# Handout for PS104A
# Understanding regressions in political science journals

In many upper division classes you'll be expected to read political science journal articles. This is also the case in many other upper division social science classes, and most of what is written here will apply there as well. A great deal of social science research relies on quantitative methods, and the most common quantitative method is OLS (ordinary least squares). This handout goes over how to interpret regression output as it is usually presented in political science journals.

The attached page has a table presenting a regression estimated to examine a political science research question. This particular table came out of a recent edition of the Journal of Politics, and this article is examining what factors lead African-Americans to identify with different types of "black nationalism." Two types of nationalism are examined, "community nationalism", where African-Americans feel they must stick together and build a strong community, and "separatist nationalism", where African-Americans feel that as a group they should avoid contact with white Americans.

Two regression models were estimated, with the measure of community nationalism as one dependent variable, and the measure of separatist nationalism as the other dependent variable. The dependent variable is the thing we're trying to explain in the regression. Listed along the side are the independent variables, or the variables we're using to try to explain the dependent variable. We haven't talked about this yet, but regression models can have multiple independent variables --- this is known as multiple regression. In other words, we've seen regression formulas like: $y = a + bx$, where y is the dependent variable and x is the independent variable --- b is the slope coefficient. Multiple regression would estimate: $y = a + bx + cz + dm \ldots$, where y is the dependent variable and x, z, and m are the independent variables, and b, c, and d are slope coefficients along different dimensions.

For each independent variable in each model, the authors list the slope coefficient. A positive coefficient indicates this independent variable has a positive effect on the dependent variable **in the sample**, while a negative coefficient indicates this independent variable has a negative effect on the dependent variable **in the sample**. The standard error (standard deviation) of the coefficient estimate is listed in parentheses below the coefficient.

Of course, what we really care about is how these independent variables affect the dependent variable **in the population**. In order to do that we need to do a hypothesis test. We use a t-test, since we don't actually observe the variation of the sampling distribution. In a regression our null hypothesis for the slope coefficients is that they're equal to zero --- if a slope coefficient is equal to zero that independent variable has no effect on the dependent variable. If we can reject this null hypothesis we would conclude this independent variable does have an effect on the dependent variable.

The stars over each coefficient estimate tell us the result of a t-test on that coefficient with a null hypothesis of zero. If there are no stars, the authors couldn't reject the null hypothesis of zero, and this independent variable is assumed to have no effect. Stars indicate we can reject the null hypothesis of zero at some commonly accepted level of significance. The key at the bottom of the table tells us what each number of stars means. For example, one star gives us a significance level of 5%. That is, coefficients

with one star over them are pretty far from zero in terms of standard deviations --- in fact, this estimate is out in the last 5% of the tail of the distribution, making the null hypothesis of zero unlikely. More stars mean that estimate is even further out --- 2 stars means the estimate is in the last 1% of the tail of a distribution centered at 0, 3 means out in the last 1/2%, and 4 means out in the last 1/10%. Generally, more stars means our null hypothesis is more unlikely.

Now you can quickly look this table over, and see what effect each independent variable had on the dependent variable (positive or negative), and see if that effect was **statistically significant**, meaning we think it is likely that this effect we found in the sample also holds true in the population. More stars means it's more likely the effect we estimated is a true effect in the population.

TABLE 2

## OLS Estimates of Predictors on Community and Separatist Nationalism Factors in the 1993 National Black Politics Study

| Independent Variables | (1) Community Nationalism Factor | (2) Separatist Nationalism Factor |
|---|---|---|
| Linked fate | .37**** | .07 |
| | (.10) | (.10) |
| Family income | .28* | −.64**** |
| | (.12) | (.12) |
| Years of education | 1.45 | −2.70* |
| | (1.15) | (1.15) |
| Age | −.06** | −.04* |
| | (.02) | (.02) |
| Female | −.04 | −.22*** |
| | (.07) | (.07) |
| Respondent from South | −.15* | −.12 |
| | (.07) | (.07) |
| Liberal-conservative ideology | .20* | −.18 |
| (1 = strong liberal) | (.10) | (.10) |
| Member of a black organization | .27*** | −.21* |
| | (.08) | (.08) |
| Perceived race of interviewer | −.23** | −.08 |
| (1 = white or other) | (.09) | (.09) |
| Constant | −.54*** | 1.01**** |
| | (.17) | (.17) |
| Adjusted $R^2$ | .10 | .09 |
| Standard error of the estimate | .95 | .95 |
| Number of cases | 749 | 749 |

Entries are unstandardized regression estimates with standard errors in parentheses. Dependent variables in Models 2, 3, 4, and 5 have been recoded on a 0–1 interval, with 0 indicating strong disagreement and 1 indicating strong agreement. All independent variables were recoded on a 0–1 interval, unless indicated otherwise. Values are rounded to 0.01, with a few exceptions. All tests are two-tailed. * indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .005$. **** indicates $p < .001$.