Professor Raj Chetty and Dr. Gregory Bruich
Department of Economics, Harvard University

Empirical Project 1

Stories from the Atlas: Describing Data using Maps, Regressions, and Correlations

Posted on Thursday, February 7, 2019

Due at midnight on Thursday, February 21, 2019

The Opportunity Atlas was publicly released on October 1, 2018, and an accompanying article appeared on the front page of the *New York Times*. The Opportunity Atlas is a freely available interactive mapping tool that traces the roots of outcomes such as poverty and incarceration back to the neighborhoods in which children grew up.

Policymakers, journalists, and the public have begun to explore the Opportunity Atlas, casting new light on the geography of upward mobility in communities across the country. As an example, see Jasmine Garsd's recent analysis for the New York City neighborhood of Brownsville in Brooklyn.

In this first empirical project, you will use the Opportunity Atlas mapping tool and the underlying data to describe equality of opportunity in your hometown and across the United States. (If you grew up outside the United States, you may select a community in which you have spent some time, such as Boston, MA.)

The end product will be a 4-6 page narrative (or story) in which you describe what you have learned from the Atlas. The next page lists specific analyses and questions that your narrative must address. It should be double spaced with references, graphs, and maps.

This project focuses on the following methods for descriptive data analysis. (The later empirical projects you will do in this class will be focused on causal inference and prediction).

1. *Data visualization.* Maps are a powerful way to present descriptive statistics for data with a geographic component. You will use maps to display upward mobility statistics for the Census tracts in your hometown.

2. *Regression and correlation analysis.* You will use linear regressions and correlation coefficients to quantify the statistical relationship between upward mobility and potential explanatory variables.

The Stata data file that you will use in this assignment, atlas.dta, contains an extract of the Opportunity Atlas data. I have also merged on several other variables, which you may use for the correlational analysis.

We will invite 5-10 students who produce the most compelling and insightful stories/analyses to discuss them with Professor Chetty and his team members at a lunch hosted at Opportunity Insights.

Instructions

Please submit your Empirical Project on Canvas. Your submission should include three files:

1. A 4-6 page narrative as a word or pdf document (double spaced and including references, graphs, maps, and tables)
2. A do-file with your STATA code or an .R script file with your R code
3. A log file of your STATA or R output

Specific questions to address in your narrative

1. Start by looking up the city where you grew up on the [Opportunity Atlas](). Zoom in to the Census tracts around your home.

   Figure 1 in your narrative should be a map of the Census tracts in your hometown from the Opportunity Atlas. Examples for Milwaukee, WI (where Professor Chetty grew up) and Los Angeles, CA (discussed in Lecture 1) are shown on the next page. The text of your narrative should describe what you see, and what data are being visualized.

   Examine the patterns for a number of different groups (e.g., lowest income children, high income children) and outcomes (e.g., earnings in adulthood, incarceration rates). Only choose one or two of these to include in your narrative.

2. (To answer this question, read the [Opportunity Atlas manuscript]()) What period do the data you are analyzing come from? Are you concerned that the neighborhoods you are studying may have changed for kids now growing up there? What evidence do Chetty et al. (2018) provide suggesting that such changes are or are not important? What type of data could you use to test whether your neighborhood has changed in recent years?

3. Now turn to the atlas.dta data set. How does average upward mobility, pooling races and genders, for children with parents at the 25th percentile (`kfr pooled_p25`) in your home Census tract compare to mean (population-weighted, using `count_pooled`) upward mobility in your state and in the U.S. overall? Do kids where you grew up have better or worse chances of climbing the income ladder than the average child in America?

   Hint: The Opportunity Atlas website will give you the tract, county, and state FIPS codes for your home address. For example, searching for "Lynwood Road, Verona, New Jersey" will display Tract 34013021000, Verona, NJ. The first two digits refer to the state code, the next three digits refer to the county code, and the last 6 digits refer to the tract code. In Stata, listing this observation can be done as follows:

   ```
   list kfr_pooled_p25 if state == 34 & county == 013 & tract == 021000
   ```

4. What is the standard deviation of upward mobility (population-weighted) in your home county? Is it larger or smaller than the standard deviation across tracts in your state? Across tracts in the country? What do you learn from these comparisons?

5. Now let's turn to downward mobility: repeat questions (3) and (4) looking at children who start with parents at the 75th and 100th percentiles. How do the patterns differ?

6. Using a linear regression, estimate the relationship between outcomes of children at the 25th and 75th percentile for the Census tracts in your home county. Generate a scatter plot to visualize this regression. Do areas where children from low-income families do well generally have better outcomes for those from high-income families, too?

7. Next, examine whether the patterns you have looked at above are similar by race. If there is not enough racial heterogeneity in the area of interest (i.e., data is missing for most racial groups), then choose a different area to examine.

8. Using the Census tracts in your home county, can you identify any covariates which help explain some of the patterns you have identified above? Some examples of covariates you might examine include housing prices, income inequality, fraction of children with single parents, job density, etc. For 2 or 3 of these, report estimated correlation coefficients along with their 95% confidence intervals.

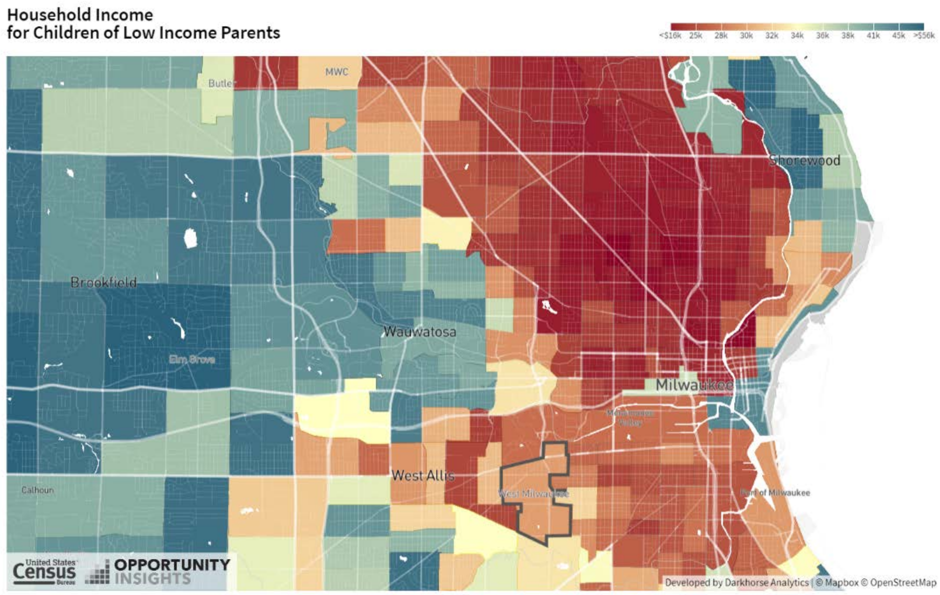
9. Open question: formulate a hypothesis for why you see the variation in upward mobility for children who grew up in the Census tracts near your home and provide correlational evidence testing that hypothesis.

   For this question, many covariates have been provided to you in the atlas.dta file, which are described under the "Characteristics of Census tracts" header in Table 1.

   You are welcome to use outside data that are not included in atlas.dta, but this is *not* required. Diane Sredl has created a research guide for our class that contains links to other data sources. You may wish to read this tutorial on how to add variables to a data set in Stata.

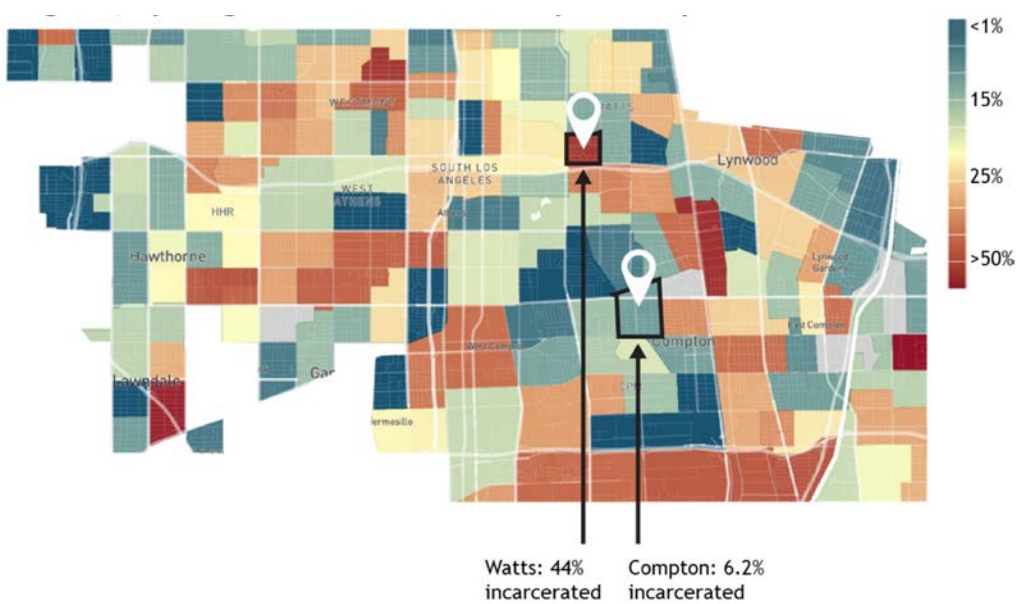
10. Putting together all the analyses you did above, what have you learned about the determinants of economic opportunity where you grew up? Identify one or two key lessons or takeaways that you might discuss with a policymaker or journalist if asked about your hometown. Mention any important caveats to your conclusions; for example, can we conclude that the variable you identified as a key predictor in the question above has a causal effect (i.e., changing it would change upward mobility) based on that analysis? Why or why not?

Figure 1
Household Income in Adulthood for Children Raised in Low-Income Households in Milwaukee, WI



Notes: This figure shows household income at ages 31-37 for low income children who grew up in Census tracts near Milwaukee, WI. The image was saved from www.opportunity-atlas.org by first searching for "Milwaukee, WI" and then clicking on the "download as image" button.

Figure 2
Incarceration Rates for Black Men Raised in the Lowest-Income Households in Los Angeles, CA



Notes: This figure is from the non-technical summary of the Opportunity Atlas and was discussed in Lecture 1.

DATA DESCRIPTION, FILE: atlas.dta

The data consist of *n* = 73,278 U.S. Census tracts. For more details on the construction of the variables included in this data set, please see Chetty, Raj, John Friedman, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. 2018. "The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility." NBER Working Paper No. 25147.

Table 1
Definitions of Variables in atlas.dta

Variable name	Label	Obs.
(1)	(2)	(3)
1. Geographic identifiers		
tract	Tract FIPS Code (6-digit) 2010	73,278
county	County FIPS Code (3-digit)	73,278
state	State FIPS Code (2-digit)	73,278
cz	Commuting Zone Identifier (1990 Definition)	72,473
2. Characteristics of Census tracts		
hhinc_mean2000	Mean Household Income 2000	72,302
mean_commutetime2000	Average Commute Time of Working Adults in 2000	72,313
frac_coll_plus2010	Fraction of Residents with a College Degree or More in 2010	72,993
frac_coll_plus2000	Fraction of Residents with a College Degree or More in 2000	72,343
foreign_share2010	Share of Population Born Outside the U.S.	72,279
med_hhinc2016	Median Household Income in 2016	72,763
med_hhinc1990	Median Household Income in 1999	72,313
popdensity2000	Population Density (per square mile) in 2000	72,469
poor_share2010	Poverty Rate 2010	72,933
poor_share2000	Poverty Rate 2000	72,315
poor_share1990	Poverty Rate 1990	72,323
share_black2010	Share black 2010	73,111
share_hisp2010	Share Hispanic 2010	73,111
share_asian2010	Share Asian 2010	71,945
share_black2000	Share black 2000	72,368
share_white2000	Share white 2000	72,368
share_hisp2000	Share Hispanic 2000	72,368
share_asian2000	Share Asian 2000	71,050
gsmn_math_g3_2013	Average School District Level Standardized Test Scores in 3rd Grade in 2013	72,090
rent_twobed2015	Average Rent for Two-Bedroom Apartment in 2015	56,607

singleparent_share2010	Share of Single-Headed Households with Children 2010	72,564
singleparent_share1990	Share of Single-Headed Households with Children 1990	72,196
singleparent_share2000	Share of Single-Headed Households with Children 2000	72,285
traveltime15_2010	Share of Working Adults w/ Commute Time of 15 Minutes Or Less in 2010	72,939
emp2000	Employment Rate 2000	72,344
mail_return_rate2010	Census Form Rate Return Rate 2010	72,547
ln_wage_growth_hs_grad	Log wage growth for HS Grad., 2005-2014	51,635
jobs_total_5mi_2015	Number of Primary Jobs within 5 Miles in 2015	72,311
jobs_highpay_5mi_2015	Number of High-Paying (>USD40,000 annually) Jobs within 5 Miles in 2015	72,311
nonwhite_share2010	Share of People who are not white 2010	73,111
popdensity2010	Population Density (per square mile) in 2010	73,194
ann_avg_job_growth_2004_2013	Average Annual Job Growth Rate 2004-2013	70,664
job_density_2013	Job Density (in square miles) in 2013	72,463
3. Measures of Upward Mobility from the Opportunity Atlas		
kfr_pooled_p25	Household income ($) at age 31-37 for children with parents at the 25th percentile of the national income distribution	72,011
kfr_pooled_p75	Household income ($) at age 31-37 for children with parents at the 75th percentile of the national income distribution	72,012
kfr_pooled_p100	Household income ($) at age 31-37 for children with parents at the 100th percentile of the national income distribution	71,968
kfr_natam_p25	Household income ($) at age 31-37 for Native American children with parents at the 25th percentile of the national income distribution	1,733
kfr_natam_p75	Household income ($) at age 31-37 for Native American children with parents at the 75th percentile of the national income distribution	1,728
kfr_natam_p100	Household income ($) at age 31-37 for Native American children with parents at the 100th percentile of the national income distribution	1,594
kfr_asian_p25	Household income ($) at age 31-37 for Asian children with parents at the 25th percentile of the national income distribution	15,434
kfr_asian_p75	Household income ($) at age 31-37 for Asian children with parents at the 75th percentile of the national income distribution	15,360

kfr_asian_p100	Household income ($) at age 31-37 for Asian children with parents at the 100th percentile of the national income distribution	13,480
kfr_black_p25	Household income ($) at age 31-37 for Black children with parents at the 25th percentile of the national income distribution	34,086
kfr_black_p75	Household income ($) at age 31-37 for Black children with parents at the 75th percentile of the national income distribution	34,049
kfr_black_p100	Household income ($) at age 31-37 for Black children with parents at the 100th percentile of the national income distribution	32,536
kfr_hisp_p25	Household income ($) at age 31-37 for Hispanic children with parents at the 25th percentile of the national income distribution	37,611
kfr_hisp_p75	Household income ($) at age 31-37 for Hispanic children with parents at the 75th percentile of the national income distribution	37,579
kfr_hisp_p100	Household income ($) at age 31-37 for Hispanic children with parents at the 100th percentile of the national income distribution	35,987
kfr_white_p25	Household income ($) at age 31-37 for white children with parents at the 25th percentile of the national income distribution	67,978
kfr_white_p75	Household income ($) at age 31-37 for white children with parents at the 75th percentile of the national income distribution	67,968
kfr_white_p100	Household income ($) at age 31-37 for white children with parents at the 100th percentile of the national income distribution	67,627
3. Counts of number of children under 18 in 2000 (to calculate weighted summary statistics)		
count_pooled	Count of all children	72,451
count_white	Count of White children	72,451
count_black	Count of Black children	72,451
count_asian	Count of Asian children	72,451
count_hisp	Count of Hispanic children	72,451
count_natam	Count of Native American children	72,451

Note: This table describes the variables included in the atlas.dta file.

Table 2a
STATA Hints

STATA command	Description
**clear the workspace* *clear* *set more off* *cap log close* **change working directory and open data set* *cd "C:\Users\gbruich\Ec1152\Projects\"* *use atlas.dta*	This code shows how to clear the workspace, change the working directory, and open a Stata data file. To change directories on either a mac or windows PC, you can use the drop down menu in Stata. Go to file -> change working directory -> navigate to the folder where your data is located. The command to change directories will appear; it can then be copied and pasted into your .do file.
**Summary stats* *sum yvar [aw = count_pooled]* **Summary stats for Wisconsin* *sum yvar if state == 55 [aw = count_pooled]* **Summary stats for Milwaukee County* *sum yvar if state == 55 & county == 079 [aw = count_pooled]* (Last two lines all go on one line in Stata)	These commands report means and standard deviations for *yvar*, weighted by the variable *count_pooled*. The first line calculates these statistics across the full sample. The second line calculates these statistics for observations in Wisconsin. The third line calculates these statistics for observations in Milwaukee County.
reg yvar xvar1 xvar2 xvar3, robust	This command estimates an OLS regression of *yvar* against *xvar*1, *xvar*2, and *xvar*3, using heteroskedasticity-robust standard errors.
**Report correlation coefficients* **Method 1* *sum yvar* *gen y_std = (yvar - r(mean))/r(sd)* *sum xvar* *gen x_std = (xvar - r(mean))/r(sd)* *reg y_std x_std , robust* **Method 2* *corr yvar xvar*	These commands show two methods for estimating correlation coefficients. The first block of code shows how to first generate standardized versions of the variables *yvar* and *xvar* by subtracting from each its mean and then dividing each by its variance (which are stored temporally by Stata as r(mean) and r(sd)). The last line reports an OLS regression of these transformed variables, with heteroskedasticity robust standard errors. The second method is to use the *corr* command, which does not report standard errors.
twoway (scatter yvar xvar) (lfit yvar xvar) *graph export figure1.png, replace*	This pair of commands first draws a scatter plot of *yvar* against *xvar*. The second line saves the graph as a .png file. Also see this tutorial on graphs in Stata.
**start a log file* *log using milwaukee.log, replace* **commands go here* **close and save log file* *log close*	These commands show how to start and close a log file, which will save a text file of all the commands and output that appears on in the command window in stata.

Table 2b: R Commands

R command	Description
#clear the workspace *rm(list=ls())* *#Install and load haven package* *install.packages("haven")* *library(haven)* *#Change working directory and load stata data set* *setwd("C:/Users/gbruich/Ec1152/Projects")* *atlas <- read_dta("atlas.dta")*	This sequence of commands shows how to open Stata datasets in R. The first block of code clears the work space. The second block of code installs and loads the "haven" package. The third block of code changes the working directory to the location of the data and loads in atlas.dta.
# summary stats, unweighted *summary(atlas$yvar)* *mean(atlas$yvar, na.rm=TRUE)* *sd(atlas$yvar, na.rm=TRUE)*	These commands show how to calculate unweighted summary statistics.
# Install and load package *install.packages("SDMTools")* *library(SDMTools)* *#Report weighted summary statistics* *wt.mean(atlas$yvar, atlas$count_pooled)* *wt.sd(atlas$yvar,atlas$count_pooled)*	These commands show how to calculate weighted summary statistics.
## subset observations to Wisconsin *wisconsin <- subset(atlas,state == 55)* *## subset observations to Milwaukee County* *milwaukee <- subset(atlas,state == 55 & county == 079)*	These commands show how to subset the data to observations in only Wisconsin and in only Milwaukee county.
#Install and load sandwich and lmtest packages *install.packages("sandwich")* *install.packages("lmtest")* *library(sandwich)* *library(lmtest)* *#Run regression with homoskedasticity-only standard errors* *mod1 <- lm(yvar~xvar1+xvar2 + xvar3, data = milwaukee)* *summary(mod1)* *#Report coefficients with heteroskedasticity robust standard errors* *coeftest(mod1, vcov = vcovHC(mod1, type="HC1"))*	This sequence of commands shows how to estimate an ordinary least squares regression with heteroskedasticity-robust standard errors. The first block of code first loads the necessary packages. The second block of code estimates a regression of *yvar* against *xvar*1, *xvar*2, and *xvar*3, then reports the estimated coefficients, *homoskedasticity-only* standard errors, and regression diagnostics (R^2, adjusted R^2, RMSE/SER which is referred to in the output as the Residual standard error). The last block of code reports the coefficients with heteroskedasticity-robust standard errors.
#Method 1 *##Standardize variables* *milwaukee$x_std <- (milwaukee$yvar - mean(milwaukee$yvar))/sd(milwaukee$yvar)* *milwaukee$y_std <- (milwaukee$xvar - mean(milwaukee$xvar))/sd(milwaukee$xvar)* *#Report correlation coefficients* *#Using a regression* *mod2 <- lm(y_std ~ x_std, data = milwaukee)*	These commands show how to estimate correlation coefficients. The first block of code shows how to first generate standardized versions of the variables *yvar* and *xvar* by subtracting from each its mean and then dividing each by its variance. The last line reports a OLS regression of these transformed variables,

summary(mod2) *coeftest(mod2, vcov = vcovHC(mod2, type="HC1"))* *#Note that regression output matches the following output* *cor(milwaukee$kfr_pooled_p25, milwaukee$job_density_2013)*	with heteroskedasticity robust standard errors. The second method is to use the *cor* command, which does not report standard errors.
# Install and load ggplot2 package *install.packages("ggplot2")* *library(ggplot2)* *# Draw scatter plot with linear fit line* *ggplot(data = milwaukee) + geom_point(aes(x = xvar1, y = yvar)) +* *geom_smooth(aes(x = xvar, y = yvar), method = "lm", se = F)* *#Save graph as figure1a.png* *ggsave("milwaukee_scatter.png")*	These commands show how to draw a scatter plot of *yvar* against *xvar1*. The *geom_smooth* part of the code adds an OLS regression line. The last line saves the graph as a .png file.
sink(file="milwaukee_log.txt", split=TRUE) *sink()*	The first line starts a log file. The last line closes and saves the log file.