STAT 4840 Final Project
County Level Poverty

Spring 2025

Introduction

The U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program produces single-year estimates of income and poverty for all U.S. states and counties as well as estimates of school-age children in poverty for all 13,000+ school districts. In this assignment, you will:

- Acquire and clean poverty data for each county in one U.S. state.
- Create various models for the number of people in poverty in each county.
- Use accuracy measures for model selection.
- Provide a 5-year forecast for each county.
- Report your results with tables and visualizations.

You will hand in one (or more) **knit** markdown documents, which should show code and output but be free of extraneous material (unused code, dumps of data).

This is an individual assignment, but you are welcome to help each other with it during or outside of class. Internet resources are allowed as well. The assignment is divided into parts. Everything requires Part 1. Part 2 is independent of Parts 3 and 4.

Due Dates

- ✓ Tuesday, April 22: Part 1 should be complete.
- ✓ Thursday, May 1: Progress check. Hand in what you've got. You should have completed Part 2 or most of Parts 3 and 4.
- ✓ Tuesday, May 13 at 1:50pm: Project due, no late work accepted.

1 Data Preparation

Start this project by reading about the SAIPE program on the official US Census web site. Here are some key pages (links are active here and also available on Canvas).

- About the SAIPE program.
- SAIPE FAQ.
- County-level estimation methodology.
- Interactive data page.
- Model Input Data and descriptions.

For this project, you'll need to combine data from three sources.

- The SAIPE data, which provides the number of people in poverty for each county, as well as the county population.
- The County SNAP Benefits Data, which provides the number of people served by the Supplemental Nutrition Assistance Program ("food stamps") for each county.
- The State IRS Data, which provides the number of exemptions on tax forms filed by households below the poverty threshold.

1.1 SAIPE data

Get the SAIPE data for your state from the interactive data page as follows: Select your state. View by counties. In the table, in the Year column, there is a filter icon (▽). Select it, select all years. At this point, you can remove the values that correspond to your entire state and the entire United States, or do it later in R. Use the download icon at the top of the table to download the table data as a CSV.

This data is pretty clean already, and should read into R nicely. The variables you need are the Year, ID, Name, Poverty Universe, and Number in Poverty. You may wish to rename them, especially Poverty Universe which is really just the county population.

The ID variable is a 5-digit FIPS code. The first two digits describe your state, the next three digits describe the county. It is better to use the FIPS code than the county name when working with county data, since county names may not be exactly the same across all sources.

For this section:

- Give some basic information: The name, abbreviation, and FIPS code for your state.
- Find the total number of counties, the largest county, and list the nine largest counties (by current population).
- Make a map of your state with each county colored by its current population. I recommend the `usmap` package for this.
- Make a time plot showing the number in poverty for each of the nine largest counties. (You can make one plot and facet it.)

1.2 County SNAP Benefits

The County SNAP Benefits Data is on the Model Input Data page. You'll want to get the CSV file which is available under Additional File Formats/Archived Datasets as `cntysnap.csv`.

This file will take some work to clean up. It has stray header lines which you can avoid with `read.csv`'s skip option. You'll need to filter down to just your state, and avoid the state totals (with county FIPS of zero). Combine the state and county FIPS codes into a five-digit FIPS code.

Finally, this data is not tidy: the year is contained in variable names. You need the data to have one row per county and year. The tool for this is `pivot_longer`.

For this section:

- Make a time plot showing the number receiving SNAP benefits for each of the nine largest counties.

1.3 State IRS Data

The State IRS Data is on the Model Input Data page. You'll want to get the CSV file which is available under Additional File Formats/Archived Datasets as `irs.csv`.

The only variable you need from this file is "Poor exemptions," which gives the total number of exemptions on IRS forms filed by households below the poverty line. Since most families file one exemption per person, this is a good measure of the number of people in poverty statewide.

For this section:

- Make a time plot showing the number of poor exemptions filed in your state.

1.4 Merging the data

Once you have three clean data frames, you need to merge them into one table which has the year, the FIPS, the county name, the number of people in poverty, and the three input variables population, snap, and poor exemptions.

If your FIPS codes match, you can perform this merge by doing `left_join` twice, attaching the snap and poor exemptions variables to the SAIPE file.

At this point, check to make sure you don't have any problem counties. Each county should have the same number of rows (years) and check for counties with many NA entries. If you have a couple of small counties with problems, remove them (or find a creative way to repair them). The US census maintains a web page listing changes to counties that might impact your work: `https://www.census.gov/programs-surveys/geography/technical-documentation/county-changes.html`

You may have noticed that some early years are missing from the data. Although it's possible to fill the gaps, to keep things simple you can remove all data from before 1997.

Finally, make your data into a tsibble by year, with the FIPS code and county name as keys.

For this section:

- Make some visualizations (your choice) to explore the relationship between the number in poverty and the three input variables.

2 Linear models

As background, read the SAIPE web page on County-level estimation methodology, especially the model for the total number of people in poverty. In particular, the census bureau argues that all of our variables should be logged. Build all your models using log variables.

In this portion of the project, the goal is to create a time series linear model of the number of people in poverty from the three input variables population, SNAP, and poor exemptions.

2.1 Variable selection

With three input variables, there are seven possible models if we choose to use one, two, or all three of them. Chapter 7.5 of our textbook discusses the situation.

Build all seven linear models and choose the one that does best across all counties.

For this section:

- Which of the three input variables did the best model include?
- For each of the nine biggest counties, make a plot showing the actual number in poverty as well as the predictions made by your best linear model. (You can use `facet_wrap` to make all nine at once).

2.2 Residual analysis

At this point, you've fit one linear model to every county in your state. Let's investigate the residuals.

For this section:

- Make a time plot of the innovation residuals (since we took logs) for the nine biggest counties.
- Run the Ljung-Box test on every county's innovation residuals. How many counties residuals are significantly different from white noise?
- Do you think the linear model does a good job of predicting the number in poverty?

3 Stochastic models

The linear model is a good approximation to US Census methodology. However, to forecast future poverty counts we would need to first forecast all the input variables. In this part of the assignment, we forget about the three other variables and just build stochastic models for the number in poverty.

The end goal is to choose one model that we can then fit to every county and use for forecasting.

3.1 Single county forecasts

For the moment, only consider data from your state's largest county. Fit a bunch of models to the number in poverty:

- The NAIVE model.
- The mean model.
- Simple exponential smoothing.
- Holt's additive linear method both damped and not.
- Auto ARIMA.

Use additive errors in the exponential smoothing models, because we are fitting to the log of number in poverty already.

For this section:

- For each model, plot the number in poverty data along with a five-year forecast. (Again, you can do this easily with facets.)
- Use some measure of model quality to decide which of these models is the best for this county.

3.2 Exponential smoothing models

Take the three exponential smoothing models (SES and the Holt's variations) and fit them to every county in your state. Use some measure of model quality to decide which of these models does the best across the entire state.

For this section:

- Which exponential smoothing model did you select and why?

3.3 ARIMA models

Fit an auto-ARIMA model to every county in your state. Which models are most commonly selected?

Now take the most common ARIMA models and fit each one to every county in your state. Use some measure of model quality to decide which of these models does the best across the entire state.

For this section:

- Which ARIMA model did you select and why?

3.4 Cross validation

You now have two contenders to use as the one model for your whole state – an ETS and an ARIMA model. Let's have a showdown with cross validation (see FPP3 section 5.10).

Use `stretch_tsibble` to build testing sets for a 5-year forecast on every county. Fit both models to it, and check the accuracy of their predictions using RMSE. You'll need to combine the per-county RMSE to form a measure of how well they fit the state overall.

For this section:

- Which model performed the best on cross validation?

4 Forecasts

Finish this project by determining which five counties in your state are predicted to have the highest percentage increase in poverty over the next five years.

Use the winning model from section 3.4 to make five year forecasts for each county. From the forecast, calculate the predicted increase in number of persons living in poverty. Divide that by the current county population to get a percentage increase.

Use `plot_usmap` from the `usmap` library to make a chloropleth map of all counties in your state, colored by the predicted increase in poverty.

For this section:

- Which five counties (by name) do you predict will have the highest percentage increase in poverty over the next five years?
- Map the forecast poverty increase.