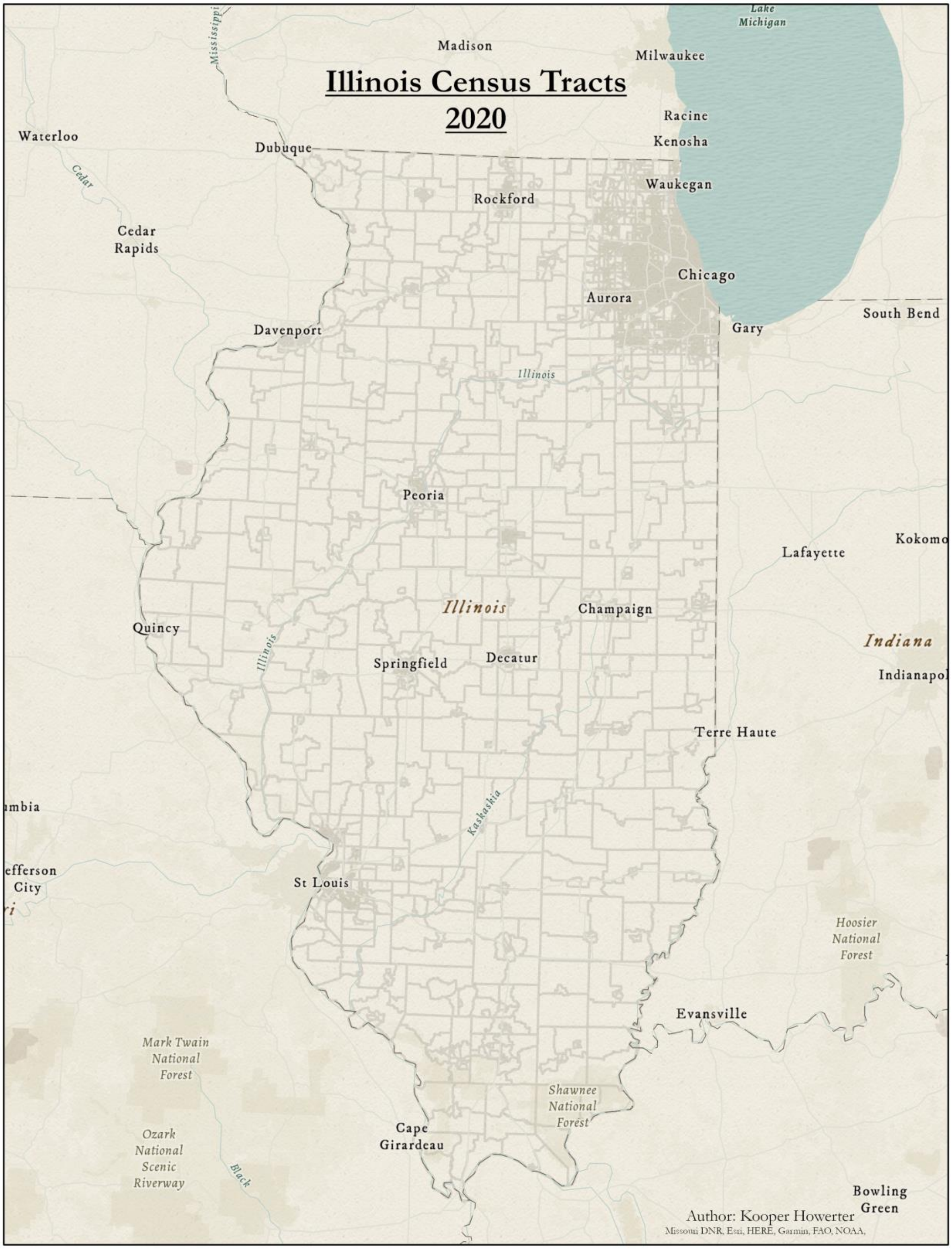
Kooper Howerter

Analyzing Relationship Between Income and Higher Education in Illinois



Factors relating to educational attainment have long been studied, with strong relationships being of particular interest in improving educational opportunities across the country. Similarly, income can have many long-reaching impacts on social, economic, and medical factors. In this study, the goal was specifically to analyze the relationship and theoretically the impact of income on educational attainment. Before analyses is done, we hypothesize that the relationship will be very strong; higher incomes will correlate strongly with higher rates of educational attainment.

Our independent variable is median household income by census tract in the year 2020. These ranged from below $5000 to above $250,000. Our dependent variable is the percentage of residents in each tract that have higher education experience, calculated by adding up all of the residents that attained some higher education (bachelors, masters, doctorates, professional degrees) and dividing by the total number of residents. These percentages almost filled the entire possible range, with some tracts near 0% and others near 100%. Higher incomes not only provide the means to pay for increasingly expensive higher education, but more affluent areas typically have lower-level educational institutions that better prepare residents for higher education. Those with higher educational attainment make more money on average than those without, so as a result the areas they live generally have higher incomes, especially when they choose more homogeneous areas with others of similar means and experiences. For all these reasons, we expect median income to be an excellent predictor for educational experience.

The main analysis was done by combining Census data with a simple linear regression in SciPy, a python package. The census data provided the income and college attainment statistics for each Census tract, so after properly setting up the data, SciPy could use an input data frame to calculate the regression statistics. These statistics could then be used to calculate the predicted and residual values for each tract, then put them back into the Illinois feature class for mapping and visualization (Page 5). All of the necessary code is shown on Page 3 and 4. Note that the data setup requires combining a shapefile with census data as an excel, then calculating or locating needed fields to be used in the regression.

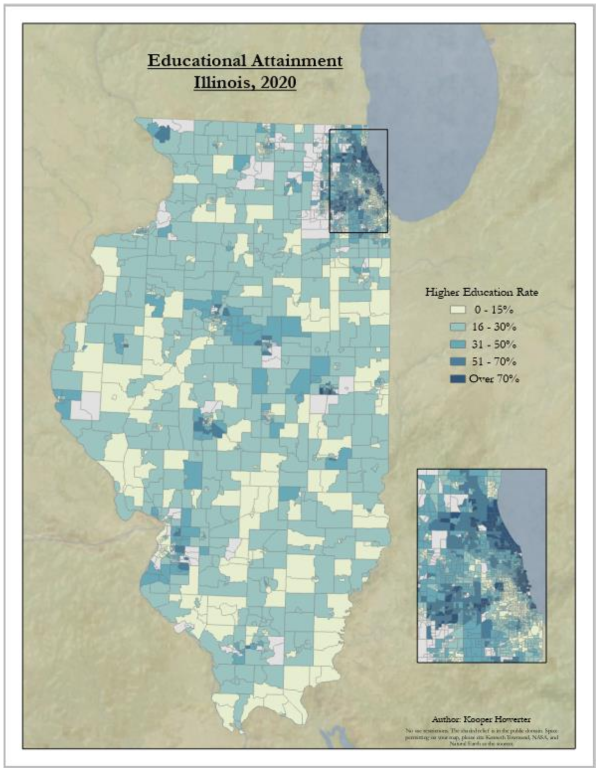
Looking at the ouput data and maps of the linear regression model, its clear that income was a solid predictor, with an r-value of 0.76. However, we might infer that an income skew may have caused slight hiccups with our results. Income is very skewed compared to education attainment; there are a decent amount of tracts in Illinois with multiple times more income than the norm, but do not have an overwhelmingly higher education level. On the other end, there are not that many counties with drastic amounts less income than the norm (around 45,000), so the monetary data is definitely skewed right. For further analysis, some normalization of the income data could improve the model greatly.

The predicted values map lines up well with the actual data map, showing that the model was not terribly far off. However, the most obvious difference is much more education expected in the more rural counties with moderate incomes. Compare the lightest two colors on each map to see that there was much less prediction to have low education rates than is the reality. Similarly, in the regression map we see that the models cool spots are in larger, more rural tracts and the hotspots are all nearby the cities (compare to map on page 1 of major urban areas). Many of the rural areas have small populations with moderate to good incomes, so the model predicted these would have moderate rates of educational attainment. However, by looking at the data we see that the education rates were very low in many of these counties. As for the urban postive residuals, these are often areas with solid incomes, but the rates of education are some of the highest. If I had to speculate, these are areas with lots of younger people who are educated at increasingly high rates, and are choosing to live nearby the cities while making slightly above average incomes. I think age would be a good variable to add in the future, as this will account for a higher education rate without a necessarily higher income level. The tan color is what we are looking for to prove the model's success as it shows the tracts predicted within +/- 10 percentage points of the actual data. There was definitely a good split of residuals with more than half within 10, even though it looks dominated by cool spots since those tracts are often the biggest ones.

```
import arcpy
import pandas as pd
import numpy as np
from arcpy import env
from scipy import stats
env.workspace = "C:\\GISProjects\\Illinois\\Illinois.gdb"
arcpy.env.overwrite = True
#hard code needed parameters, the commented out parameters could be
used to incorporate as tool entirely in ArcGIS
#feature = arcpy.GetParameterAsText(0)
feature = 'Illinois_Tracts'
#fldf1 = arcpy.GetParameterAsText(1)
fldf1 = 'S1903_C03_001E'
#fldf2= arcpy.GetParameterAsText(2)
fldf2 = 'PercCollege'

#create numpy array with two fields
data = arcpy.da.TableToNumPyArray(feature, [fldf1, fldf2])
#array = arcpy.da.FeatureClassToNumPyArray(in_table = feature,
field_names = [field1,field2], null_value = -9999)
#turn numpy frame into pandas frame
frame = pd.DataFrame(data)
frame.dropna(inplace=True)

#run the regression to return intercept, slope, r, p, error
b1, b0, r_value, p_value, std_err = stats.linregress(frame[fldf1],
frame[fldf2])

#add field for predictions
arcpy.management.AddField(feature, 'Predicted', 'DOUBLE')
#predict the value of education level from income
#formula for predictions = b0(slope) + b1(y-intercept) * (income)
expression = "b0+b1*!S1903_C03_001E!"
arcpy.management.CalculateField('Illinois_Tracts', 'Predicted',
expression)

#add another field for residual
arcpy.management.AddField(feature, 'Residual', 'DOUBLE')
#calculate residuals (observed - predicted)
expression2 = "!PercCollege! - !Predicted!"
arcpy.management.CalculateField('Illinois_Tracts', 'Residual',
expression2)
```

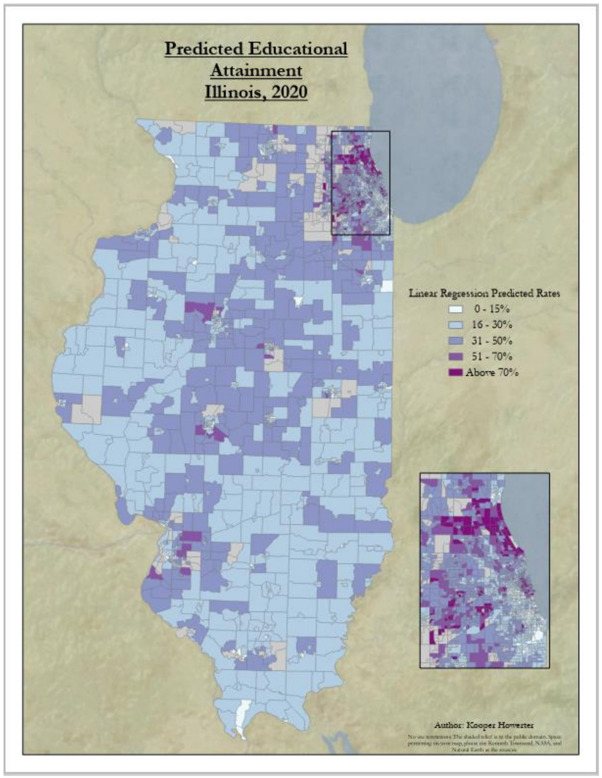
```
#code to ensure missing values are handled properly in Arc
preds = []
field = arcpy.da.SearchCursor(feature, ['Predicted'])
for item in field:
    preds.append(item[0])
del field

resids = []
field2 = arcpy.da.SearchCursor(feature, ['Residual'])
for item in field2:
    resids.append(item[0])
del field2

cols = ['S1903_C03_001E', 'PercCollege']

cursor = arcpy.da.SearchCursor(feature, cols)
edit = []
for rownum, row in enumerate(cursor):
    for i in range(len(cols)):
        if row[i] == None:
            break
    else:
        edit.append(rownum)
del cursor

cursor = arcpy.da.UpdateCursor(feature, ['Predicted', 'Residual'])

rw = 0
for rownum, row in enumerate(cursor):
    if rownum in edit:
        row[0] = preds[rw]
        row[1] = resids[rw]
        cursor.updateRow(row)
        rw += 1
del cursor
```

Created Maps. Note that linear regression was run using median income to predict education attainment rate. Areas with missing data shown in gray.



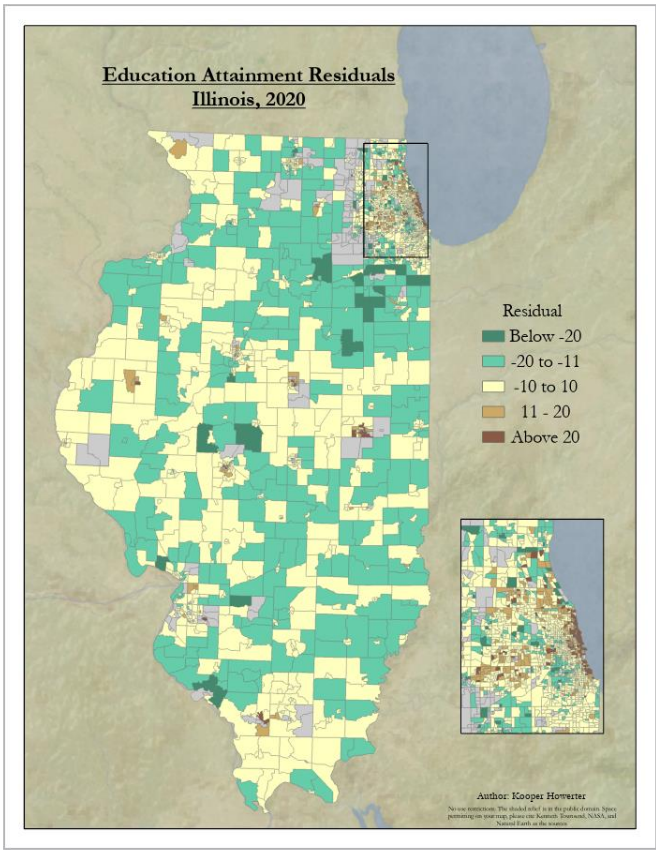
Author: Kooper Howerter

No use restrictions. The shaded relief is in the public domain. Space permitting on your map, please cite Kenneth Townsend, NASA, and Natural Earth as the sources.



Author: Kooper Howerter

No use restrictions. The shaded relief is in the public domain. Space permitting on your map, please cite Kenneth Townsend, NASA, and Natural Earth as the sources.



Author: Kooper Howerter

No use restrictions. The shaded relief is in the public domain. Space permitting on your map, please cite Kenneth Townsend, NASA, and Natural Earth as the sources.