**Bowen Li**
**Natasia Engeline**

# How R&D Fund impacts Migration and Economic Growth in the US

*In this project, our goal is to find how the amount of R&D funds from the US government to the Universities each year is affecting the county-level growth in economics and population. Each year, the US government spends millions of dollars to fund the universities to conduct researches in a variety of areas including Engineering, Computer Science, Biochemical Studies, to name a few. We are interested to find out if those money funds into the universities will actually attract people outside of where the universities are located and more broadly if the combined effect of population attraction along with funding will ultimately influence the county's overall economic growth in the long term.*

## Our main data sources are:

**All data collected are selected from year 2010 to year 2019**

- R&D spending
  National Science Foundation Dataset - https://ncsesdata.nsf.gov/home
- Population Data based on Nativity and Place of Birth
  Census Data: API Code B05002 – ACS 5
  https://data.census.gov/cedsci/table?q=native%20and%20foreign%20born%20place%20of%20birth&g=1400000US17031330100&tid=ACSDT5Y2019.B05002&hidePreview=true
- Income Level
  Census Data: API Code B19301 – ACS 5
  https://data.census.gov/cedsci/table?q=PER%20CAPITA%20INCOME&g=1400000US17031330100&tid=ACSDT5Y2019.B19301&hidePreview=false
- Population data and Income Data prior to 2005 - we decided not to use this because this would cause missing data in between. The detail population nativity was a part of the Decennial survey. Hence if we combine the data, it would only be 1990, 2000, and start from 2010 onwards.
  https://www.nhgis.org/
- Shape Files:
  https://catalog.data.gov/dataset/tiger-line-shapefile-2019-nation-u-s-current-county-and-equivalent-national-shapefile
  https://hifld-geoplatform.opendata.arcgis.com/datasets/geoplatform::colleges-and-universities/about

## General steps:

- We first downloaded the data. We are using API for the ACS data while the data for R&D must be parse manually using the csv file. Shape file are from census bureau (US map) and arcgis open source data (university geo data).
- We then clean the dataset consisting of income, population based on nativity, and also the R&D fund for top 50 universities in USA.
- We then merge the dataset for the regression purposes and the plotting purposes.
- For the plotting, we opt for 3 different style of plotting:

**Bowen Li**
**Natasia Engeline**

  - Bokeh with Geopanda and interactive plot.
  - Bokeh line with interactive plot (widget option).
  - Static line to check the distribution of dataset and general trend.
- For the interactive plotting, we only select the state and counties where the top universities reside.
- For the interactive plotting using bokeh, we tried using the interactive feature that came with bokeh application handler and also using the ipywidgets. Both has its own charm and also complexities in terms of usage.
- We found some issue when using the interaction capabilities in Bokeh (please refer to the limitation and future improvement of this project)

## Findings:

- **Regression:**
  - We first create dummy variables of year to control for yearly effect after data cleaning.
  - Based on the simple OLS regression, we can see that the fund is only significant in affecting the income for the past 12 months. The t-statistic for the fun when we regress income on fund is 3.366 with P-value of 0.001. Some of the dummy variable is also statistically significant which is year 2017, 2018, 2019 at 5% significance level.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         income_past12m   R-squared:                       0.100
Model:                            OLS   Adj. R-squared:                  0.082
Method:                 Least Squares   F-statistic:                     5.402
Date:                Tue, 07 Dec 2021   Prob (F-statistic):           1.33e-07
Time:                        19:27:11   Log-Likelihood:                -5217.6
No. Observations:                 495   AIC:                         1.046e+04
Df Residuals:                     484   BIC:                         1.050e+04
Df Model:                          10
Covariance Type:            nonrobust
================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept       2.738e+04   1575.143     17.381      0.000    2.43e+04    3.05e+04
C(year_2011)[T.1]   641.9013   1851.380      0.347      0.729   -2995.834    4279.636
C(year_2012)[T.1]   898.0218   1851.300      0.485      0.628   -2739.556    4535.600
C(year_2013)[T.1]  1195.7350   1861.390      0.642      0.521   -2461.669    4853.138
C(year_2014)[T.1]  1740.7645   1861.006      0.935      0.350   -1915.884    5397.413
C(year_2015)[T.1]  2274.9771   1861.373      1.222      0.222   -1382.393    5932.348
C(year_2016)[T.1]  3157.7291   1852.932      1.704      0.089    -483.055    6798.514
C(year_2017)[T.1]  4654.5666   1854.344      2.510      0.012    1011.008    8298.125
C(year_2018)[T.1]  6306.9849   1856.890      3.397      0.001    2658.424    9955.546
C(year_2019)[T.1]  8181.3573   1880.235      4.351      0.000    4486.926    1.19e+04
fund               0.0043      0.001      3.366      0.001       0.002       0.007
==============================================================================
Omnibus:                      212.499   Durbin-Watson:                   0.180
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              801.233
Skew:                           1.992   Prob(JB):                    1.03e-174
Kurtosis:                       7.793   Cond. No.                     8.51e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.51e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

- **Static Plotting:**
  - We found that all the 50 counties are experiencing positive income growth which might explain why the regression result of income on fund is significant.

  - From the static plot, we could also see that the share of foreigner growth is relatively flat while R&D fund show positive growth. This suggests that R&D fund might not have any effect on the population growth.
  - We also found that the counties where 50 top universities reside more or less receive similar amount of funding except one or two counties that received more funding that the other counties.
  - Some counties also experience significant income growth that might not be contributed by the R&D fund (the growth is more significant than other counties). While majority of the counties experience similar growth.
  - See: fund_trend, income trend, and foreigner share trend static plot in folder static png.
- **Interactive Plotting:**
  - Using line plots, by testing how either population growth or the income per capita on the county level, we have seen a trend that those counties where the target universities are located do get a fast and steady increase in income per capita and population growth.
  - Using an interactive map plot, we can see that the finding is aligned with the above results. The county, where the universities are located, is marked with darker color compared to elsewhere; and in some cases, counties that are adjacent to the target share similar economic status. There are certainly correlations between schools and county economics.
  - In terms of population growth on the interactive map plot. We do not visually detect any significant change in the 10-year range, due to vastly different county size in population and social-economic status.

## Limitations:

We have encountered many limitations along the way due to time constraint and abilities to use more efficient tools to analyze and present our hypothesis

- Our research is limited to 10 years from 2010 to 2019 from our initial plan which ranges from 1990 to 2019. This is due to the inconsistency of data indexing and lack of data from the Census Data which is our main use of data source. We comprised to use years 2010 to 2019 so we would be able to align county FIPS with NSF dataset; also, data collected from 2010 has more valid data to use than the year 1990.
- With the limited data and the gap between ACS survey and Decennial data, it would be worth exploring in the future to do some interpolation of dataset.
- We cannot confirm whether our hypothesis or model is correct using university funds as an independent variable to determine the growth of county-level economy and population. We have seen metropolitan counties like Los Angeles and Manhattan are huge in scale in terms of economy and population already, there would be insignificant effects of R&D fund on the overall county development, if not at all.
- We also find that given our rather short time period of observations, the overall population change in some counties is insignificant in absolute numbers, so it is hard for our project to determine if whether such fund has a real-life impact; Moreover, due to immigration is an elongated phenomenon, if our hypothesis were true that school fund did attract people from outside of states, then the materialization period would require us to expand our research time period at least a decade into the future to obtain a better understanding of the true impact.

**Bowen Li**
**Natasia Engeline**

- We have derived significant p values, but this is based on assuming that the model is correctly specified. If the model we are using y ~ a + x1 + x2 + e, isn't the "correct" model to explain y, then the p-value has no meaning. This is a very difficult thing to test and requires a solid theoretical justification.
- In terms of OLS result. we might suspect that this result is a naïve estimation since it could be the case that the income is also experiencing overall positive growth which coincides with the fund growth.
- Hence, we might need to use other method like placebo test to check difference between county that received R&D funds and counties that do not received R&D funds.

## Future Improvement:

- Full information of census on county-level beyond the year range of our study is crucial to further our testing on the hypothesis.
- Our regression analysis is constrained by our limited sample size and regression model. We may also need to conduct T-test and use a better model to define the relationships
- In the map plots, we fixated the maximum and minimum colormapper bar which cause our population map at county-level to fail to reflect true state-level population density. Some counties like Los Angeles are so populated, they become an outlier of the average county population sample. Instead of putting a fixed number, those ranges should be varied by state.
- When creating the bokeh interactive plot, we encounter some difficulties especially regarding the widget interaction, several things that we found and finally resolve:
  - We manage to solve the error whereby initially when you choose the year first or the county first the plot would not be working. Now it works perfectly. It turns out because we are using the drop down with different name (to make the naming more beautiful instead of using the snake format e.g total_population) when updating either the year or state, the variable names are using the "Income in Past 12 Months" instead of using the column name 'income_past12m' which we initially use for creating the base plot.
  - We hope to be more proficient in the Bokeh, however with the time limitation and also the complexity of interactive plot using Bokeh the project duration might not be sufficient to gain deep understanding of Bokeh complex interactions.