U.S. Census Modeling

Your Name

Due: dd/mm/yyyy

Project tasks

Article 1 of the United States Constitution actually mandates that a census of the US population be taken every 10 years. This has various purposes, including reallocation of seats in the U.S. House of Representives based on relative population of each state. Population growth is typically not linear (as we'll see).

You will scrape and clean the US Census data from Wikipedia, and then fit an exponential model of the form:

$$y = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 * x)}$$

where y* is the total population in the US at the time of the census taken in year x. This is the model family, and you will use R's `optim` function to estimate the parameters $\beta_1, \beta_2, \& \beta_3$ for the non-linear model that best fits the data.*

0. Get ready for the census taker!

Task 0.1 Just for fun, embed the image "farmers.jpg" showing an 1940's advertisement from the US Census Bureau here. Be sure to include the following image caption: "Image source: https://catalog.archives.gov/id/514239"

1. Data Preparation

1.1 [2 pts] Scrape the source data

Task 1.1.1 Scrape and inspect the population data for all decennial U.S. Census results from 1790 through 2010 from Wikipedia (link: https://en.wikipedia.org/wiki/United_States_Census). Call the resulting data structure `USCensusRaw`.

Task 1.1.2 Be sure to inspect the `USCensusRaw` data structure and show a few rows of the data before any processing/cleaning.

1.2 [5 pts] Data cleansing

A bit of cleaning is necessary to prepare the data for further analysis. It's a good idea to keep an unchanged copy of the complete source data around (i.e., `USCensusRaw`) when possible. The cleaned/processed data should be given a new name: `USCensus`.

Task 1.2.1 For example, your code should use regular expressions to identify and eliminate all footnoote references in the date column. You can also drop the notes column, since we won't be using it for analysis. Other simple cleaning operations will be necessary as you complete your analysis; you'll figure them out along the way, but your code for those steps belongs here.

Task 1.2.2 Be sure to inspect the `USCensus` data structure and show a few rows of the cleaned table.

2. [8 pts] Exploratory Data Analysis

We should almost always plot our data to try and better understand it's structure before proceding with analysis.

Task 2.1 show a scatter plot of population (in millions, rounded to 3rd decimal place; call it `popMillions`) and add a smoother (e.g. loess) to the plot. Make sure your plot includes descriptive axis labels and an informative title.

Task 2.2: show the same plot with a line of best fit (rather than the smoother).

Task 2.3: explain why the linear model is NOT useful here. it's okay to state the obvious, but try to be specific.

3. [8 pts] Non-linear model fit (`optim` with RMS distance metric)

For the purpose of modeling we will use an index for each census as our explanatory variable instead of `year`. Call the index `census`, and add it to the data back in your "data cleansing" task earlier in the assignment. The `census` index should start at 0 (zero) for the 1790 census, and increment for each subsequent census (1 for 1800, 2 for 1810, and so on). In other words, we want to assume that 1790 is "time 0" because we can use that (with a little algebra) to solve for reasonable estimates for our three model parameters. I'll spare you the trouble, but you should start with the following estimates:

- $\beta_1 \approx 400$*; you can think of* β_1 *as something like an upper-bound for the US poplation. Certainly there are many countries populated far more densely than the U.S., but we also wouldn't want to extrapolate* too* dramatically and some have even begun to argue that the US (and other countries) has already exceeded it's sustainable carrying capacity. [1]*
- $\beta_2 \approx 4.5$*; reflects size of population at time 0, adjusted for the upper-bound of the population.*
- $\beta_3 \approx -0.3$*; controls growth rate of the population.*

Task 3.1 Store predicted values for the initial model of `popMillions` vs `census` (our index variable) using the provided starting values as a vector called `initialPreds` in the `USCensus` data. What population does this initial model estimate for year 2010?

Task 3.2 Store residuals for the initial model of `popMillions` vs `census` (our index variable) using the provided starting values as a vector called `initialResids` in the `USCensus` data. Interpret the residual for the 2010 census.

Task 3.3 Calculate and show the root-mean-square (RMS) distance metric for our initial model.

Task 3.4 Produce and report optimal parameter estimates for beta1, beta2, and beta3 using the RMS distance metric. Interpret the estimates for beta1 and beta3 in the context of the study.

4. [4 pts] Non-linear model fit (`optim` with MAD distance metric)

Task 4.1 Calculate and show the mean absolute deviation (MAD) distance metric for our initial model.

Task 4.2 Produce and report optimal parameter estimates for beta1, beta2, and beta3 using the MAD distance metric. Interpret the estimates for beta1 and beta3 in the context of the study.

[1] Ehrlich, Paul R; Ehrlich, Anne H (2004), One with Nineveh: Politics, Consumption, and the Human Future, Island Press/Shearwater Books, pp. 137, 182, see also pages 76–236

5. Bootstrap confidence intervals for parameter estimates

As always, it's important to consider the uncertainty of our estimates. We'll do so with Bootstrapping.

5.1 RMS model fit

Task 5.1.1 [2 pts] Generate 1000 (on thousand) bootstrap model fits for the US Census data using RMS distance metric. Use the `head` function to show a few rows of our Bootstrap results, and clearly explain what each column represents.

Task 5.1.2 [1.5 pts] Calculate and report the bootstrap standard errors for beta1, beta2, and beta3.

Task 5.1.3 [3 pts] Show a bootstrap sampling distribution for EACH parameter estimate (with informative label or title) and share any pertinent observations.

Task 5.1.4 [3 pts] Calculate and report the bootstrap 95% confidence intervals for beta1, beta2, and beta3.

5.2 MAD model fit

Here, you will largely repeat the tasks associated with 5.1, but now using the mean absolute deviation (MAD) distance metric.

Task 5.1.1 [2 pts] Generate 1000 (on thousand) bootstrap model fits for the US Census data using MAD distance metric. Use the `head` function to show a few rows of our Bootstrap results, and clearly explain what each column represents.

Task 5.1.2 [1.5 pts] Calculate and report the bootstrap standard errors for beta1, beta2, and beta3 using your MAD model fit.

Task 5.1.3 [3 pts] Show a bootstrap sampling distribution for EACH parameter estimate (with informative label or title) and share any pertinent observations.

Task 5.1.4 [3 pts] Calculate and report the bootstrap 95% confidence intervals for beta1, beta2, and beta3.

6. [4 pts] Get ready for the census taker again!

Recall the model:

$$y = \frac{\beta_1}{1 + \exp(\beta_2 + \beta_3 * x)}$$

** Task 6.1 Use the RMS model fit to estimate the expected US population (in millions) for the 2020 Census**

Task 6.2 Use the MAD model fit to estimate the expected US population (in millions) for the 2020 Census