Final Project

Birth data: births and infant deaths in the U.S. 2010-present across some demographic subgroups

Hon Chi Chan

1. Introduction

This report concerns births and infant mortality in the United States. Our analysis focuses on understanding the factors influencing birth rates using a linear regression model with three predictors: the proportion of mothers without a high school diploma, the proportion of Medicaid-covered births, and the proportion of mothers with college education. In this analysis, I aim to thoroughly investigate the relationship between fertility rates and educational attainment using linear regression models. Given the academic nature of this work, it is essential to present a detailed methodology to ensure a comprehensive understanding of the analytical approach. This not only supports educational objectives but also provides a foundation for replicability and deeper insights into the analytical process

Research Question:

How do educational attainment and Medicaid coverage among mothers influence the infant birth rates across different states and time periods?

The first few rows of the test data are shown below: Births and population by state 2016-2023: The columns are:

Column name	Meaning
`state`	51 states and DC
`year`	Calendar year of the data
`bmcode`	Bimonthly code. Data is aggregated into 2 month chunks. 1 means January 1 through end of Feb. 6 corresponds to data from November-December.

Column name	Meaning
`births_nohs`	Births to mothers without high school education
`births_coll`	Births to mothers with college degree
`births_medicaid`	Births covered by Medicaid
`births_total`	Total births across all categories
`pop_total`	Total population in the area
`pop_nohs`	Population without high school education (women age 15-54 population)
`pop_medicaid`	Population covered by Medicaid (women age 15-54 population)
`pop_coll`	Population with college degree (women age 15-54 population)

2. Simple Linear Regression Explanation

This section provides a theoretical foundation for the analysis, explaining the matrix formulation of the linear regression model. It's beneficial if your audience needs clarity on the mathematical underpinnings of the model.

To start, I'll fit a simple linear model regressing birth rates on educational attainment and Medicaid coverage. First, I need to organize our data into the necessary components—the response and explanatory variables—required for fitting the model in the correct format. Recall that the linear regression model, in matrix form, is represented as:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad ,$$

- Represents the birth rates for different observations(states and time periods).

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{bmatrix}$$

- Contains a column of ones for the intercept and columns representing:

x_{ij1}: Proportion of mothers without a high school diploma.

x_{ij2}: Proportion of Medicaid-covered births.

x_{ij3}: Proportion of mothers with a college education.

This is the model coefficients:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

This is the captures the error terms:

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

This setup allows us to systematically evaluate the influence of education and Medicaid coverage on birth rates. The matrix includes a column of ones for the intercept, which is crucial for estimating the model coefficients.

3. Preparing the Design Matrix for Linear Regression Analysis

This is a part of preparing data for a linear regression analysis. It involves creating a design matrix that includes both the response variable and explanatory variables.

Organizing Data: This setup organizes the data into a format suitable for fitting a linear regression model, where y is the dependent variable and x contains the independent variables.

Modeling Framework: By using the design matrix, the regression model can systematically estimate the relationships between birth rates and the explanatory factors of educational attainment and Medicaid coverage.

4. Estimation

In this section, I apply the linear regression model to our data, focusing on extracting key statistical parameters such as coefficients, variance estimates, and interpreting the results. This process involves fitting the model to our dataset and using statistical software to derive meaningful insights.

```
Call:
lm(formula = birth_rate ~ prop_nohs + prop_medicaid + prop_coll,
    data = birth_data_clean)

Residuals:
       Min          1Q      Median          3Q         Max
-2.431e-03 -2.526e-04 -5.310e-06  2.665e-04  2.373e-03

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.148e-03  7.173e-05   29.94   <2e-16 ***
prop_nohs     1.625e-03  4.871e-05   33.36   <2e-16 ***
prop_medicaid 1.108e-03  4.576e-05   24.21   <2e-16 ***
prop_coll     5.345e-03  5.809e-05   92.02   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0004758 on 2444 degrees of freedom
  (51 observations deleted due to missingness)
Multiple R-squared:  0.8222,    Adjusted R-squared:  0.822
F-statistic:  3767 on 3 and 2444 DF,  p-value: < 2.2e-16
```

The regression analysis indicated that all predictors have a significant impact on birth rates, with (p)-values less than (2.2e-16). The coefficients suggest that increases in the proportion of mothers without a high school diploma, Medicaid-covered births, and college education are associated with increased birth rates.

5. Extracting Estimates

The coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained using:

```
The cefficient estimates :
```

```
(Intercept)     prop_nohs prop_medicaid     prop_coll
0.002147788   0.001624656   0.001107720   0.005345208
```

- Intercept: This is the estimated value of the birth rate when all the explanatory variables are zero. A positive value suggests a slight increase in birth rates as this proportion increases.
- prop_nohs: This coefficient indicates the estimated change in birth rates for a one-unit increase in the proportion of Medicaid-covered births, with all other factors held constant. A positive value means a slight increase in birth rates as this proportion increases.
- prop_medicaid: This coefficient indicates the estimated change in birth rates for a one-unit increase in the proportion of Medicaid-covered births, with all other factors held constant. A positive value implies a slight increase in birth rates as Medicaid coverage increases.
- prop_coll: This coefficient says that the estimated change in birth rates for a one-unit increase in the proportion of mothers with a college education, holding other variables constant. The positive value means a more significant increase in birth rates associated with higher educational attainment.

6. Error Variance Estimate

The error variance estimate $\hat{\sigma}^2$ can be retrieved as:

```
Error Variance Estimate :

[1] 2.264064e-07
```

This statistic measures the variance of the residuals, indicating how much the observed birth rates deviate from those predicted by the model. A smaller value suggests that the model's predictions closely fit the actual data, while a larger value would indicate more variability.

7. Variance-Covariance Matrix

The variance-covariance matrix of the estimated coefficients is given by $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ which can be retrieved in R using:

```
                (Intercept)     prop_nohs prop_medicaid     prop_coll
(Intercept)    5.145761e-09 -1.086138e-09 -9.313719e-10 -3.473114e-09
prop_nohs     -1.086138e-09  2.372259e-09 -7.339025e-10 -2.433991e-10
prop_medicaid -9.313719e-10 -7.339025e-10  2.093994e-09  1.775701e-10
prop_coll     -3.473114e-09 -2.433991e-10  1.775701e-10  3.374033e-09
```

8. Model Interpretation R-Squared:

A standard metric often reported with linear models is the R^2 score, which quantifies the proportion of variation in the response explained by the model:

```
R-Squared :
```

```
[1] 0.822175
```

9. Fitted values and Residuals

The fitted value for y_i is the value along the line specified by the model that corresponds to the matching explanatory variable x_i. In other words:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_p x_{ip}$$

This formula includes the intercept $\hat{\beta}_0$ and the contributions from each predictor variable x_{ij}.

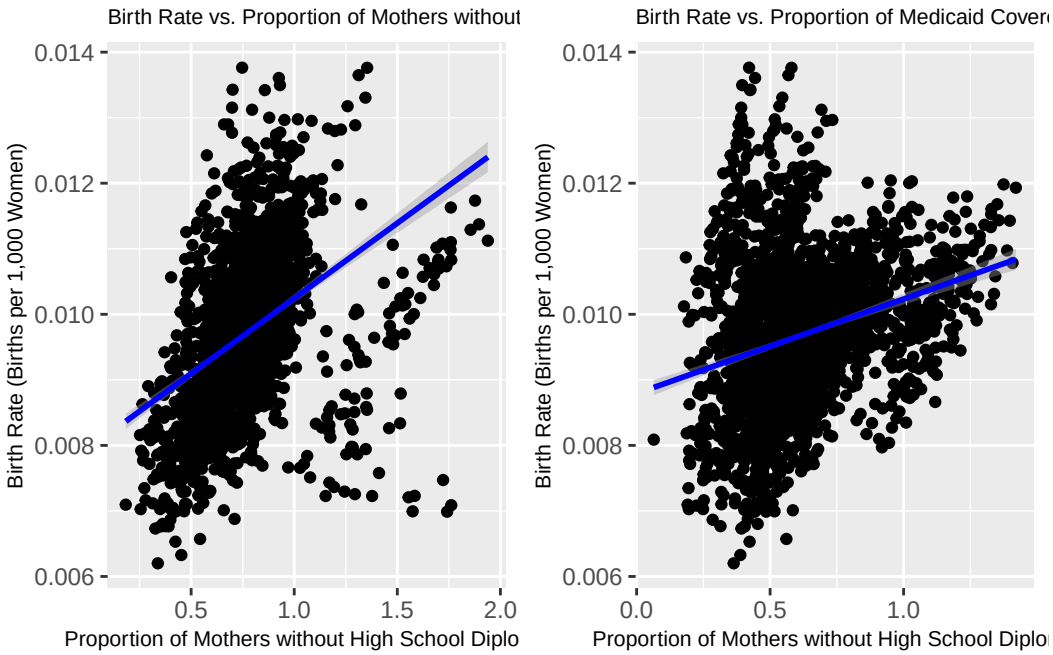
Residuals, denoted as e_i, are the differences between the observed values of the response variable and their corresponding fitted values. Residuals indicate how much the model's predictions deviate from the actual data.

The residual for the i-th observation is calculated as: $e_i = y_i - \hat{y}_i$ Here, y_i is the actual observed value, and $\hat{y}_i$ is the predicted value by the model.

Fitted Values: These provide an estimate of the response variable based on the model. They are crucial for understanding the model's predictions and can be used to assess the overall fit of the model.

Residuals: They measure the accuracy of predictions. Small residuals indicate that the model's predictions are close to the actual values, suggesting a good fit. Analyzing residuals helps in diagnosing model performance and checking assumptions, such as homoscedasticity and normality.

10. Visualizing the Model



Scatter Plot Analysis Description The scatter plots visually represent the relationship between birth rates and two key variables: the proportion of births covered by Medicaid and the proportion of mothers without a high school diploma. Each point in the scatter plots corresponds to an observation in the dataset, such as a state or region.

Observed Relationships:

Birth Rate vs. Proportion of Medicaid Covered Births:

Positive Trend: The scatter plot indicates a positive correlation between the proportion of Medicaid-covered births and the birth rate. As the proportion of Medicaid-covered births increases, the birth rate also tends to increase.

Trend Line: The blue trend line in the plot confirms this positive relationship, suggesting that higher proportions of Medicaid-covered births are associated with higher birth rates.

Density and Variability: There is a dense clustering of points around the trend line, although the points are spread out horizontally. This indicates a consistent pattern with some variability across the data.

Birth Rate vs. Proportion of Mothers without High School Diploma:

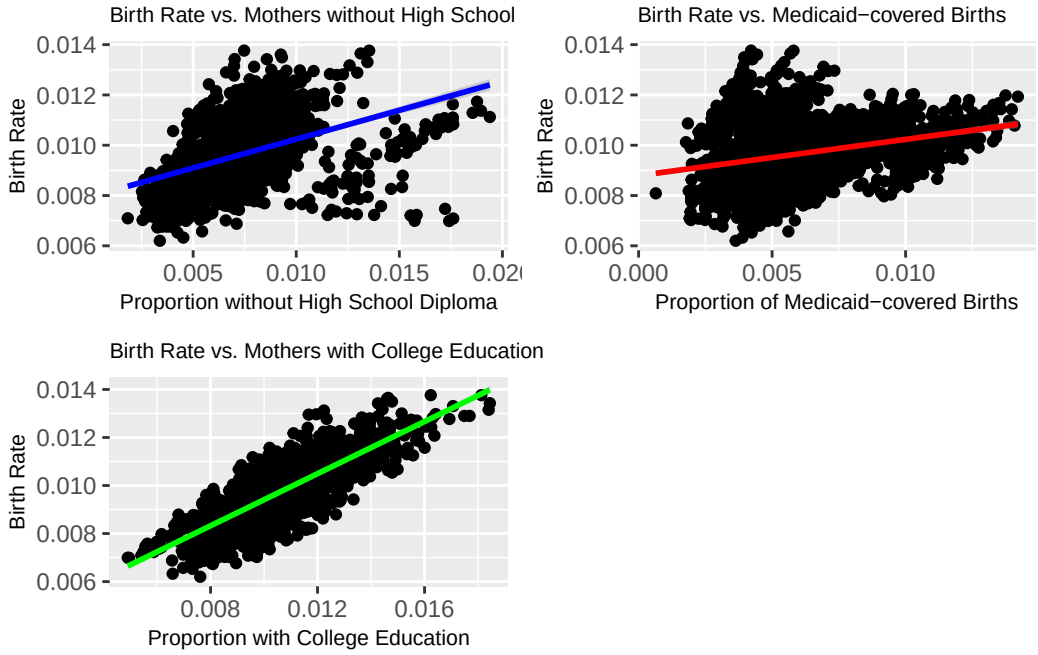
Positive Trend: Similar to the Medicaid plot, there is a positive correlation between the proportion of mothers without a high school diploma and birth rates. As the proportion increases,

the birth rate tends to rise. Trend Line: The blue trend line illustrates this relationship, indicating that a higher proportion of mothers without a high school diploma is associated with higher birth rates.

Density and Spread: There is a noticeable clustering of points around the trend line, suggesting a strong pattern across the data. However, the variability is more pronounced, as indicated by the wider spread of points.

The calculated coverage probability (approximately 94.77%) implies that the prediction intervals provided by the model are effectively capturing the true variability of the birth rates across observations. This high coverage rate indicates that the intervals are appropriately wide, considering both model uncertainty and inherent data variability, thereby providing reliable predictions for new observations.

11. Multiple Linear Regression



12. Correlation Analysis of Birth Rates with Socio-Economic Factors

```
Correlation with prop_nohs: 0.4268599
```

```
Correlation with prop_medicaid: 0.2849148
```

```
Correlation with prop_coll: 0.8000969
```

The correlation analysis reveals the strength and direction of the linear relationship between birth rates and the studied socio-economic factors:

Proportion Without High School Diploma (prop_nohs): The correlation coefficient of 0.4269 indicates a moderate positive relationship. This suggests that as the proportion of mothers without a high school diploma increases, the birth rate tends to increase moderately.

Proportion of Medicaid-Covered Births (prop_medicaid): With a correlation coefficient of 0.2849, there is a weak positive relationship between Medicaid coverage and birth rates. This suggests a slight tendency for birth rates to increase with higher Medicaid coverage, though the relationship is not strong.

Proportion with College Education (prop_coll): The high correlation coefficient of 0.8001 indicates a strong positive relationship. This suggests that higher proportions of mothers with a college education are strongly associated with higher birth rates.

These correlations provide insights into how different educational and health coverage factors relate to birth rates, highlighting the varying degrees of influence each factor has. The strong correlation with college education, in particular, points to significant socio-economic implications.

13. Linear Regression Analysis of Birth Rate with Proportion Without High School Diploma

```
[1] "Birth Rate Based on Proportion Without a High School Diploma"
```

```
Coefficients:
 0.007946722 0.002296114
```

```
Standard Errors:
 7.734393e-05 9.835601e-05
```

```
Variance of Residuals:
 1.040358e-06
```

```
[1] "Birth Rate Based on Proportion of Medicaid-Covered"
```

```
Coefficients:
 0.008791115 0.001442021


Standard Errors:
 6.474169e-05 9.809447e-05


Variance of Residuals:
 1.168888e-06



[1] "The Influence of Mothers with College Education on Birth Rates"


Coefficients:
 0.003984824 0.005423688


Standard Errors:
 8.751738e-05 8.222078e-05


Variance of Residuals:
 4.577792e-07
```

1. Birth Rate Based on Proportion Without a High School Diploma The analysis reveals a coefficient of 0.002296 for the proportion of mothers without a high school diploma, indicating a positive relationship between this factor and birth rates. The standard error is relatively small at 9.835601e-05, suggesting precise estimates. The variance of residuals is 1.040358e-06, pointing to a reasonable fit of the model.
2. Birth Rate Based on Proportion of Medicaid-Covered A coefficient of 0.001442 for Medicaid coverage suggests a positive yet weaker relationship with birth rates compared to educational factors. The standard error is 9.809447e-05, indicating reliable estimates, while the variance of residuals is 1.168888e-06, showing a fair model fit.
3. The Influence of Mothers with College Education on Birth Rates The coefficient of 0.005423 for mothers with college education is notably higher, reflecting a strong positive association with birth rates. The standard error is 8.222078e-05, denoting precise estimates. The variance of residuals, at 4.577792e-07, suggests a good model fit, highlighting the significant impact of higher education levels on birth rates.

Conclusion: Overall, these analyses demonstrate varying degrees of influence that educational attainment and Medicaid coverage have on birth rates. The positive association with college education is particularly strong, suggesting that higher education levels may lead to higher birth rates, perhaps due to increased socio-economic stability.

14. Model Fit and R^2 Statistic

The following R2 values represent the proportion of variance in birth rates explained by each model, highlighting the influence of educational attainment and Medicaid coverage among mothers:

```
Model 1: Impact of Low Educational Attainment on Birth Rates:  0.1822093

Model 2: Influence of Medicaid Coverage on Birth Rates :  0.08117643

Model 3: Effect of Higher Education on Birth Rates :  0.6401551
```

Model 1: Impact of Low Educational Attainment on Birth Rates

(R^2)=0.1822: This indicates that approximately 18.2% of the variance in birth rates is explained by the proportion of mothers without a high school diploma. This suggests a moderate impact of low educational attainment on birth rates.

Model 2: Influence of Medicaid Coverage on Birth Rates

(R^2)=0.0812: About 8.1% of the variance in birth rates is explained by Medicaid coverage. This relatively low (R^2) suggests a weaker influence of Medicaid coverage on birth rates compared to educational factors.

Model 3: Effect of Higher Education on Birth Rates

(R^2)=0.6402: This model explains 64.0% of the variance in birth rates, indicating a strong positive relationship between higher education levels among mothers and birth rates.

The multiple linear regression model, by considering all these factors, captures more variance than any individual model, providing a comprehensive understanding of how these socioeconomic factors jointly influence birth rates.

15. Discussion

In this project, I studied how educational attainment and Medicaid coverage among mothers influence birth rates across different states in the U.S. The main goal was to understand if these socioeconomic factors have significant impacts on birth rates and how strong these impacts might be.

The strongest finding was that mothers with a college education had a clear positive relationship with birth rates (r=0.8001). This factor alone explained about 64% of the differences in

birth rates. This suggests that mothers with college education tend to have higher birth rates, possibly due to better economic stability and resources.

Mothers without a high school diploma also showed a positive but weaker relationship with birth rates (r=0.4269), explaining about 18.2% of the variation. Medicaid coverage had the weakest relationship of the three factors (r=0.2849), explaining only about 8.1%.

When considering all three factors together in a multiple regression model, I explained around 82.2% of the variation in birth rates, showing that these factors together significantly influence birth rates.

The study has some limitations. The data used was aggregated at the state level, which may hide important variations within states, such as differences between urban and rural areas or between different socioeconomic groups. Another limitation is the absence of other potentially influential factors such as income levels, race or ethnicity, healthcare accessibility, or policy differences.

Future research could use more detailed data, like county or city-level statistics, and include additional factors. This would allow for a more complete understanding of what drives birth rates, enabling better-informed policy decisions and interventions.

Overall, educational attainment, especially college education, has a significant positive impact on birth rates. Medicaid coverage and lower educational attainment also influence birth rates but to a lesser extent.