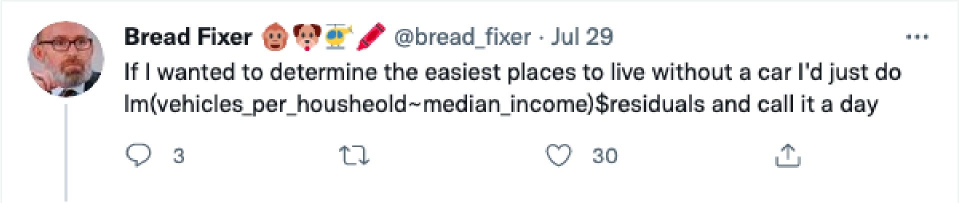
Problem set 2: Money and cars

S470/670 Fall 2021

Upload a HTML/PDF/Word document with your graphs and write-up to Canvas by 11:59 pm, Tuesday 14th September.



Bread Fixer 🐵🐶🚁🖍 @bread_fixer · Jul 29

If I wanted to determine the easiest places to live without a car I'd just do lm(vehicles_per_housheold~median_income)$residuals and call it a day

Background

Former Indiana University student and popular Twitter user `@bread_fixer` proposed that one could find the "easiest places to live without a car" by fitting a linear model predicting vehicles per household using median income, then looking at the residuals. Current Indiana University faculty member and unpopular Twitter user `@bradluen` suggested that it might be better to use the log of income as the predictor instead. Does it make a difference?

Data

The file `vehicles.txt` contains tab-separated long-form data for U.S. counties from the 2019 American Community Survey. (See the appendix to learn how to get this and similar data for yourself.) The variables are:

- `GEOID`: The ID number for the county.
- `NAME`: The county name and state.
- `variable`: One of `total`, `cars0`, `cars1`, `cars2`, `cars3`, or `cars4`.
- `estimate`: The value of that variable for that county.
- `moe`: The margin of error for that estimate.

The variables measured for each county are:

- `total`: The number of households in the county.

- `cars0`: The number of households in the county with no motor vehicles.
- `cars1`: The number of households in the county with 1 motor vehicle.
- `cars2`: The number of households in the county with 2 motor vehicles.
- `cars3`: The number of households in the county with 3 motor vehicles.
- `cars4`: The number of households in the county with 4 or more motor vehicles. (For the purpose of this exercise, treat "4 or more" as "4.")

Your task

1. By transforming the data to "wide form" or otherwise, estimate the mean number of vehicles owned per household for each county in the data set. Plot the distribution of this variable.
2. Plot vehicles owned (y-axis) against median income (x-axis) for each county. Add the linear regression line. How well does the line fit?
3. Plot vehicles owned against median income using a log scale on the x-axis for each county. Add the linear regression line. Does this line fit any better?
4. Fit a linear regression that predicts vehicles owned using median income for each county. List the ten counties with the lowest (most negative) residuals. Do these seem like the easiest places to live without a car?
5. Fit a linear regression that predicts vehicles owned using log median income for each county. List the ten counties with the lowest (most negative) residuals. Do these seem any more like the easiest places to live without a car?

Grading

Two points per question, ten points total. Remember to label things and discuss results to get full credit.

Appendix: How I got the data

I got the data using the `tidycensus` R package. You can learn more about this package at `https://walker-data.com/tidycensus/articles/basic-usage.html` . Using it requires a Census Data API key. To sign up for one, enter your email at `https://api.census.gov/data/key_signup.html` and they'll send you a key (you'll have to wait a few minutes for it to activate.)

```
vehicles = get_acs(geography = "county",
  variables = c(total = "B08201_001",
                median_income = "B19013_001",
                cars0 = "B08201_002",
                cars1 = "B08201_003",
                cars2 = "B08201_004",
                cars3 = "B08201_005",
                cars4 = "B08201_006"),
  year = 2019)
```