My title*

My subtitle if needed

Ahnaf Alam

March 24, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

*Code and data are available at: LINK.

1 Introduction

You can and should cross-reference sections and sub-sections. We use R Core Team (2023) and (**rohan?**).

The remainder of this paper is structured as follows. Section 2….

2 Data

2.1 Software and R-packages

We create this project using statistical software, `R` (R Core Team 2023). For cleaning and re-purposing the data, we used `tidyverse` (Wickham et al. 2019) package and graphs, we relied on `ggplot2` (Wickham 2016). The data used in this paper comes from `fredr` (Boysel and Vaughan 2021) package. We further used `rstanarm` (Goodrich et al. 2022) for modelling. Lastly, we used `kableExtra` (Zhu 2021) and `viridis` (Garnier et al. 2024) for aesthetics purposes.

2.2 Incorporating FRED data

The data comes from FRED or Federal Reserve Economic Data, which is an online database, consisting of hundreds of thousands time series economic data on both US national level and international level. From the database, we incorporated six different datasets using `fredr` package. These were titled:

- Real personal consumption expenditure: Food
- Real disposable personal income
- Real personal consumption expenditure: Durable Goods
- Real personal consumption expenditure: Nondurable Goods
- Real personal consumption expenditure: Services
- Real personal consumption expenditure: Healthcare

A detailed description of what each of these datasets reports on can be found on Table 1. There are few key features that are present in all of the datasets. Firstly, we only incorporate data between 2007 and 2022 on a quarterly basis. We used as an 2007 as an anchor because data on durable/nondurable goods is only available from year and we wanted to model our data on 15 year period, hence 2022.

All the datasets are also seasonally adjusted and chained to 2017 dollars. Seasonally adjusted time series eliminates effect of seasonal influences. Seasonal influences like strikes, abnormal weather patterns, or events Boxing day sale, can distort the real underlying movements in the business cycle and adjusting for these variation, provides us with much clearer understanding

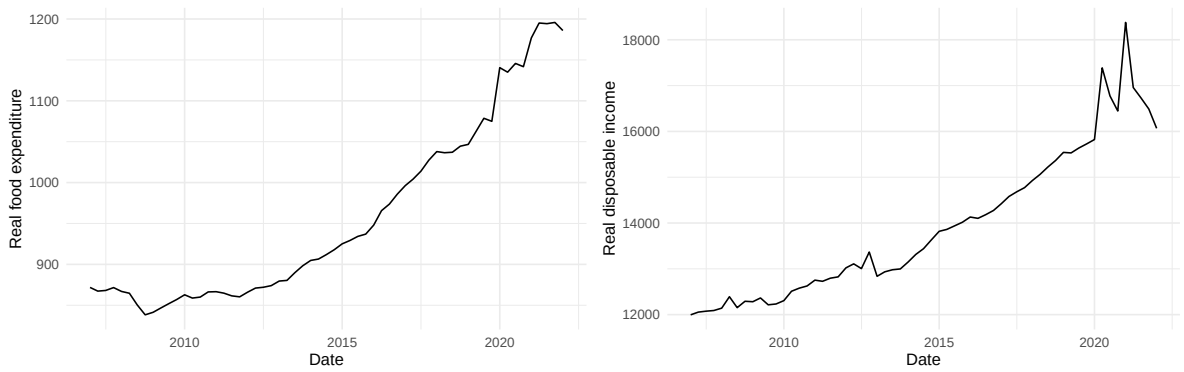
of the dataset from period to period. The datasets used 2017 price level as reference point. This adjusts for inflation across time, allowing us to accurately compare economic data over multiple periods. We further adjust for inflation by using real economic data, as opposed to nominal data. This enables us to create valid comparison groups, allowing us to compare a category with another category across time.

2.3 Dataset characteristics

Each dataset contains five different columns, with key ones being date and value. The data variable reports on the specific day on which data was collected. With quarterly data, we only see data on the first day from the months of January, April, August and October. Although quarterly, FRED does not include data for quarter months of August, or indeed December. This is more due to convenience as some data becomes available only after end of the quarter and updating the database on before that is not productive. Lastly, value columns reports on expenditure in billions of US dollars. For a better understanding of how each category measure up against one another, please see Figure 2. Table 2 reports in cleaned data that is being used for modelling and analysis. Datasheet is available at Section A.1.

2.4 Measurement

FRED does not collect data itself but relies on public and private organizations to provide the database with data. Except for first and last observations of the month, FRED ignores missing value when it average, sum and end-of-period aggregation ("Getting to Know FRED" 2024). In this context, missing values often arise during statutory holidays, when federal offices are closed. On those weeks, FRED only reports data on 6 days of the week, excluding the holiday and end-of-period calculation will be conducted based whatever the corresponding days are in that month, minus one.



(a) Food expenditure (b) Income

Figure 1: Levels of real consumption expenditure and income expenditure, in billions of dollars, chained to 2017 prices

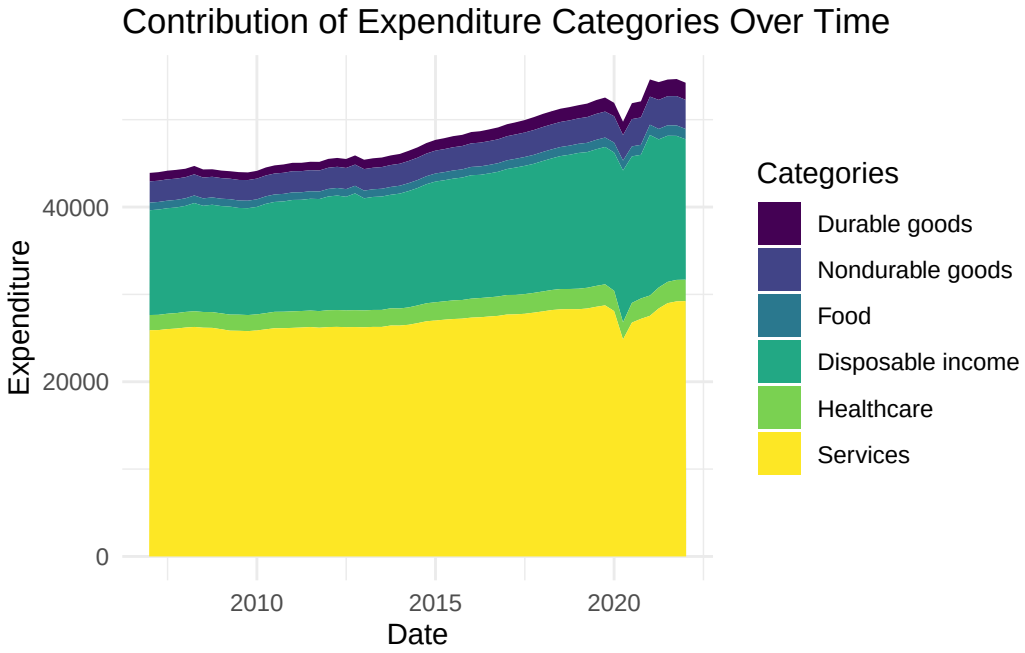


Figure 2: Expenditure for different categories, in billions of 2017 dollars

Table 1: Description of the data variables

Expenditure variable	Description
Date	Date of data collected
Durable goods	Durable goods are goods that are more for future consumption than immediate consumption. These types of goods provides utlity over a length of period. Examples include machinery, tools, appliances among others
Non-durable goods	Durable goods are anything that are generally consumed within a short period of time. Examples include food, clothing, cosmetics etc
Food	Expenditure in food by all Americans in a time period
Disposable income	Refers to total income that is availble to individuals for consumption after deducing taxes
Healthcare	The category reports on total expenditure on healthcare services, including medical treatments, medicine cost, physician services among other services
Services	This category encomapasses a variety of services, inluding education, transportation, utilities, hospitality and many others

3 Model

In this section, we briefly discuss Bayesian models that are being used in this analysis. Background details and model diagnostics can be found under Appendix B.

3.1 Model set-up

Using `rstanarm` library, we evaluated two Bayesian model, with one being simple linear regression, and another being multiple linear regression. The simple linear regression explores whether an increase in income leads to increase in expenditure in food. Multiple linear regression evaluates the same topic, however, controlling for various other predictors.

3.1.1 Simple linear regression

Define y_i as the expenditure in food in year i. Then $income_i$, is level of disposable income in year i, both in billions of US dollars.

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \beta_0 + \beta_1 \times \text{income}_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 279) && (3)\\
\beta_1 &\sim \text{Normal}(0, 0.17) && (4)\\
\sigma &\sim \text{Exponential}(0.009) && (5)
\end{aligned}
$$

Table 2: Cleaned data showing real expenditure by different categories

Date	Durable goods	Non-durable goods	Food	Disposable income	Healthcare	Services
2007-01-01	969.90	2435.00	871.80	11995.90	1736.72	25900.00
2007-04-01	980.10	2429.90	867.30	12055.30	1745.91	25913.00
2007-07-01	992.10	2437.50	868.10	12075.60	1762.55	26024.00
2007-10-01	999.50	2435.50	871.60	12090.30	1770.73	26088.00
2008-01-01	967.20	2416.60	866.90	12141.60	1788.97	26202.00
2008-04-01	960.30	2420.40	864.70	12391.20	1793.99	26273.00
2008-07-01	927.90	2385.10	850.40	12152.80	1800.01	26186.00
2008-10-01	859.60	2362.30	838.30	12291.70	1804.94	26167.00
2009-01-01	861.10	2361.40	841.40	12282.00	1817.91	26012.00
2009-04-01	854.90	2347.30	846.70	12364.40	1838.01	25848.00
2009-07-01	896.40	2354.70	851.90	12214.70	1849.31	25830.00
2009-10-01	875.30	2362.30	857.00	12232.60	1840.44	25795.00

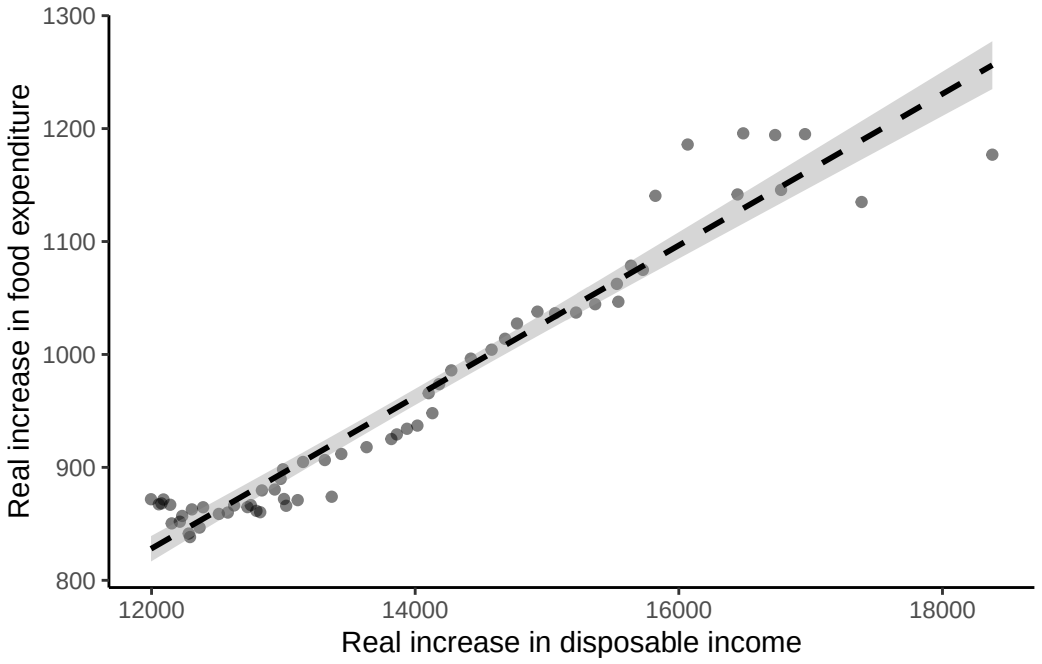


Figure 3: Relationship between increases in disposable income and increases in food expenditure, between 2007 and 2022.

Table 3: Summary results for both models

	Simple linear	Multiple linear
(Intercept)	24.07	−97.50
	(30.88)	(114.73)
income_expenditure	0.07	0.02
	(0.00)	(0.00)
durable_expenditure		−0.15
		(0.04)
nondurable_expenditure		0.47
		(0.04)
healthcare_expenditure		0.00
		(0.04)
services_expenditure		−0.01
		(0.01)
Num.Obs.	61	61
R2	0.936	0.992
R2 Adj.	0.933	0.990
Log.Lik.	−289.419	−224.943
ELPD	−292.9	−233.0
ELPD s.e.	8.0	8.0
LOOIC	585.7	466.0
LOOIC s.e.	16.0	16.0
WAIC	585.6	464.3
RMSE	27.49	10.36

3.1.2 Multiple linear regression

Define y_i as the expenditure in food in year i. Then $income_i$, is level of disposable income in year i. Model further controls for durable goods with $durable_i$, non-durable goods with $nondurable_i$, levels of health care expenditure with $healthcare_i$ and levels of expenditure in services with $services_i$. All the variables are in billions of US dollars.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{6}$$
$$\mu_i = \beta_0 + \beta_1 \text{income}_i + \beta_2 \text{durable}_i + \beta_3 \text{nondurable}_i + \beta_4 \text{healthcare}_i + \beta_5 \text{services}_i \tag{7}$$
$$\beta_0 \sim \text{Normal}(0, 279) \tag{8}$$
$$\beta_1 \sim \text{Normal}(0, 0.17) \tag{9}$$
$$\beta_2 \sim \text{Normal}(0, 0.84) \tag{10}$$
$$\beta_3 \sim \text{Normal}(0, 0.95) \tag{11}$$
$$\beta_4 \sim \text{Normal}(0, 1.25) \tag{12}$$
$$\beta_5 \sim \text{Normal}(0, 0.27) \tag{13}$$
$$\sigma \sim \text{Exponential}(0.009) \tag{14}$$

3.2 Model justification

We expect a positive relationship between between income and food expenditure. Figure 3 shows that for an increase an disposable income, food expenditure increases by approximately equal amount. Further, Figure 1 compares both of these variables over time and we see similar patterns of growth, with steady exponential increase in between 2010 and 2020 and both measures veer off after 2020, presumably due to Covid-19 downturn. A paper by Parker and Wong (1997) looks at data from Mexico and finds that income and expenditure are correlated. In fact, lower income uninsured groups reduces cash expenditure on health care during economic crisis. Mahadea and Rawat (2008) further considers relationship between between happiness and incomes and finds that economic growth and increased income contributes to happiness. Therefore, based on exploratory data analysis and theories, we believe that there is positive relationship between food expenditure and income levels.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

A.1 Datasheet

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

> It was created in order to analyze how food expenditure changes when income fluctuates. While all the datasets can be found individually on FRED website, they however, are not complied together and my datasets fills that void. The dataset complies different types of expenditure into one data frame, which I believe can help to answer the proposed question.

Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

> Ahnaf Alam, an undergraduate student at University of Toronto.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

> No direct funding was received for this project.

Any other comments?

> No.

Composition

What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

> Each row of the dataset composed on valuation in billions of US dollars, on a specific date. The data provides information of cumulative spending habits by Americans throughout the year.

How many instances are there in total (of each type, if appropriate)?

> There are about 366 instances in total.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).

> The dataset is a sample, howeverm it isn't random. The larger dataset consiststs of all the observations for expenditure on a daily basis from the time datasets were made available to FRED database. In that sense, sample is representative of the larger dataset as we see the patterns of consumptions and expenditure over 25 year period match the patterns we see our sample.

What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.

> Each instance consists of value of expenditure in billions of USD, across different categories.

Is there a label or target associated with each instance? If so, please provide a description

> Yes, the unique consists of specific date

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.

> There are no missing values in the dataset.

Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

> Yes, using the 'year' column.

Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

> No.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

> No.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The data is self-contained. The data will exist and won't change over time.

Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

No. These are publicly avaiable data, released by public organizations

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.

No. The data reports on national level.

Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

No.

Collection process

How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

> The data was collected from FRED website, using `fredr` package. The data was not directly observable but based on audits, taxes and surveys.

What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?

> We downloaded the data using `fredr` package on R.

If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?

> The dataset was collected based on specific years on a quarterly basis.

Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

> Ahnaf Alam and no one else.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

> The dataframe was created over period of one week.

Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

> No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?

> No. We relied on third-party website for in all cases.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

No.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Consent was not needed as we are dealing with data on country level.

Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Any other comments?

No.

Preprocessing/cleaning/labelling

Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.

Yes. The data was cleaned. There were no missing values in the dataset. From the raw data, we only selected columns that were pertinent to our question.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

No. However, if one were to run the codes avaliable on `01-download_data.R` file, they can access the raw data.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.

R was used.

Any other comments?

No.

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Not that I am aware of.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

No.

What (other) tasks could the dataset be used for?

we could use differnt types model to see how they compare to the ones we have performed in our paper.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

No.

Any other comments?

No.

Distribution

Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

The dataset will be avaiable on Github for later uses.

How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

It will be distributed through Github.

When will the dataset be distributed?

The dataset is avaiable now.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

No.

*Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.8

None.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

No.

Maintainence

Who will be supporting/hosting/maintaining the dataset?

Ahnaf Alam

How can the owner/curator/manager of the dataset be contacted (for example, email address)?

ahnaf.alam@mail.utoronto.ca

Is there an erratum? If so, please provide a link or other access point.

No.

Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?

No.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

Not applicable.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

No.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

Pull request on Github.

Any other comments?

No.

B Model details

B.1 Posterior predictive check

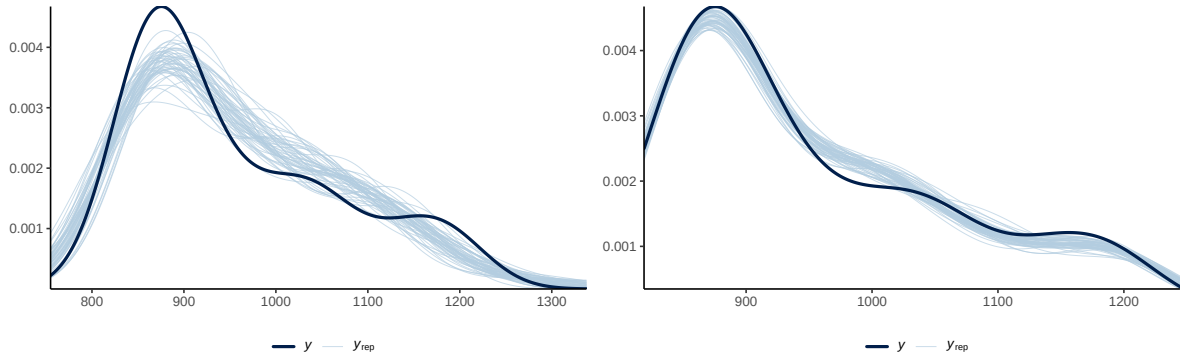
In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

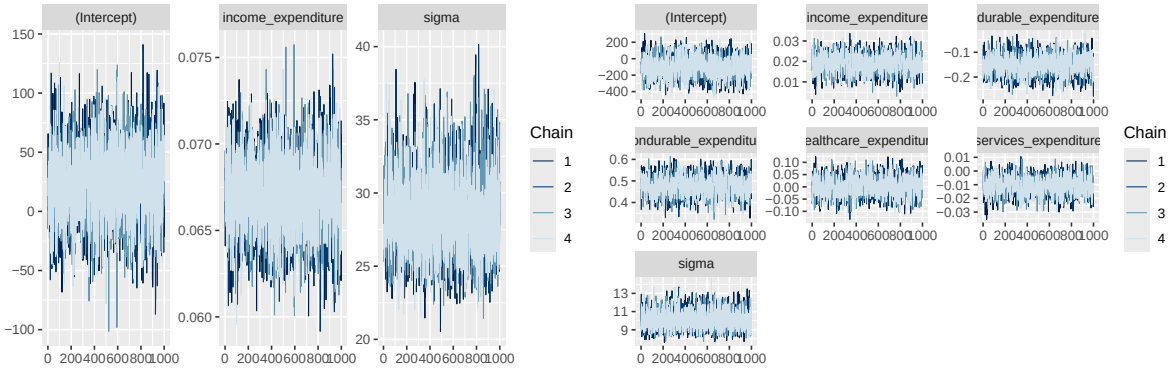
?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…



(a) Posterior prediction check of simple linear regression

(b) Posterior prediction check of multiple linear regression

Figure 4: Examining how the model fits, and is affected by, the data



(a) Trace plot of Model 1

(b) Trace plot of model 2

Figure 5: Checking the convergence of the MCMC algorithm

References

Boysel, Sam, and Davis Vaughan. 2021. *Fredr: An r Client for the 'FRED' API*. https://CRAN.R-project.org/package=fredr.

Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, et al. 2024. *viridis(Lite) - Colorblind-Friendly Color Maps for r*. https://doi.org/10.5281/zenodo.4679423.

"Getting to Know FRED." 2024. *Getting To Know FRED*. https://fredhelp.stlouisfed.org/fred/data/understanding-the-data/how-are-missing-values-treated-in-average-sum-and-end-of-period-aggregation-methods-2/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Mahadea, D, and T Rawat. 2008. "Economic Growth, Income and Happiness: An Exploratory Study." *South African Journal of Economics* 76 (2): 276–90.

Parker, Susan W, and Rebeca Wong. 1997. "Household Income and Health Care Expenditures in Mexico." *Health Policy* 40 (3): 237–55.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.