Regression Analysis of the U.S. Teen Birth Rate in 2023

Abstract

Adolescent childbearing carries significant social, health, and economic consequences for individuals and society. Although teen birth rates have declined by approximately 68% from 2007 to a historic low of 13.1 in 2023, this study examines factors associated with variation in state-level teen birth rates in 2023. We construct a cleaned, multi-source state-level dataset by integrating five data sources and converting rates into counts using population data. Predictors include demographic variables (`Divorce.rate`, `Dropout.Rate`, `Political.Favor`, `Abortion`), socioeconomic indicators (`Median.Income`), and geographic `Region`. We estimate Poisson and Negative Binomial regression models and apply backward elimination for variable selection. Model performance is evaluated using the Akaike Information Criterion (AIC), residual deviance, and incidence rate ratios. Due to evidence of overdispersion, the Negative Binomial model provides a better fit, although overall performance remains limited (AIC = 700.07, deviance = 52.52). Results suggest that region and median income are the most consistent predictors of variation in teen births across states.

Introduction

Adolescent childbearing remains an important public health and social issue in the United States, given its well-documented associations with adverse health outcomes, reduced educational attainment, and long-term economic challenges for young parents and their families. Despite these concerns, national trends over the past two decades show substantial progress. Data from the National Center for Health Statistics (NCHS), part of the Centers for Disease Control and Prevention (CDC), indicate that the U.S. teen birth rate has declined markedly, falling by approximately 68% between 2007 and 2023. This downward trend culminated in a historic low of 13.1 births per 1,000 females aged 15 to 19 nationwide in 2023. Despite this decline, the United States still reports higher teen birth rates compared to other industrialized countries. As a result, understanding the factors that continue to influence teen birth rates remains critical. Focusing on 2023 data, this regression analysis aims to examine the key demographic, socioeconomic, and geographic variables associated with variation in teen birth rates at this recent low point to provide insights to inform targeted state-level policy and public health interventions in the U.S. As a result, we came up with the research question: On a state level, how are geographical regions, political favor, abortion policy, median household income, divorce rate, and school dropout rate associated with teen birth rate in 2023?

Data Summary

The data were compiled from five sources for the year 2023, except for political favor. Teen birth and divorce data were obtained from the National Center for Health Statistics (NCHS) CDC annual tables. Socioeconomic variables, including median household income, dropout rate, and teen population, were sourced from the American Community Survey (ACS) 2019–2023. Geographic regions were assigned using the U.S. Census Bureau's four-region classification by manually constructing an attribute table based on CDC documentation. Abortion policy data were drawn from the KFF's "Abortion in the United States Dashboard". Political favor was measured using 2020 presidential electoral vote outcomes by state, compiled by The American Presidency Project, UCSB. With a limited number of observations for the gestational limit levels of the abortion variable, 3 categories of gestational-limit policies were combined into a single category labeled "With Some Gestational Limits."

Teen birth rate (Y) is defined as the number of live births per 1,000 females aged 15–19 in each state in 2023. The model response, teen birth count (Y*), is derived by scaling the rate using the total female teen population (`Total.Female.Teen`) following the formula: $Teen.Birth.Count = \frac{Teen.Birth.Rate}{1000} \times Total.Female.Teen$, where the total female teen population is incorporated as an offset term in the model. Categorical predictors consist of geographic region (`Region`, four levels), political favor (`Political.Favor`, two levels), and abortion policy (`Abortion`, three levels after recoding). Quantitative predictors include median household income (`Median.Income`, in dollars), divorce rate per 1,000 individuals (`Divorce.rate`), and dropout rate per 1,000 individuals (`Dropout.Rate`).

Exploratory Data Analysis (EDA)

We examined histograms for teen birth rate (Figure 1) and teen birth count (Figure 2) to guide model selection. From Figure 1, we found that `Teen.Birth.Rate` shows a slightly right-skewed shape in the histogram without outliers, which matches the Poisson Distribution. The model response `Teen.Birth.Count` histogram shows an extremely right-skewed distribution (Figure 2). Because the response variable is a nonnegative count outcome, a Poisson model was initially considered.

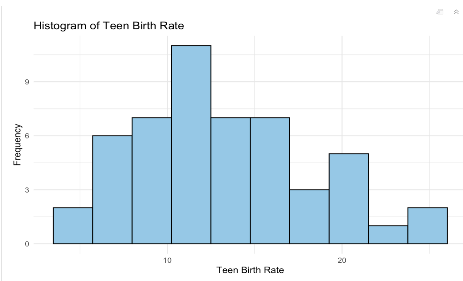


Figure 1: Histogram of Teen Birth Rate

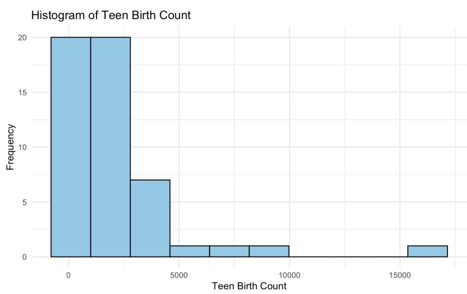


Figure 2: Histogram of Teen Birth Count

However, from the data summary (Appendix A), we can see that the variance (7170887) was much larger than the mean (2194). This suggests overdispersion, so negative binomial regression could also be considered.

From the scatterplots (Appendix B), we found that `Median.Income` showed a negative relationship with the response, meaning states with higher income levels tended to have lower teen birth rates. `Divorce.Rate` appeared to have a weak positive relationship, with a potential curved pattern. For the `Dropout.Rate`, there was a strong positive linear relationship with the response. We also examined multicollinearity among predictors and found a moderate negative correlation between `Median.Income` and `Dropout.Rate` (≈ -0.64), indicating that states with higher income levels tend to have lower dropout rates.

Model Analysis

Following EDA, we fitted an initial Poisson regression model with all predictors (Region, Political Favor, Abortion, Median Income, Divorce Rate, and Dropout Rate), using Teen.Birth.Count as the response variable and Total.Female.Teen as the offset term. As suggested by the EDA, an overdispersion test indicated substantial overdispersion ($p = 2.49 \times 10^{-7}$, dispersion estimate ≈ 22), providing evidence that the dispersion parameter exceeded 1. Therefore, the Poisson assumption was violated.

We begin with a full Negative Binomial regression model that includes all predictors and an offset term. Before model reduction, we evaluate key assumptions to ensure the validity of subsequent hypothesis tests. Given that `Teen.Birth.Count` is a count-based response, and each state is observed once; the independence assumption is satisfied. The component + residual (CR) plots indicated no major departures from linearity (Figure 3), and all VIF and adjusted GVIF values were below 2, suggesting no serious multicollinearity. Overdispersion is confirmed, with an estimated dispersion parameter of 40.21, well above 1, supporting the use of a Negative Binomial model. Finally, deviance residuals and Cook's distance are used to assess outliers and influential observations; no severe outliers are detected. Given that all key assumptions are reasonably satisfied, we proceed with model reduction using significance-based selection.

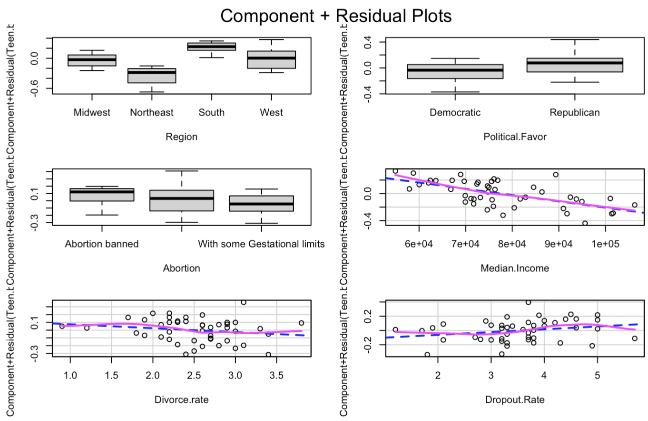


Figure 3: CR plots for the full model

After that, we used backward elimination based on the p-values of individual significance tests for each coefficient. For categorical predictors, if at least one level was significant, we kept the whole variable in the model. If none of the levels were significant, the variable was considered for removal. At the same time, we also referred to the Pr(>Chi) results to evaluate whether the variable was significant overall. `Abortion` (p = 0.59 and 0.10 for each level), `Divorce.Rate` (p = 0.37), `Dropout.Rate` (p = 0.173), and `Political.Favor` (p = 0.081) was removed due to its large p-value. Therefore, we got our final model with Region and Median Income as the main predictors, and our final Negative Binomial regression model is:

$$\log(\mu_i) = -3.285 - 0.382(Region_{Northeast}) + 0.276(Region_{South}) + 0.054(Region_{West}) - 1.458 \times 10^{-5}(MedianIncome) + \log(TotalFemaleTeen_i)$$

$$\log(Teen.Birth.Rate) = -3.285 - 0.382Region_{Northeast} + 0.276Region_{South} + 0.054Region_{West} - 1.458 \times 10^{-5}Median.Income$$

where μi represents the expected teen birth count for each state.

We also checked the IRR and Confidence Interval (Table 1). Using the Midwest as the reference category, states in the Northeast had 32% lower teen birth rates than those in the Midwest, while states in the South had 32% higher rates, controlling for other

Characteristic	IRR	95% CI	p-value
Region			
Midwest	—	—	
Northeast	0.68	0.59, 0.79	<0.001
South	1.32	1.17, 1.48	<0.001
West	1.06	0.92, 1.20	0.4
Median.Income	1.00	1.00, 1.00	<0.001
Abbreviations: CI = Confidence Interval, IRR = Incidence Rate Ratio			

Table 1: Incidence Rate Ratio (IRR) and Confidence Interval (CI) for the reduced model

variables in the model. In contrast, Western states were not significantly different from those in the Midwest. The confidence intervals for the Northeast and South effects did not include 1, indicating statistical significance. For median income, the incidence rate ratio (IRR) is approximately 1, suggesting no clear relationship with teen birth rates; however, this may reflect rounding and scaling effects, as income is measured in dollars and requires a very small coefficient to capture meaningful percentage changes. Overall, the results indicate that both regional differences and economic conditions play important roles in explaining variation in teen birth rates across states.

Model assumption diagnostics show that key assumptions are generally satisfied, including independence, absence of multicollinearity, and evidence supporting the use of a count model with overdispersion. However, a mild nonlinear pattern is observed for median income in the CR plots (Figure 4), suggesting the linearity assumption is only weakly met. Additionally, Alaska is identified as the only state with a deviance residual greater than 2.

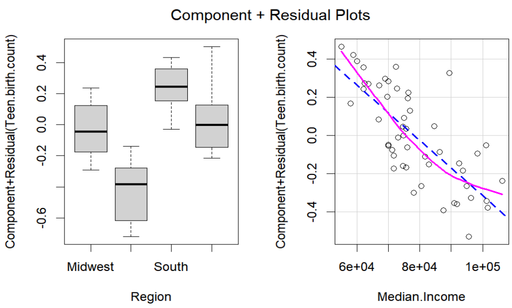


Figure 4: CR plots for the reduced model

Based on the diagnostic results and model evaluation, the final Negative Binomial model was deemed appropriate for analyzing teen birth counts and effectively addressed the issue of overdispersion.

Conclusion, Discussion, and Limitations

In conclusion, the Negative Binomial model provided a better fit than the Poisson model due to overdispersion in the teen birth counts. After the model selection process, Only `Region` and `Median.Income` remained in the final model. Our final Negative Binomial regression equation is:

$$\log(Teen.Birth.Rate) = -3.285 - 0.382Region_{Northeast} + 0.276Region_{South} + 0.054Region_{West} - 1.458 \times 10^{-5} Median.Income$$

From the estimated coefficients, we found that, compared with the Midwest, states in the Northeast had lower teen birth rates, while states in the South had higher teen birth rates, holding other variables constant. States in the West were not significantly different from the Midwest, since the West only had 5.5% ($e^{0.054} \approx 1.055$) higher teen birth rate than the Midwest. For the Median income, it also showed a negative relationship with teen birth rate; for every $10,000 increase in median income, the expected teen birth rate decreases by about 13.6% ($e^{-1.485\times10^{-5}\times1000} \approx 0.864$), holding other variables constant. Therefore, the results suggest that both economic conditions and regional factors are related to variation in teen birth rate across states. Since the model included an offset for population size, the observed associations with teen birth rate can also reflect differences in teen birth rates across states after adjusting for population size.

Although our model identified some meaningful predictors, there are still some limitations. First, for the final model performance, it has a high AIC (700.07) and large residual deviance (52.52), which suggests that there may still be unexplained variation in the data. The linearity assumption for the median income is barely met, which may affect the reliability of coefficient estimates and hypothesis tests. Future work could improve model specification by incorporating polynomial terms. Second, the study was based on a relatively small sample size (51-5=46 states) for using cross-sectional data from a single year (2023), which limits generalizability over time. Extending the dataset to multiple years or using a time-series or panel framework could better capture temporal trends and strengthen inference. Third, teen birth counts were derived from reported birth rates rather than obtained directly from the original data source, which may reduce measurement accuracy. Future studies could improve accuracy by using a better source of data that has the raw teen birth counts. Additionally, missing divorce data for several states (e.g., CA, HI, IN, MN, NM) reduced the sample size and may have impacted representativeness; future work should address missingness through improved data sources or imputation strategies.

Overall, while these limitations should be considered when we interpret our findings, the analysis still provides useful and meaningful evidence that regional and Median Income are associated with differences in teen birth rate across states.

References

Introduction:

Mickler, A., & Tollestrup, J. (2025). Teen Birth Trends: In Brief. Congress.gov. https://www.congress.gov/crs-product/R45184

Data:

CDC. (2023, June 26). *Geographic division or region - Health, United States*. Www.cdc.gov. https://www.cdc.gov/nchs/hus/sources-definitions/geographic-region.htm

CDC. (2025a, August 19). *Divorce*. Stats of the States. https://www.cdc.gov/nchs/state-stats/more-maps/divorce.html

CDC. (2025b, August 20). *Teen Births*. Stats of the States. https://www.cdc.gov/nchs/state-stats/births/teen-births.html

kffannab. (2025, July 25). *Abortion in the United States Dashboard | KFF*. KFF. https://www.kff.org/womens-health-policy/abortion-in-the-u-s-dashboard/

Median Household Income (In 2023 Inflation Adjusted Dollars) (American Community Survey (ACS) 2019--2023 (5-Year-Estimates)). In SocialExplorer.com. Retrieved 17 April 2026 from https://www.socialexplorer.com/data/ACS2023_5yr/metadata?ds=SE&table=A14006.

School Dropout Rate for Female Population 16 to 19 Years (American Community Survey (ACS) 2019--2023 (5-Year-Estimates)). In SocialExplorer.com. Retrieved 17 April 2026 from https://www.socialexplorer.com/data/ACS2023_5yr/metadata?ds=SE&table=A12003B .

UC Santa Barbara. (2021, March 3). *2020 | The American Presidency Project*. Www.presidency.ucsb.edu. https://www.presidency.ucsb.edu/statistics/elections/2020

Generative AI Statement

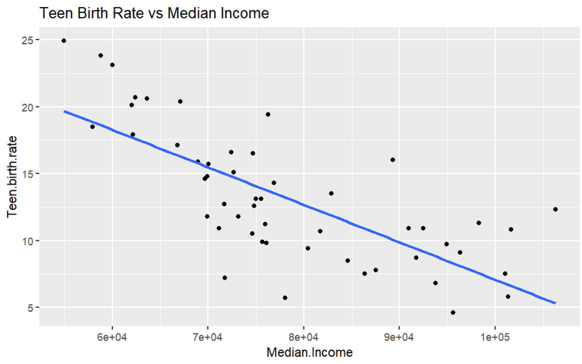
Generative AI tools were used solely to improve grammar, wording, and clarity of expression. ChatGPT was used to revise and shorten selected passages based on the students' original text. All ideas, analyses, results, and conclusions were developed, verified, and reviewed independently by the student authors.

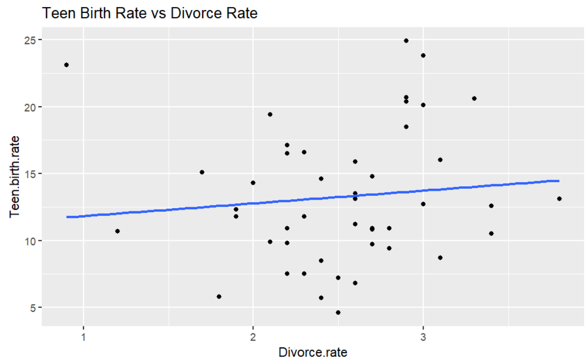
Appendix A: Data Summary

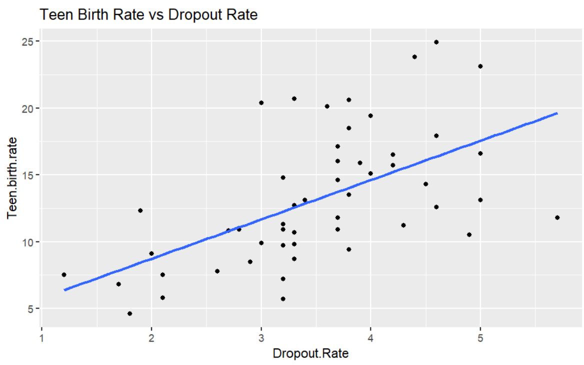
	Mean ± Std dev	Min - Max
Teen.Birth.Rate	13.18±4.95	4.6 – 24.9
Teen.birth.count	2194±2677.85	103 – 16254
Median.Income	78195±12978.45	54915 – 106287
Divorce.rate*	2.5±0.55	0.9 – 3.5
Dropout.Rate	3.5±0.95	1.2 – 5.7
Region	Midwest: 12 Northeast: 9 South: 17 West: 13	---------------------
Political.Favor	Democratic: 26 Republican: 25	---------------------
Abortion	Abortion banned: 13 No gestational limits: 10 With some Gestational limits: 28	---------------------
Total.Female.Teen	168463 ± 193055.3	14826 – 1014874

Appendix B: Exploratory Data Analysis Plots

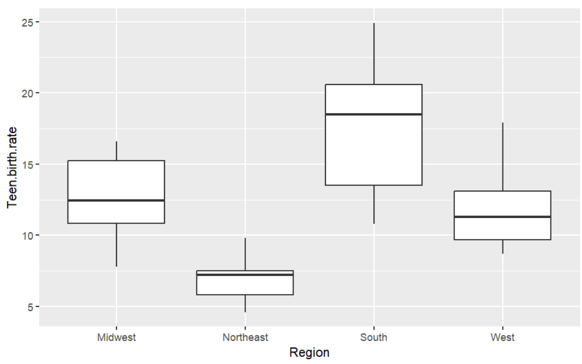
Quantitative Predictors vs. Teen Birth Rate Scatterplots

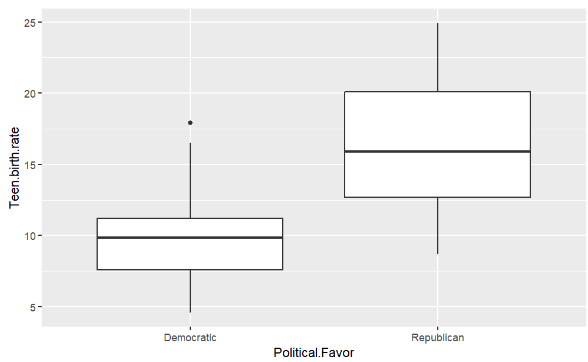


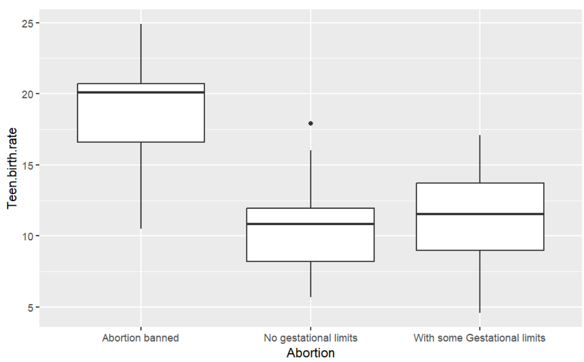




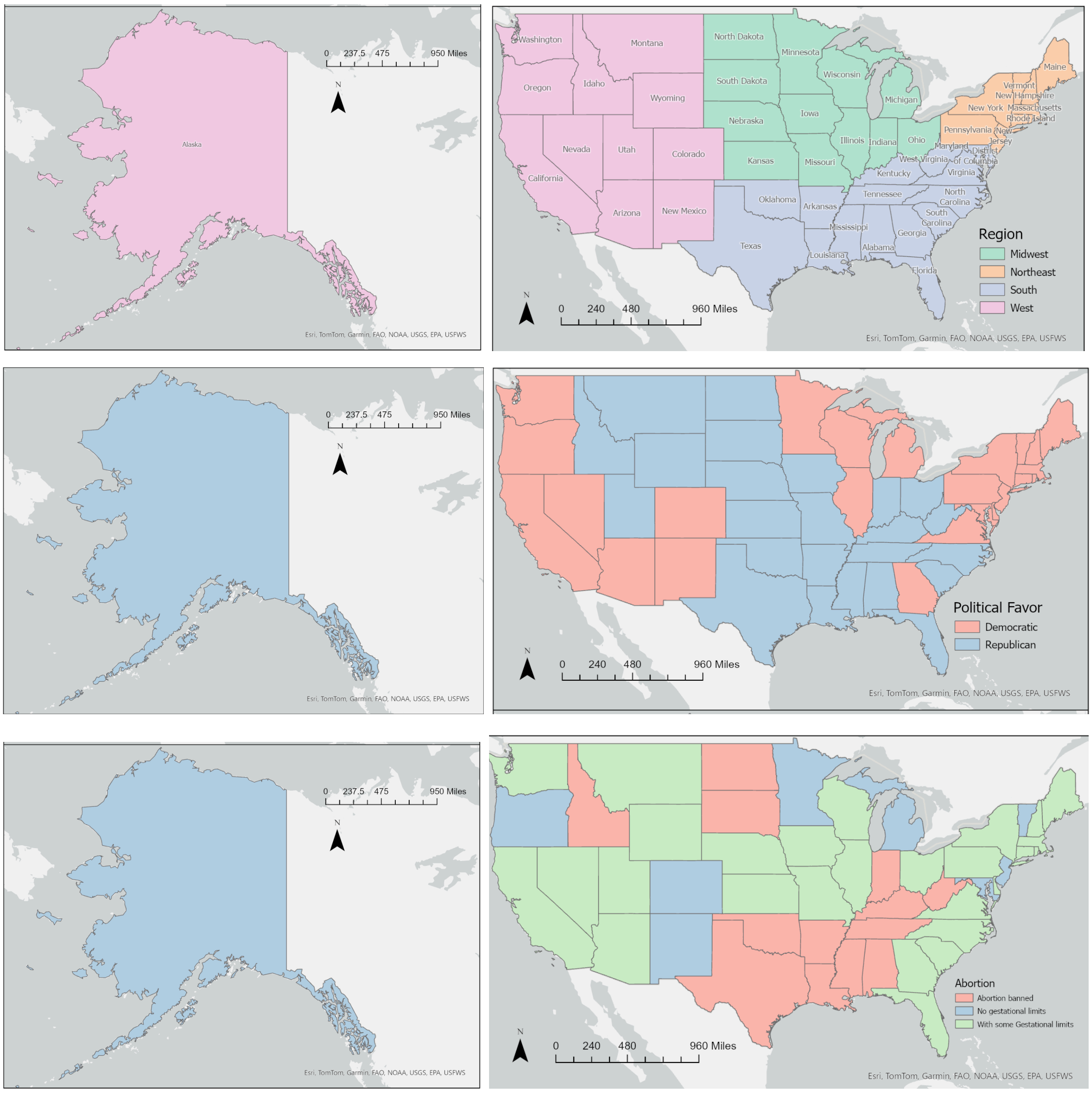
Categorical Predictors vs. Teen Birth Rate Boxplots







Categorical Predictors Distribution Maps



Multicollinearity Plots