Examining Birthweight Determinants: How Maternal Age, Education, Baby's Sex, and Gestational Period Shape Neonatal Outcomes*

An Analysis of Birthweight in the United States in 1998

Ruiyang Pang

November 26, 2024

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

1 Introduction

The challenge of aging populations continues to hinder national economic growth, prompting governments to promote higher birth rates (Christensen et al. 2009). However, modern pressures such as intense job competition, housing shortages, and high living costs have led many young adults to opt for child-free lifestyles or limit their families to a single child. Within this context, the health of newborns becomes a critical focus, with birth weight serving as a key indicator. Low birth weight is associated with an increased risk of diseases like ischemic heart disease and chronic conditions later in life. Studies from the early 1980s highlight the link between low birth weight and fetal malnutrition, which can have lasting developmental effects (Paneth 1995). While economic and technological advancements have reduced malnutrition, identifying contemporary factors influencing birth weight is crucial. This study investigates the determinants of birthweight in today's society, exploring how various socioeconomic and biological factors interplay to shape neonatal outcomes.

Estimand paragraph

My estimand focuses on the relationship between birth weight and various factors. These factors include:

- **Parental Attributes**: The age and education level of the parents and their marital status.

*Code and data are available at: https://github.com/RuiyangPang/birthweight.git.

- **Pregnancy and Infant Factors**: Gestation length, prenatal care, the number of prior live births, and the infant's sex.

My goal is to determine which factors significantly affect birth weight, and I also want to understand how these factors contribute to the infant birth weight. This analysis aims to inform strategies for improving outcomes related to birth weight.

Results paragraph

Why it matters paragraph

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

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2022), and with the help of several packages in R (R Core Team 2022) to explore these data and build regression models. These packages are include: tidyverse (Wickham et al. 2019),knitr (Xie 2023a),dplyr (Wickham et al. 2023), here (Müller 2020), tinytex (Xie 2023b),readr (Wickham, Hester, and Bryan 2024).

Our data is published in the Data and Story Library (DASL) (DASL 2008), an archive containing hundreds of datasets designed for teaching statistics and data science. I used the baby sample dataset from DASL, a randomly selected subset of 200 records from the 1998 Natality Public Dataset provided by the National Bureau of Economic Research (NBER) (Economic Research 2008). This sample was generated using set.seed(1) in R to ensure reproducibility (R Core Team 2022). Although the NBER dataset provides comprehensive information on all U.S. births in 1998, the full dataset includes over 3.9 million records and occupies nearly 200MB. Such a large dataset poses practical challenges. First, it significantly slows down computational processing. More importantly, in regression analysis, large sample sizes can lead to misleading results. Even minimal effect sizes may appear statistically significant but lack practical relevance, resulting in erroneous conclusions about predictor variables' importance. Following Alexander (2023), we consider random sampling can help us to get a valid dataset to represent the population. The selected dataset contains 7 predictors and 1 outcome variable, with 200 observations, which is sufficient for meaningful regression analysis. we effectively reduced the dataset size while maintaining its representatives of the population. After the data clean, Table 1 shows a preview of the cleaned data set.

Table 1: Preview of the cleaned 1998 birth data set

weight	MomAge	MomEduc	MomMarital	gestation	sex	prenatalstart
3175	35	17	1	39	F	1
3884	22	12	1	42	F	2
3030	35	15	1	39	F	2
3629	23	6	1	40	F	1
3481	23	13	1	42	F	2
3374	26	12	2	39	M	4

2.2 Measurement

The Natality Public Dataset is a comprehensive collection of U.S. birth records processed by the National Center for Health Statistics (NCHS). It includes detailed information on maternal and infant health, family demographics, and other birth-related variables. The dataset records births for both U.S. residents and non-residents but excludes births of U.S. citizens occurring outside the country.

In terms of measurement precision, the dataset replaces exact dates of birth with month and year to protect the privacy of newborns and parents. Geographic details are also simplified, making it less relevant for studies focusing on location-based factors. However, the focus of this dataset is on individual and family characteristics rather than geographic influence.

The dataset is collected officially through birth registration processes required for legal documentation, ensuring high coverage and minimal data missingness. Although the dataset is reliable, some variables, such as paternal age and education, have higher rates of missing data. This discrepancy is due to the lesser involvement of fathers in birth registration and some families lacking complete paternal information.

Despite these limitations, the dataset has low measurement error in variables like infant sex, gestational age, and maternal education. The 1998 dataset includes data for special areas like Northern Mariana Islands and American Samoa, providing an extensive range of records. Given its official source, the dataset serves as a robust foundation for analyzing maternal and infant health factors.

2.3 Outcome variables

Birthweight is reported in some areas in pounds and ounces rather than grams. To ensure consistency, we present statistical data using the metric unit of grams. A birth weight below 2500 grams is classified as low birth weight, while a weight below 1500 grams is considered very low birth weight (Health Statistics (NCHS) 2008). Newborns with low birth weight often require specialized care, such as admission to a Neonatal Intensive Care Unit (NICU).

Figure 1 shows the distribution of the birth weight in the sample data set. Among the observations, 2 cases are classified as low birth weight, representing approximately 1.09 % of the sample. The mean birth weight is 3318.22 grams, with the lowest recorded birth weight being 1729 grams. The distribution of birth weights appears to be approximately normal. Given that the data was collected through random sampling, a linear regression model can be appropriately applied to analyze the relationships between birth weight and the selected predictors.

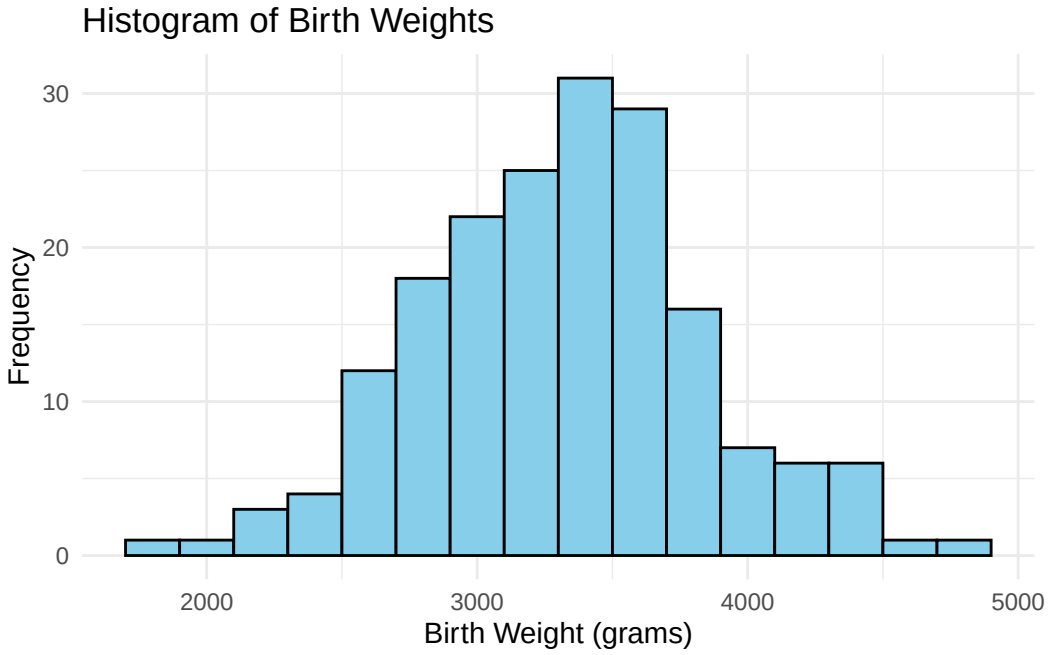

Figure 1: Birth weight of United States newborn in 1998

2.4 Predictor variables

There are 4 numerical and 2 categorical predictors in the sample. The summary of these numerical predictors are shown in Table 2 and their distribution are shown in Figure 2.

In this sample, maternal age ranges from 14 to 42 years, as shown in Figure 2. The distribution of maternal age is concentrated between 20 and 40 years, with the number of births decreasing steadily after age 30. Maternal age is obtained through official documentation, ensuring high accuracy.

Additionally, maternal education is mostly concentrated at 12 years, corresponding to high school completion. This peak likely reflects that many mothers choose to enter the workforce after high school. Another peak occurs at 16 years, aligning with the typical length of a university education, as completing a four-year degree results in 16 years of education.

The mode of gestational length is 40 weeks, with a sharp decline in births beyond 40 weeks. However, the probability of birth increases after 36 weeks, as 36 weeks marks the threshold for full-term infants, and delivery can occur anytime after this point.

Prenatalstart refers to the timing of when prenatal care begins during pregnancy. Earlier initiation of prenatal care is essential for promoting fetal development and supporting maternal health. In this dataset, the mode of prenatalstart is 2, indicating that most mothers began prenatal care during their second month of pregnancy. Furthermore, the majority of these values are less than 5, reflecting that in the United States, parents are generally able to detect pregnancy early and provide timely and necessary prenatal care.

Table 2: Statistics summary of the numerical predictors

MomAge	MomEduc	gestation	prenatalstart
Min. :14.00	Min. : 2.00	Min. :33.00	Min. :0.000
1st Qu.:22.00	1st Qu.:12.00	1st Qu.:38.00	1st Qu.:2.000
Median :26.00	Median :12.00	Median :39.00	Median :2.000
Mean :27.01	Mean :12.79	Mean :39.11	Mean :2.393
3rd Qu.:32.00	3rd Qu.:15.00	3rd Qu.:40.00	3rd Qu.:3.000
Max. :42.00	Max. :17.00	Max. :47.00	Max. :9.000

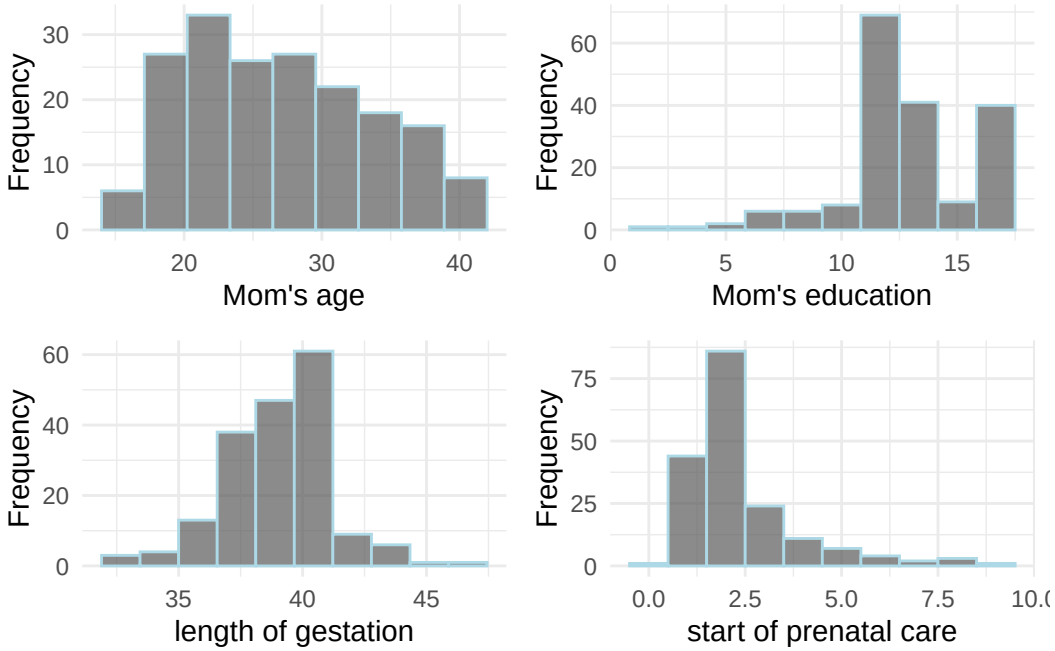

Figure 2: histogram of predictors

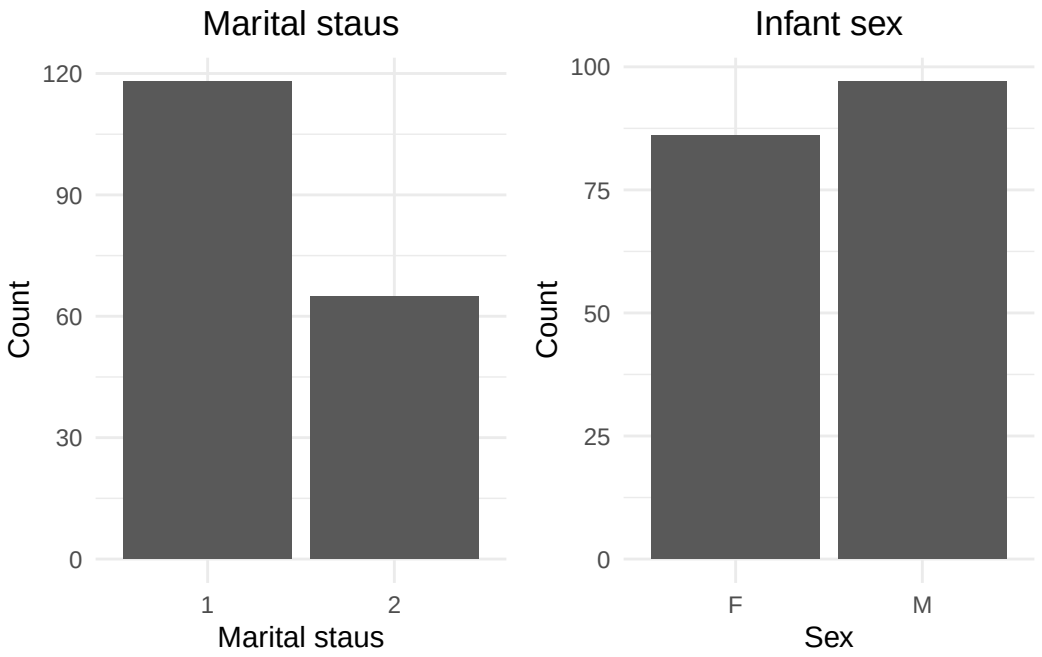

Figure 3: Barplot of predictors

The distribution of the two categorical variables are shown Figure 3. From the barplot of marital status, about of the 64.48% mom is married. The proportion of male infant is a little higher than female infant.

Figure 4 shows that as the length of gestation increases, birth weight also tends to rise. However, the correlation between gestation and birth weight is only 40.01%. This suggests that birth weight is not solely dependent on gestation but is also influenced by other factors. During the prenatal care phase, various parental attributes, pregnancy-related factors, and infant characteristics could all play a role in shaping birth weight. Therefore, in the following sections, I will use regression analysis to explore the impact of these factors on birth weight.

3 Model

The goal of our modeling strategy is twofold. First, we aim to investigate the influence of Parental Attributes (such as parental age and education) and Pregnancy and Infant Factors (like gestational age, timing of prenatal care, and infant sex) on newborn birth weight. This analysis seeks to identify the critical determinants of birth weight and their relative impacts. Second, we aspire to derive insights from these findings that can inform the public about fundamental aspects of childbirth. For instance, we explore the roles of parental characteristics, the gestation period, and prenatal care in fostering healthy birth weights. Moreover, by including

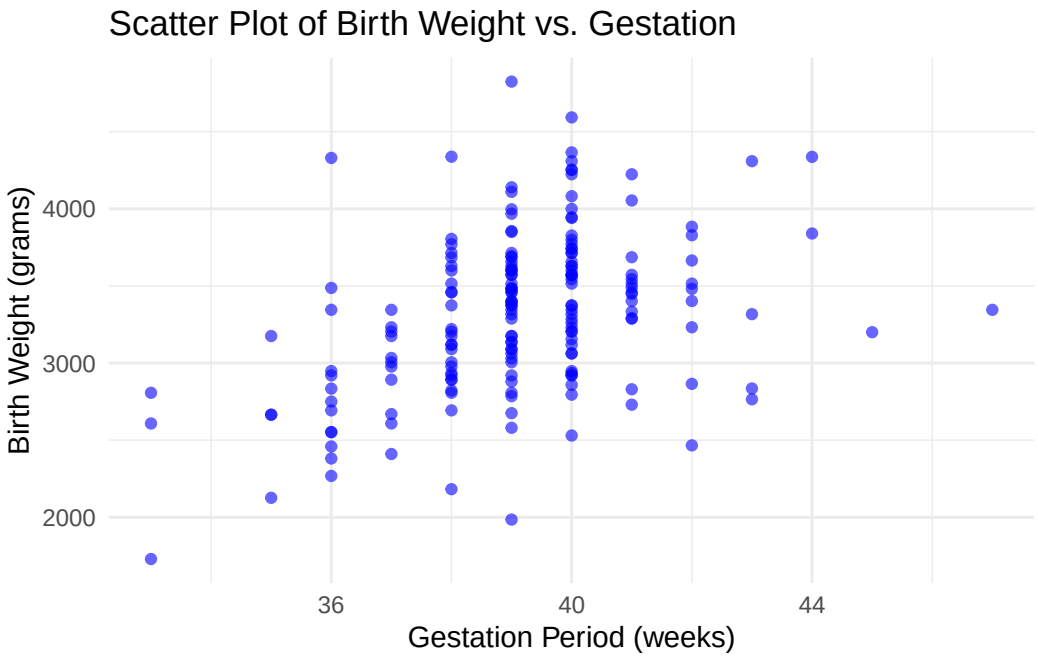

Figure 4: Correlation between birth weight and gestation

infant sex in the model, we aim to mitigate omitted variable bias, ensuring a more accurate understanding of predictor effects. This comprehensive approach not only contributes to the academic study of birth outcomes but also provides practical insights for improving maternal and neonatal health practices.

Linear regression models are often used to estimate relationships between continuous outcome variables and multiple predictors. However, the validity of linear regression can be compromised by various challenges, such as the assumption of linearity in predictors, homoscedasticity of errors, independence of observations, and the influence of outliers (Alexander 2023). To address the limitations of traditional linear regression models, this study employs a Bayesian analysis model.

In the Bayesian framework, model parameters are treated as random variables with prior distributions that reflect initial beliefs or knowledge about these parameters (Goodrich et al. 2022). By combining the information from sample data with prior distributions, we derive the posterior distributions of the parameters. This posterior distribution provides a more robust basis for inference and prediction, particularly in cases where data assumptions are violated or sample sizes are small. Using Bayesian analysis allows for integrating prior knowledge and updating it with observed data, yielding potentially more accurate and nuanced results.

3.1 Model set-up

I will use the Bayesian multiple linear regression model to build the relationship between birth weight and other predictors. Define y_i as the birth weight in grams. Then the model is shown in the following:

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \beta_0 + \beta_1 MomAge_i + \beta_2 MomEduc_i + \beta_3 gestation_i + \beta_4 prenatalstart_i + \\
&\quad \gamma_1 MomMarital_i + \gamma_2 sex_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
\beta_2 &\sim \text{Normal}(0, 2.5) && (5)\\
\beta_3 &\sim \text{Normal}(0, 2.5) && (6)\\
\beta_4 &\sim \text{Normal}(0, 2.5) && (7)\\
\gamma_1 &\sim \text{Normal}(0, 2.5) && (8)\\
\gamma_2 &\sim \text{Normal}(0, 2.5) && (9)\\
\sigma &\sim \text{Exponential}(1) && (10)
\end{aligned}
$$

We run the model in R (R Core Team 2022) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the length of gestation and birth weight, as longer gestation generally results in higher birth weights. Maternal age is another important factor influencing infant health. Older mothers, especially those over the age of 35, often face higher risks during pregnancy, which can impact fetal development (Frederiksen et al. 2018). Maternal education level also plays a crucial role, as more educated mothers are more likely to understand and follow scientifically supported prenatal care practices, benefiting fetal growth. Additionally, marital status can influence pregnancy outcomes. Married women may receive more support from their partners, leading to better nutrition and healthcare, which in turn benefits fetal development. Furthermore, the earlier prenatal care begins, the more favorable it is for the fetus's health. Given these considerations, we have selected these factors as predictors for our model. To account for potential biases due to infant gender, we have included it in the model to reduce omitted variable bias.

Table 3: Linear regressin model on the birth weight

	First model
(Intercept)	−1647.58
	(758.31)
MomAge	7.70
	(6.45)
MomEduc	26.55
	(14.22)
MomMarital	−50.12
	(81.63)
gestation	114.06
	(17.47)
as.factor(sex)M	43.90
	(70.77)
prenatalstart	0.56
	(23.78)
Num.Obs.	183
R2	0.214
R2 Adj.	0.187
AIC	2776.9
BIC	2802.6
Log.Lik.	−1380.470
RMSE	456.97

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.

Table 4: Bayesian regressin model on the birth weight

	First model
(Intercept)	−1637.58
MomAge	7.58
MomEduc	26.59
MomMarital	−51.97
gestation	114.22
as.factor(sex)M	42.34
prenatalstart	0.40
Num.Obs.	183
R2	0.223
R2 Adj.	0.145
Log.Lik.	−1381.764
ELPD	−1389.6
ELPD s.e.	11.5
LOOIC	2779.3
LOOIC s.e.	23.0
WAIC	2779.1
RMSE	457.07

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 Diagnostics

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

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

References

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

Christensen, Kaare, Gabriele Doblhammer, Roland Rau, and James W Vaupel. 2009. "Ageing Populations: The Challenges Ahead." *The Lancet* 374 (9696): 1196–1208.

DASL. 2008. "Birthweight Dataset." Online. https://dasl.datadescription.com.

Economic Research, National Bureau of. 2008. "NBER Natality Data Project." https://data.nber.org/natality/ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/DVS/natality/.

Frederiksen, Line Elmerdahl, Andreas Ernst, Nis Brix, Lea Lykke Braskhøj Lauridsen, Laura Roos, Cecilia Høst Ramlau-Hansen, and Charlotte Kvist Ekelund. 2018. "Risk of Adverse Pregnancy Outcomes at Advanced Maternal Age." *Obstetrics & Gynecology* 131 (3): 457–63.

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

Health Statistics (NCHS), National Center for. 2008. "User Guide to the 2008 Natality Public Use File." National Center for Health Statistics, Centers for Disease Control; Prevention. https://www.cdc.gov/nchs/data_access/VitalStatsOnline.htm.

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

Paneth, Nigel S. 1995. "The Problem of Low Birth Weight." *The Future of Children*, 19–34.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.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.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2023a. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

———. 2023b. *Tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents*. https://github.com/rstudio/tinytex.