

## EXPLORING THE ANGRIST-KRUEGER RETURNS TO EDUCATION DATA

Angrist and Krueger (1991) Use a data set with over 300,000 observations to measure returns to education, using quarter of birth as an instrument to eliminate selection bias in the measurement of years of school completed. They stick to a linear model, estimate a very small number of parameters, and do not examine the error distributions in their model. We will use the same data set to estimate richer models, allowing us to get additional economically interesting insights. The data will be on the course web site both as a text file and as an R data frame. In R, much of the work below is easy if you use R's `anova` command on the output of its `lm` command. You will also want to use R's ability to treat variables as factors, perhaps by creating two versions of the data frame or by creating two versions of the variables other than the wage, by using the R `as.factor` command.

- (1) Estimate by OLS a regression of log wage on years of education and year of birth, treating both of these right-hand-side variables as numeric. Compare your estimates to those in the AK paper. Are both variables "significant" by both the usual  $F$  test and by the BIC criterion?
- (2) Repeat the previous question, but now treating education and year of birth as factors (i.e., creating a separate dummy variable for each value of the education variable and for each value of the year of birth variable, though R's `lm()` does the automatically for variables that are R factors). Does it look as though the size of the years of schooling effect is much different with this specification? Is there "eyeball" evidence against the assumption of a linear years-of-schooling effect? What about linearity of the year-of-birth effect?
- (3) Estimate a model with education, year of birth, and place of birth all entered as factors, and in addition an education:place-of-birth interaction. Are place of birth and the interaction terms each significantly non-zero as a group at conventional levels? What about by the BIC criterion? What is the economic interpretation of the question of whether education:place-of-birth interactions are significant?
- (4) Plot histograms and normal qqplots (in R, using `hist()` and `qqnorm`, perhaps plus `qqline`) of the full list of residuals in the model from part 3. (In R, retrieve them with `residuals(lmout)`.)
- (5) Show 3 separate normal qqplots for residuals in observations with 0, 12, and 20 years of education. Calculate the sample variances of residuals separately for observations in each of the 21 education categories. Do they show a pattern of variation?

### REFERENCES

ANGRIST, J. D., AND A. B. KRUEGER (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 106(4), 979–1014.

---

*Date:* February 15, 2018.

©2018 by Christopher A. Sims. ©2018. This document is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

<http://creativecommons.org/licenses/by-nc-sa/3.0/>