State Names

Pranay Gundam

2023-08-27

State Name Etymology

First as some color commentary for this first TidyTuesday analysis. This data set doesn't give one much to work with. There are some interesting visualizations one can make, but I found it quite difficult to ask any meaningful well conditioned research questions. That being said the best I could think of for now was: The effects of native american vs european influence on states.

The influence of Native American vs European culture is not a very concrete term. In the context of this project, we will be using the etymology of a state's name as a proxy for Native American vs European influence on a state. This is of course not anywhere near perfectly analogous but gives some level of insight about the early formation of a given state.

EDA

There are originally two datasets that we combine together which in total has information on simple characteristics of each state such as its' population in 2020, number of representatives in the House of Representatives, land/water area metrics, and etymological characteristics of a states' name. Currently there are various languages and we characterize them into European or not.

```
mergedf = merge(states, etymo, by = "state")

eurolang = function(x){
  y = substring(x, 1, 8)
  return(grepl("Spanish", y, fixed=TRUE) |
         grepl("English", y, fixed=TRUE) |
         grepl("Latin", y, fixed=TRUE) |
         grepl("Russian", y, fixed=TRUE) |
         grepl("French", y, fixed=TRUE) |
         grepl("Greek", y, fixed=TRUE) |
         grepl("Dutch", y, fixed=TRUE))
}

findf = mergedf %>%
  mutate(eurolang = ifelse(eurolang(language), 1, 0))
```

After doing so, we can see that the division between these two classes of languages is close to even. 26 European languages and 30 non-European languages.

Main Analysis

For the first TidyTuesday we are keeping things simple and just creating a fixed effects model to determine if having a european vs native american origin has an effect on the population of a state in 2020 controlled by its land size, water size and the date at which the state's name was decided. Our regression looks like

$$pop = \beta_0 + \beta_1 X_{origin} + \boldsymbol{\beta} Z + \epsilon.$$

Our data set doesn't give a very clear variable to indicate if a state is of european native american origin but we will be using the etymology of a state's name as a proxy for its origin.

```
model = lm(data = findf, formula = population_2020 ~ admission + date_named +
             land_area_km2 + water_area_km2 + eurolang)
summary(model)
```

```
##
## Call:
## lm(formula = population_2020 ~ admission + date_named + land_area_km2 +
##     water_area_km2 + eurolang, data = findf)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -10866362  -3423590  -1704934   1874438  28824751
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -5.475e+06  3.926e+06  -1.394  0.16933
## admission      -1.772e+02  9.350e+01  -1.895  0.06382 .
## date_named      2.420e+00  5.042e+01   0.048  0.96191
## land_area_km2   2.223e+01  8.048e+00   2.762  0.00802 **
## water_area_km2 -6.610e+01  4.704e+01  -1.405  0.16609
## eurolang        9.579e+05  2.149e+06   0.446  0.65777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6830000 on 50 degrees of freedom
## Multiple R-squared:  0.1729, Adjusted R-squared:  0.09014
## F-statistic:  2.09 on 5 and 50 DF,  p-value: 0.08215
```

The results indicate that there isn't any statistically significant effect of the eurolang variable (and therefore our proxy for the cultural origins of a state) on the population of a state in 2020. As an explanation for our result, other than the obvious potential explanation that there simply is no relationship between the two it could also be that the amount of time that has passed since the founding of a state has overturned any residual effects of its cultural founding.

We should obviously discuss some limitations of the discussion above. There are undoubtedly some untracked variables that could contribute to omitted variable bias: for example, it is likely that the distance to a body of water and ports is connected to both the population in that state and the etymological origins of that state (since these are the states that were often colonized first by europeans).

In addition, the relationship between the etymology of a state's name and it's cultural origins may not be a strong link which would make the supposed causal relationship we are trying to discern between the cultural origins of a state and its population in 2020 not the relationship we are exploring.

Future TidyTuesday Posts

This current post personally feels a bit lackadaisical in terms of what I am actually doing but I just wanted to get one out the door and I have to start somewhere before I can release a product that is something I've worked on only for a week and I am also satisfied with.

Future posts, conditional on the specific data set that we are working with and the type of questions we can ask, will range from building machine learning models to questions about causality or forecasting. Also, as a preface for the next two posts, the data sets we will be using are regarding spam emails and fair use cases.