Department of City and Regional Planning
CPLN 505 Planning By Numbers/ Spring 2019
Professor: Megan Ryerson
TA: Joshua Davidson

Assignment 3: Regression

(due March 2, 2019, 11:59 pm)

Mayor Kenney's infrastructure investment analysis team has awarded you additional scope to develop statistical models to explain reasons for Philadelphia's high reported official need for water infrastructure. The infrastructure investment team is more statistically savvy than the average informed layperson, and you can assume that they will read your report looking to point out potential flaws in the analysis, so it is good to be proactive about pointing out limitations and shortcomings of the analyses, but also not discredit your own work.

For students who would like to map this data in GIS software, latitude and longitude data are provided. Export the provided .csv file to .dbf format using the R package `foreign`. In ArcGIS, Add the .dbf file, then in the Layer tree, right click on the tabular dataset, and select 'Display XY Data'. When selecting the spatial reference, choose: 'Geographic Coordinate System NAD 1983 (2011)'**.** From here, you can do spatial joins to other datasets (you don't have to for this assignment, but this is often very useful).

Data Key:

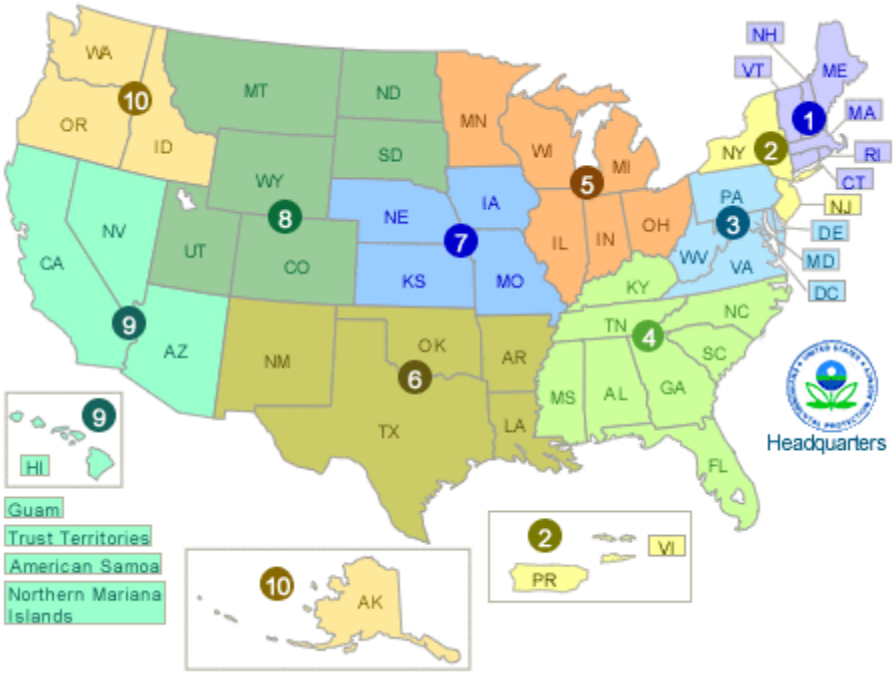
Column Name	Description	Source
STCOU	State and county identifier	US Census County-level data
CWNS_NUMBER	Facility identifier	EPA 2012 CWNS
TOTAL_OFFICIAL_NEED	Total reported need for water infrastructure	EPA 2012 CWNS
PRES_RES_REC_COLLCTN	Present (2012) residential population receiving collection	EPA 2012 CWNS
PRES_RES_REC_TRMT	Present (2012) residential population receiving treatment	EPA 2012 CWNS
PRES_N_RES_REC_COLLCTN	Present (2012) non-residential population receiving collection	EPA 2012 CWNS
PRES_N_RES_REC_TRTM	Present (2012) non-residential population receiving treatment	EPA 2012 CWNS
PROJ_RES_REC_COLLCTN	Projected residential population receiving collection	EPA 2012 CWNS
PROJ_RES_YR	Projection year for residential population	EPA 2012 CWNS
PROJ_RES_REC_TRMT	Projected residential population receiving treatment	EPA 2012 CWNS
PROJ_N_RES_REC_COLLCTN	Projected non-residential population receiving collection	EPA 2012 CWNS
PROJ_N_RES_YR	Projection year for non-residential population	EPA 2012 CWNS

PROJ_N_RES_REC_TRMT	Projected non-residential population receiving treatment	EPA 2012 CWNS
EPA_REGION	EPA Region (1-10), see map below	EPA 2012 CWNS
STATE	State where facility is located	EPA 2012 CWNS
FACILITY_NAME	Name of facility	EPA 2012 CWNS
OWNER_TYPE	Public, Private, Federal	EPA 2012 CWNS
FACILITY_CITY	City where facility is located	EPA 2012 CWNS
LATITUDE	Latitude of facility	EPA 2012 CWNS
LONGITUDE	Longitude of facility	EPA 2012 CWNS
TMDL_INDICATOR	Y=contributing to a TMDL receiving water body	EPA 2012 CWNS
STATEFP	State code	US Census County-level data (via spatial join)
COUNTYFP	County code of location of facility	US Census County-level data (via spatial join)
COUNTYNAME	County name of location of facility	US Census County-level data (via spatial join)
ALAND	Land area (m^2) of county	US Census County-level data (via spatial join)
AWATER	Water area (m^2) of county	US Census County-level data (via spatial join)
MEDINC69	Median income in 1969 in county	US Census County-level data (via spatial join)
MEDINC79	Median income in 1979 in county	US Census County-level data (via spatial join)
MEDINC89	Median income in 1989 in county	US Census County-level data (via spatial join)
MEDINC99	Median income in 1999 in county	US Census County-level data (via spatial join)
MEDINC09	Median income in 2009 in county	US Census County-level data (via spatial join)
MEANINC09	Mean income in 1969 in county	US Census County-level data (via spatial join)
POP80	Total population in county in 1980	US Census County-level data (via spatial join)
POP90	Total population in county in 1990	US Census County-level data (via spatial join)
POP00	Total population in county in 2000	US Census County-level data (via spatial join)
POP10	Total population in county in 2010	US Census County-level data
PCTWHITE00	Percent of county population identifying as white in 2000	US Census County-level data (via spatial join)
PCTWHITE10	Percent of county population identifying as white in 2010	US Census County-level data (via spatial join)
CSS	1=facility associated with a combined sewer system (CSS)	EPA 2012 CWNS (via T. Lim recategorization)
MS4	1=facility associated with a municipal separate sewer system (MS4)	EPA 2012 CWNS (via T. Lim recategorization)

Data References

US EPA. "Clean Watersheds Needs Survey 2012: Report to Congress," January 2016. https://www.epa.gov/cwns/clean-watersheds-needs-survey-cwns-2012-report-and-data.

USA Counties Data File Downloads. http://www.census.gov/support/USACdataDownloads.html#IPE



EPA Regions 1-10. Source: usepa.gov

I. DATA PREPARATION AND PRELIMINARY TESTS

Suspecting that Philadelphia's high infrastructure need can be operationalized in different ways, you create the following variables to represent 'need':

Total Official Need ($, untransformed)
Log of Total Official Need (log $)
Residential Burden ($/person)
Log of Residential Burden (log $/person)

Give an explanation for why you choose to operationalize 'need' in the above ways (what each can reveal, and the interpretations of coefficients from regression change). Boxplots and density/histograms of these variables can help illustrate very skewed distributions.

Hint: Be careful of dividing by zeros in communities that have either reported zero residential treatment or zero residential collection values. You can assume that most residents that receive treatment also receive collection and vice versa, but that non-standard reported has resulted in

some inconsistency in the data. To deal with this while avoiding double counting, try something like:

```
dat$resburden<- dat$TOTAL_OFFICIAL_NEED/max(dat$PROJ_RES_REC_COLLCTN,
   dat$PROJ_RES_REC_TRMT)
```

Also calculate the following transformations of potential explanatory variables:

- logs of Median Income variables
- logs of Population variables
- Dummy variables for 'groups' such as 'growing cities' or 'cities with non-residential populations'

You could also try to incorporate dummy variables such as 'growing cities' or 'cities with non-residential populations' or any other categorizations you think would explain facilities' systematically higher or lower reported needs. Remember: dummy variables are essentially intercept adjustments in a regression that based on the entire group having a higher/lower mean than non-group facilities.

From what you found from your previous scope (i.e. any significant associations or correlations you found in Assignment #2), you may also wish to include some of the change and percent change metrics you calculated to explain 'need'. Explain theoretical reasons why you want to include certain variables in your regressions.

Run pairwise correlations between variables that you may want to include in your regression to make sure to avoid multicollinearity, or included highly correlated variables in the regressions.

II. BIVARIATE REGRESSIONS

Run bivariate regressions of explanatory variables against your 'need variables.' Make scatterplots showing the linear relationships between them (convention is to put the independent variable on the x-axis and the dependent variable, 'need' in this case, on the y-axis). Interpret the coefficient estimates of significant variables. How well do these simple bivariate models perform in predicting the various measures of 'need'?

III. MULTIVARIATE REGRESSIONS

Build multivariate regressions for each of the measures of 'need'. Explain the use of, and report the results of automated variable selection techniques, such as Forward and Backward Selection. In your final models for each of the measures of 'need', report which variables you chose to include in your final "leanest and meanest" models, interpret the significant coefficients of the final models. Make sure that you interpret interesting results in full sentences, as well as presenting regression results in a table summarizing your work. Compare models against each other, using R squared values and F-tests.

Remember to take care in interpreting log-transformed variables.

Try interactions between variables. Do any have significant effects?

Report the VIF values to convince suspicious readers that you have considered the problem of multicollinearity in your analysis.

V. CONCLUSIONS

Draft a conclusion to the report to Mayor Kenney and his infrastructure investment analysis team summarizing your findings.