My title*

My subtitle if needed

First author Another author

November 16, 2024

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

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2....

2 Data

2.1 Overview Check packages

For this analysis, we combined four data sets all together into one. All these four data sets comes from **Worldbank** open data platform (Arel-Bundock (2022)). We employed **R** (R Core Team (2023)), a coding platform to download, clean and conduct statistical analysis. Besides, we also utilized R packages **tidyverse** (Wickham et al. (2019)), **rstanarm** (Goodrich et al. (2022)), **ggplot2** (Wickham (2016)), **knitr** (Xie (2014)), **arrow** (Richardson et al. (2024)), **here** (Müller (2020)), and **dpylr** (Wickham et al. (2023)). The paper is outlined in github using starter folder provided in **Telling Stories With Data** (Alexander (2023)).

[*] Code and data are available at: https://github.com/RohanAlexander/starter_folder.

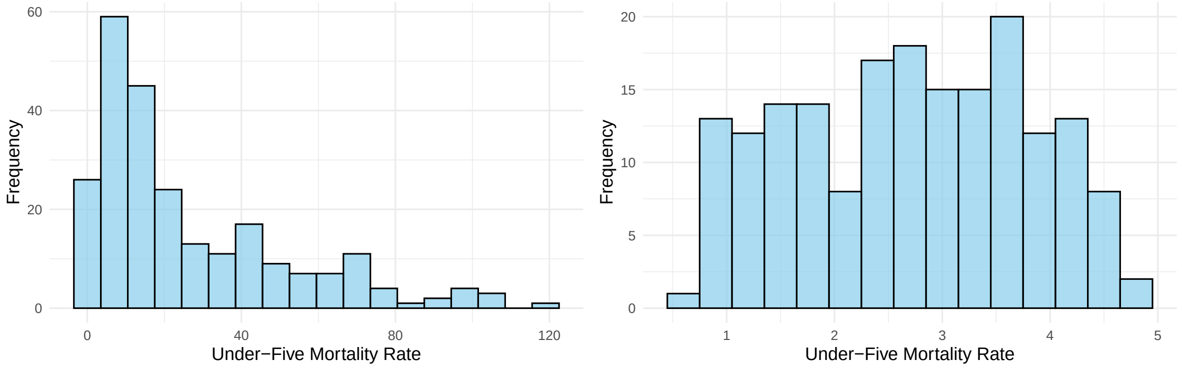
2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Outcome variables

The response variable in our analysis is the mortality rate below 5 year-old children for all countries. Here Mortality rate is calculated based on number of deaths of children under 5 years old per 1000 person. The histogram of this variable is shown in Figure 1a. We observe that the data ranges from 0 to 120, with a peak at 10-15.

However, this data is extremely right skewed and have several outlyers, thus we decided to perform log transformation to stabilize the variance and in-proving the normality of our dataset for a better regression model latter. The distribution after we perform log transformation is shown in Figure 1b, here the distribution is approximately normal ranging from 2 to 5 with a peak at 3.



(a) Dsitribution before log tranformation

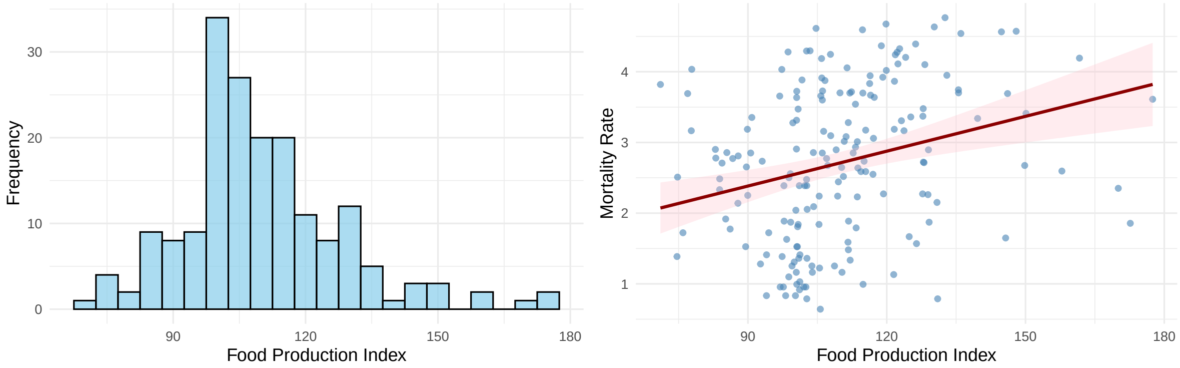
(b) Distribution after log transformation

Figure 1: Data Analysis for Mortality Rate of Different Countries

2.4 Predictor variables

In our mdoel, we have three Preditor variables, namely, Food production index, which is the relative level of agricultural production for each nation compared with the base period 2014-2016 (Arel-Bundock (2022)); Current health expenditure per capita (current US$); and DPT vaccine percentage.

We will first look at the food production index. From Figure 2a, we see that it ranges from 60 to 180, with a center at 110, and the shape is approximated normal distributed. Figure 2b is the scatter plot between Food production index and log of Mortality rate, with a best line of fit and standard error. We obsereve a slightly positive linear relationship between these two variables. Detailed relationship will be studied in Section 3.



(a) Histogram of each countries' food production index

(b) Scatter plot of each countries' Food production index VS. Log of Mortality

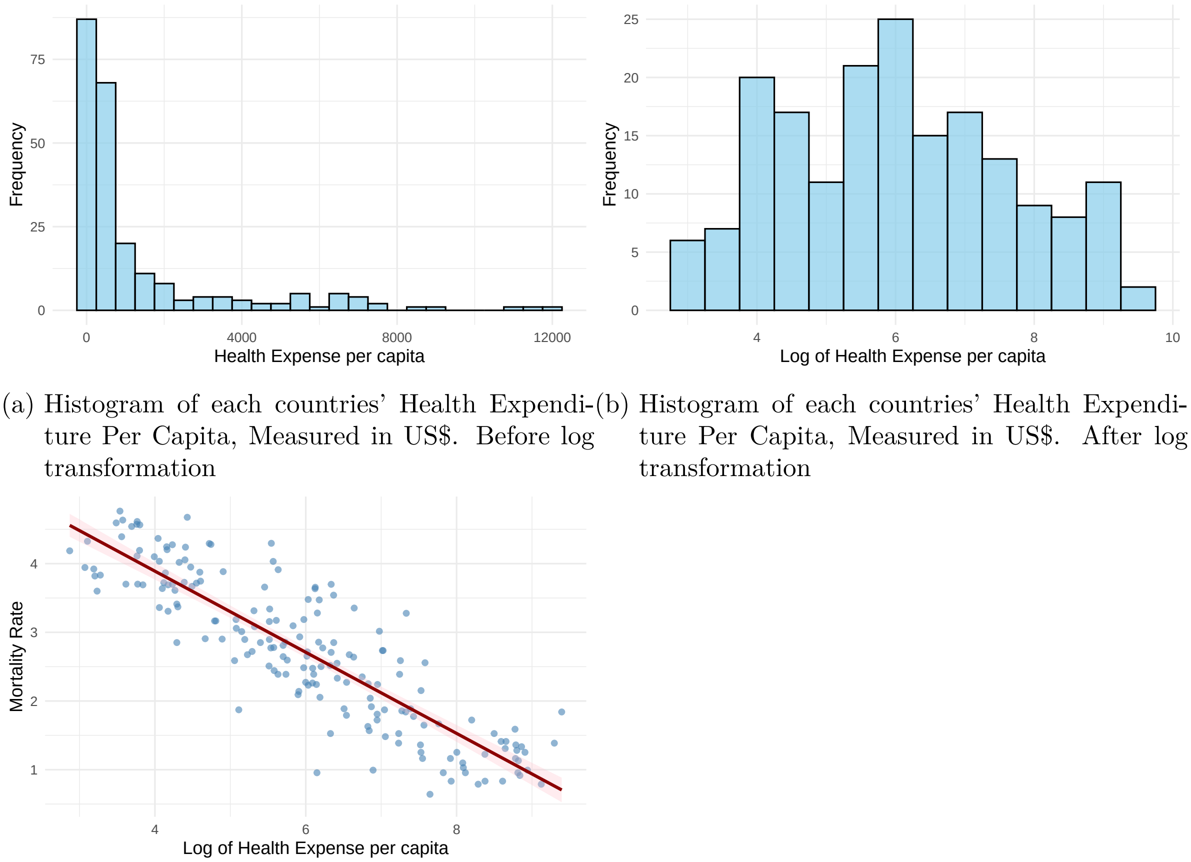
Figure 2: Data Analysis of Predictor Food Production Index

Next is the Current health expenditure. From Figure 3a, we see that it ranges from 0 to 12000, but the shape of the distribution is extremely skewed to the right. Thus we decided to perform a log transformation to stabilize the normal shape. Figure 3b is the histogram after the log transformation, we see that it now ranges from 0 to 10 with a center at 6 and the shape a approximately normal now. Figure 3c is a scatter plot between Log of Mortality rate and Log of Current health Expenditure, with a best line of fit and standard error. We observe a significant negative linear relationship between these two variables. Detailed relationshape will be studied in Section 3.

Finally is the DPT vaccine percentage. From Figure 4a, we see that it ranges from 30 to 100, but the shape of the distribution is skewed to the left, we will discuss this in Section 4. Here we do not have efficient tecnics to stablize the distribution. Figure 4b is a scatter plot between DPT vaccine percentage and Log of Mortality rate, with a best line of fit and standard error. We observe a significant negative linear relationship between these two variables. Detailed relationship will be studied in Section 3.

2.5 Missing Data and Time Inconsistancy

In our file, the varible Health Expendure is collect for year 2021 while the others is collected in year 2022. We done this is because there is n data avliable for year 2022 of this data. We believe that these data are recent and so the data collected in year 2021 could still do a good job associating data in 2022. Also there are a few data missing for some minor countries. We made the decision to drop them and continue our modeling. We will discuss further implications of these two in Section 4

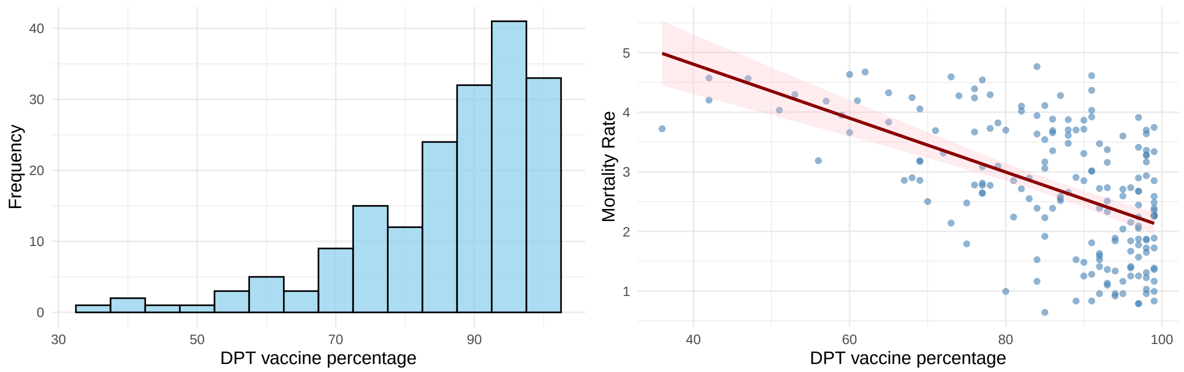


(a) Histogram of each countries' Health Expenditure Per Capita, Measured in US$. Before log transformation

(b) Histogram of each countries' Health Expenditure Per Capita, Measured in US$. After log transformation

(c) Scatter plot of each countries' Health Expenditure Per Capita, Measured in US$ VS. Log of Mortality

Figure 3: Data Analysis of Predictor Current Health Expenditure



(a) Histogram of each countries' DPT vaccine percentage

(b) Scatter plot of each countries' DPT Vaccine Percentage VS. Log of Mortality

Figure 4: Data Analysis of Predictor DPT Vaccine Percentage

3 Model

The goal of our modeling strategy is to investigate how the DDT vaccine coverage, food production index, and current health expenditure per capita (current US$) relate to the under-five mortality rate for each nation.

We aim to use this model to understand how the above mentioned three factors' impact on child mortality and identify opportunities for improving health outcomes, especially for high mortality rate nations. Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Let y_i be the logrithm of under-five mortality rate for nation i. We define the following predictors:

x_{1i}: DDT vaccine coverage for nation i. x_{2i}: Food production index for nation i. x_{3i}: Logrithm of current health expenditure per capita (current US$) for nation i. The model is specified as:

$$y_i \mid \mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$

$$u_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \tag{2}$$

$$\alpha \sim \text{Normal}(5,) \tag{3}$$

$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$

$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{5}$$

$$\beta_3 \sim \text{Normal}(0, 2.5) \tag{6}$$

$$\sigma \sim \text{Exponential}(1) \tag{7}$$

We implement this model in R (R Core Team 2023) using the rstanarm package (Goodrich et al. 2022), employing its default priors for predictors and a distribution of Normal(_i,).

3.1.1 Model justification

We use a multivariate linear model to capture the relationship between the predictors and the response variable. This choice is justified by the linear trends observed in the data (see Section 2). The logarithmic transformation of the under-five mortality rate and current health expenditure per capita is applied to stabilize variance and linear relationships, ensuring model validity.

The Bayesian framework is employed due to its ability to incorporate prior knowledge, improve uncertainty quantification, and handle small sample sizes effectively.

4 Results

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

###Intercept: The intercept estimate of 6.643 represents the expected logarithm of the under-five mortality rate when all predictors are held constant at their reference or baseline values (e.g., average food production, average vaccine coverage, and average health expenditure). This serves as a baseline for interpreting the effects of the predictors.

###Food: Food Production Index: The coefficient for the food production index is 0.003, indicating a very slight positive relationship between food production and the under-five mortality rate. This naïvely suggests that reducing food production could also reduce mortality, which is counter-intuitive.

However, a closer examination of the confidence interval for this predictor shows that it includes 0. This implies that the relationship is not statistically significant, and food production is not strongly related to overall under-five mortality rates. This finding suggests that while food production is essential for societal well-being, its immediate impact on reducing child mortality might depend on other factors such as food access, distribution systems, and nutritional quality.

###Vaccine Coverage: The coefficient for vaccine coverage is -0.013, meaning that for every 1 percentage point increase in vaccine coverage, the logarithm of the under-five mortality rate decreases by 1.3%, holding other variables constant.

This highlights the critical role of vaccination programs in reducing child mortality. For example, increasing vaccine coverage by 10 percentage points could reduce the mortality rate by approximately 13%, emphasizing the importance of robust immunization initiatives.

###Health Expenditure (Log-Transformed): The coefficient for the log-transformed health expenditure is -0.529, indicating that a 1% increase in health expenditure per capita is associated with a 0.529 unit decrease in the log of the under-five mortality rate, holding all other predictors constant.

Term	Estimate	Std. Error	2.5% CI	97.5% CI
(Intercept)	6.643	0.363	6.047	7.222
Food	0.003	0.002	0.000	0.007
Vacinne	-0.013	0.003	-0.019	-0.008
Health_expense	-0.529	0.029	-0.578	-0.482

“‘

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

Please don’t use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

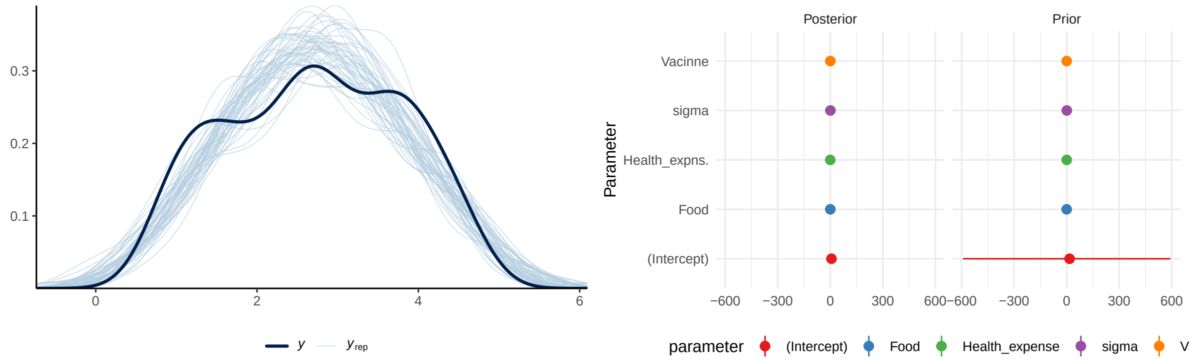
A Additional data details

B Model details

B.1 Posterior predictive check

In Figure 5a we implement a posterior predictive check. This shows...

In Figure 5b we compare the posterior with the prior. This shows...



(a) Posterior prediction check

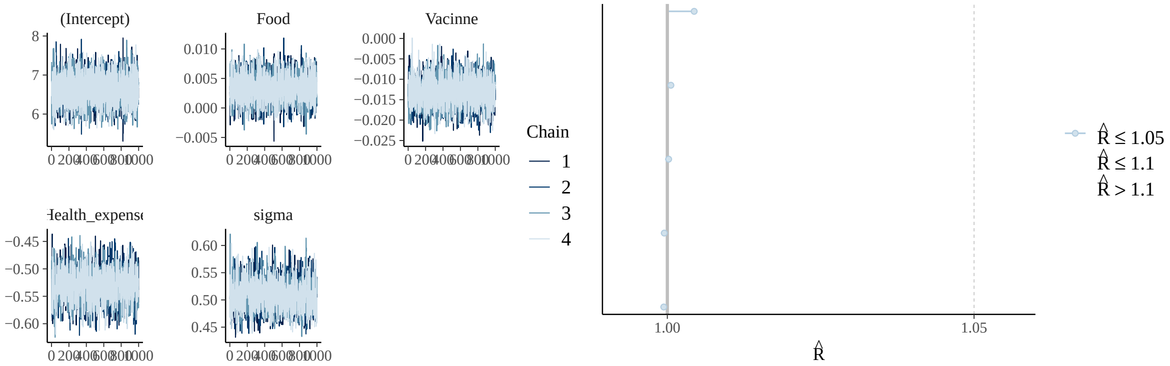
(b) Comparing the posterior with the prior

Figure 5: Examining how the model fits, and is affected by, the data

B.2 Diagnostics

Figure 6a is a trace plot. It shows... This suggests...

Figure 6b is a Rhat plot. It shows... This suggests...



(a) Trace plot

(b) Rhat plot

Figure 6: Checking the convergence of the MCMC algorithm

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Arel-Bundock, Vincent. 2022. *WDI: World Development Indicators and Other World Bank Data*. https://CRAN.R-project.org/package=WDI.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.