the awkward interpretation of the intercept). In the second regression, the intercept is interpreted as the expected level of rule of law for a level of income inequality that is common in our data: the mean level.

References

Enders, C. K., & Tofighi, D. (2007). Centering Predictor Variables in Cross-Sectional Multilevel Models: A New Look at an Old Issue. *Psychological Methods*, *12*(2), 121–138.

Glaeser, E., Scheinkman, J., & Shleifer, A. (2003). The Injustice of Inequality. *Journal of Monetary Economics*, *50*(1), 199–222.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The Effect of Different Forms of Centering in Hierarchical Linear Models. *Multivariate Behavioral Research*, *30*(1), 1–21.

Paccagnella, O. (2006). Centering or Not Centering in Multilevel Models? The Role of the Group Mean and the Assessment of Group Effects. *Evaluation Review*, *30*(1), 66–85.

Figure 2: Scatterplot with centered inequality and rule of law