Social and Economic Effects on Infant Mortality

Dylan S. Eggemeyer

Northwest Missouri State University, Maryville MO 64468, USA
S543038@nwmissouri.edu

Abstract. A low infant mortality rate is a sign of a healthy nation, not only physically but both socially and economically. This project looks to find the factors that contribute to infant mortality rate to discover where improvements can be made that would cause the most impact in its reduction. The data set analyzed was sourced from various government agencies, including the CDC and World Health Organization. Since this data was from disparate sources, the data was collated first before being cleaned and loaded into a final data set. The effects of Poverty Rate, Median Income, Uninsured Rate, Graduation Rate, and Violent Crime Rate on Infant mortality were analyzed using Linear Regression, Multiple Linear Regression, and Polynomial Regression. R squared and mean absolute error were used to determine the best models for each feature and all features together. While R-squared scores were low, it was observed that Poverty Rate and Median Income were factors in infant mortality rate, and the combined features could be used to predict the rate as well.

Keywords: infant mortality · birth rates · social impact · economic impact

1 Introduction

In healthcare, the infant mortality rate is defined as the number of deaths per 1,000 live births. This is an indication of the overall health of the country. [4] This project seeks to understand the social and economic effects of infant mortality rate, which resides in the healthcare domain. Data sources for this project will be sourced from cdc.gov, the world health organization, and other government sources. The first step of the project was to create a GitHub repository and Overleaf account to store code and data as well as the report and outline. The outline was created and data sources to aid in the analysis were gathered. Next, was to clean the data and store in a CSV format to later apply machine learning techniques (most likely regression analysis) learned in previous courses to the cleaned dataset, analyze and visualize the findings, and draw conclusions to include in the final report. The key components of this research will be the data sources that are found and the features in those data sources, the analysis that is done on those data sources, as well as the visualizations and presentation of the conclusions drawn from the analysis. The analysis will include regression analysis of multiple years of data per state to determine the factors that drive infant mortality rate.

1.1 Goals of this Research

This project seeks to identify social and economic factors that contribute to the infant mortality rate by state, both positively and negatively.

2 Dataset

This project began with many disparate data sources. The main data source used was infant mortality rate by state from the CDC[5], downloaded in CSV format. This was the base for the project, as finding the causes of infant mortality rate is the basis of this study. From the basis of infant mortality rate by state and year, a selection of social determinants of health was researched and pulled into the data set in CSV form. Social determinants of health include Economic Stability, Education Access and Quality, Health Care Access and Quality, Neighborhood and Built Environment, and Social and Community Context[7].

2.1 Data Selection

The first step in data selection was finding which social determinants of health were available for consumption. The availability of reliable data lead to the choice of gathering poverty rate[1] and median annual income[8] for Economic Stability, graduation rates[6] for Education Access and Quality, uninsured percent of the population[3] for Health Care Access and Quality, and violent crime rate[2] for Neighborhood and Build Environment and Social and Community context. These selected attributes will help to solve which factors contribute to the infant mortality rate across the United States.

2.2 Data Cleaning

Data cleaning began with the manual combination of all sources of data into a single dataset. Although all sources were in CSV format, the data structures were vastly different and needed to be updated to the level of state and year. After manually combining all data sources, the data source contained ten attributes and 400 records. The data set was then loaded into a panda data frame in Python for further cleaning and analysis.

With the dependent variable of infant mortality rate assumed to be affected by the independent variable of poverty rate, median annual income, graduation rate, uninsured percent, and violent crime rate, the dependent variable and the independent variables were analyzed for any missing data. Using the isnull() function in pandas it was discovered that all independent variables contained null values. With the dropna() function, all rows with missing values were removed. Since the data only spanned a few years and it didn't make sense to average an attribute across states, it was decided to remove these rows. After the function was used, the data set went from 400 records to 200 records.

Next, each attribute was checked to ensure that the data types matched the expected data types. It was found that the infant mortality rate data type was

object instead of float64. It was found that text was in one record and this record was removed as it would not have a dependent variable when it came time to analyze the data. Lastly, the now 199 records were visualized into a histogram for any outliers. No outliers were found in the data and the data set was finalized.

2.3 Data Set

The final data set contains 199 records and 10 attributes. The 10 attributes are Year, Infant Mortality Rate, Infant Deaths, Poverty Rate, Median Income, Uninsured Rate, Graduation Rate, and Violent Crime Rate. These 199 records describe the infant mortality rate and assumed causes by year and state. The data set spans all 50 states for 2016 through 2019, with one record for Idaho removed due to a missing infant mortality rate.

The dependent variables that are assumed to affect the infant mortality rate are Poverty Rate (the percent of people below the federal poverty line), Median Income (median household income), Uninsured Rate (percent of the population without any form of insurance), Graduation Rate (percent of students who enter ninth grade as a cohort and receive their diploma in four years), and Violent Crime Rate (number of violent crimes committed per 100,000 people).

3 Data Analysis

3.1 Exploratory Data Analysis

During exploratory data analysis, the final data set was analyzed in 3 ways. Univariate Analysis was conducted to analyze each variable for any outliers, Bivariate analysis was conducted to show the correlation between each independent variable against the dependent variable, and Multivariate analysis was conducted to analyze each variable's correlation to every other variable. In order to conduct this analysis, the final data set was loaded into a Jupyter Notebook using pandas. The code below was used to import all necessary packages and load the CSV file into a pandas data frame.[9]

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
df = pd.read_csv('final_dataset.csv')
```

To conduct a univariate analysis, each variable was put in a box plot to check for large outliers. The code below produced a box plot for each individual variable[9]. An example using graduation rate is given. There were no major outliers, so the decision was made to continue the analysis without any further data cleaning. Code can be seen below and an example in Figure 1.

```
df.boxplot(column='GRADUATION_RATE')
```

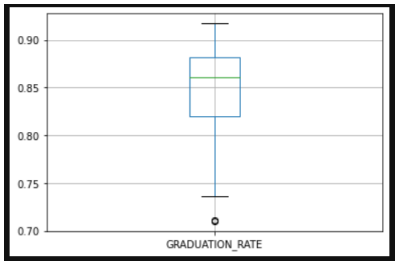

Fig. 1. Box Plot of Graduation Rate

In bivariate analysis, each independent variable was plotted against the independent variable, utilizing the seaborn package. After analyzing the scatter plots, Poverty Rate, Violent Crime Rate, and Uninsured Rate showed a positive correlation to Infant Mortality Rate, while Median income showed a negative correlation. Graduation Rate did not reveal any strong correlation. An example of the code is below and the output can be seen in Figure 2.

```
sns.relplot(x='INFANT_MORTALITY_RATE', y='POVERTY_RATE', data=df)
```

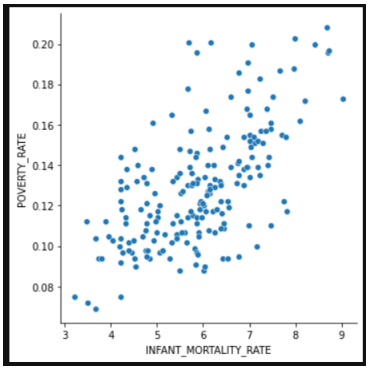

Fig. 2. Poverty Rate plotted against Infant Mortality Rate

In multivariate analysis, each variable was compared to every other variable to analyze which variables correlated with one another. This analysis confirmed

the results of the bivariate analysis. In addition, it confirmed proper data collection by showing a strong negative correlation between median income and poverty rate. Below shows the code, and the output can be seen in Figure 3.

```
corr = df.corr()
sns.heatmap(corr, annot=True, square=True)
plt.yticks(rotation=0)
plt.show()
```

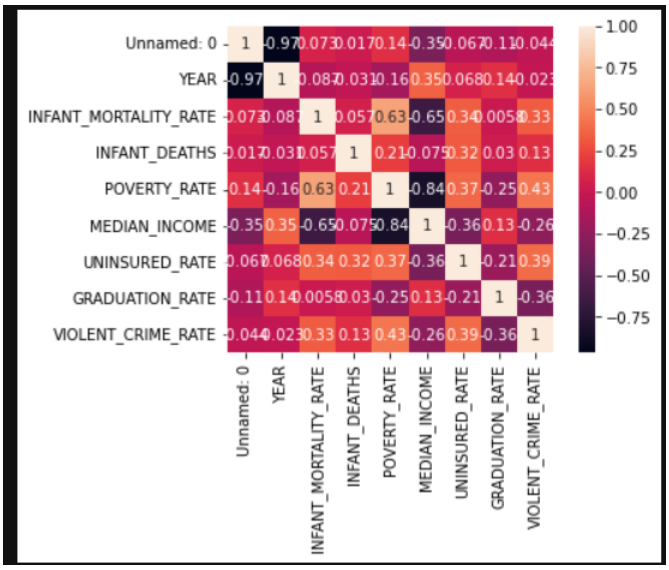

Fig. 3. Multivariate Analysis of Final Data Set

To see all code and output, view the Jupyter Notebook on GitHub

3.2 Predictive Data Analysis

Because all dependent variables are numeric values, this research conducted Regression Analysis to evaluate the factors that contribute to infant mortality rate. Linear Regression, Multiple Linear Regression, and Polynomial Linear Regression were used to train multiple models and assess their accuracy. For each model, the final data set was split into two training and testing data sets. 80 percent of the data was used to train and 20 percent was used to test.

```
X_train, X_test, y_train, y_test
= train_test_split(X, y, test_size=0.2, random_state=42)
```

For Linear Regression, each independent variable was plotted against the dependent variable, and a best-fit line was found. In the below example, poverty rate was plotted against infant mortality rate and a model was created to predict infant mortality rate based on the poverty rate.

```
X = df[['POVERTY_RATE']]
y = df['INFANT_MORTALITY_RATE']

# Split the Data
X_train, X_test, y_train, y_test
= train_test_split(X, y, test_size=0.2, random_state=42)

#fit the model
lm = LinearRegression()

lm.fit(X_train, y_train)

#predict training data
y_pred = lm.predict(X)
```

In multiple linear regression, all independent variables were used to predict infant mortality rate. Next, all but graduation rate was used, and finally, both graduation rate and uninsured rate were removed. Using a similar code to linear regression, multiple linear regression was tested by defining many independent variables in the data set.

```
# Multiple Linear Regression for All Independent Variables
X = df[['POVERTY_RATE', 'MEDIAN_INCOME', 'UNINSURED_RATE',
       'GRADUATION_RATE', 'VIOLENT_CRIME_RATE']]
y = df['INFANT_MORTALITY_RATE']

# Split the Data
X_train, X_test, y_train, y_test
= train_test_split(X, y, test_size=0.2, random_state=42)

#fit the model
lm = LinearRegression()

lm.fit(X_train, y_train)
```

```
#predict training data
y_pred = lm.predict(X)
```

In polynomial regression, both polynomials 3 and 4 were used to create a polynomial linear regression model. In both instances, all independent variables were used.

```
# Polynomial Linear Regression (poly = 4)
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import PolynomialFeatures
from sklearn.preprocessing import StandardScaler

X = df[['POVERTY_RATE', 'MEDIAN_INCOME', 'UNINSURED_RATE',
       'GRADUATION_RATE', 'VIOLENT_CRIME_RATE']]
y = df['INFANT_MORTALITY_RATE']

# Split the Data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Run the Model
imp_mean = SimpleImputer(missing_values=np.nan, strategy='mean')
poly4 = PolynomialFeatures(degree=4, include_bias=False)
scale = StandardScaler()
lr_model = LinearRegression()

stages = [('imp_mean', imp_mean),
          ('poly4', poly4),
          ('scale', scale),
          ('lr_model', lr_model),
         ]
pipe_model = Pipeline(stages)

pipe_model.fit(X_train, y_train)

y_pred = pipe_model.predict(X_train)
```

After each model was run for the training data, it was tested with the test data. Each model was then scored and the Mean Average Error, Root Mean Square Error, Mean Square Error, and R Squared were recorded for both training and test sets.

```
print('Results for linear regression on training data')
```

```
print('  Default settings')
print('Internal parameters:')
print('   Bias is ', lm.intercept_)
print('   Coefficients', lm.coef_)
print('   Score', lm.score(X,y))

print('MAE is  ', mean_absolute_error(y, y_pred))
print('RMSE is ', np.sqrt(mean_squared_error(y, y_pred)))
print('MSE is ', mean_squared_error(y, y_pred))
print('R^2    ', r2_score(y,y_pred))
```

Scores for both test and training models were recorded into an Excel file and are included in the GitHub link and table below. Upon analysis of the R squared of both test and training data sets, it was found that most of the variables account for only a small portion of the causes of infant mortality rate. Linear Regression showed that the model created with Poverty Rate and Median Income explains about 40 percent of the variance in infant mortality rate. Multiple Linear regression gave the best results when all independent variables were used, but did not significantly improve the model from some individual features. In polynomial regression, the training data sets produce a strong correlation, but the test data sets were not accurately predicted, indicating that the polynomial regression over-fitted the data.

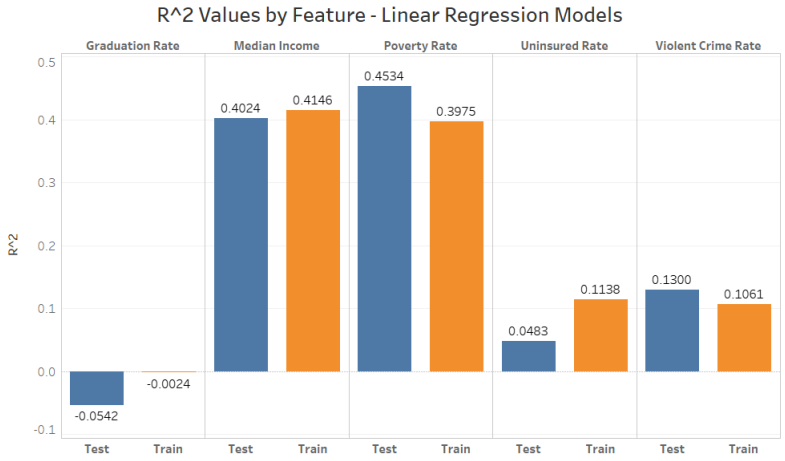

Fig. 4. Poverty Rate and Median Income explain about 40 percent of the variance in infant mortality rate.

Model	Feature(s)	Train/Test	MAE	RMSE	MSE	R^2
Linear Regression	Poverty Rate	Train	0.741997933	0.912275768	0.832247077	0.397494192
Linear Regression	Poverty Rate	Test	0.639632978	0.784472003	0.615396324	0.453394977
Linear Regression	Median Income	Train	0.736462831	0.899212746	0.808583562	0.414625409
Linear Regression	Median Income	Test	0.658123052	0.820277199	0.672854683	0.402359528
Linear Regression	Uninsured Rate	Train	0.928944676	1.106422407	1.224170544	0.113760945
Linear Regression	Uninsured Rate	Test	0.857932486	1.035111588	1.071455999	0.048315357
Linear Regression	Graduation Rate	Train	0.937731716	1.176697654	1.384617368	-0.002394637
Linear Regression	Graduation Rate	Test	0.810130025	1.089416847	1.186829066	-0.054160877
Linear Regression	Violent Crime Rate	Train	0.897770064	1.111219241	1.234808202	0.106059806
Linear Regression	Violent Crime Rate	Test	0.756055395	0.989701541	0.97950914	0.129984053
Mulitple Linear Regression	Poverty Rate, Median I	Train	0.679363216	0.841217672	0.707647172	0.487698374
Mulitple Linear Regression	Poverty Rate, Median I	Test	0.60358425	0.77449515	0.599842737	0.467209926
Mulitple Linear Regression	Poverty Rate, Median I	Train	0.705283101	0.866987107	0.751666644	0.455830448
Mulitple Linear Regression	Poverty Rate, Median I	Test	0.617669416	0.78019654	0.60870664	0.459336863
Mulitple Linear Regression	Poverty Rate, Median I	Train	0.701833418	0.869572653	0.756156599	0.452579943
Mulitple Linear Regression	Poverty Rate, Median I	Test	0.599722232	0.767272867	0.588707652	0.47710029
Pipelined Linear Regression (poly = 4)	Poverty Rate, Median I	Train	0.201440332	0.301186127	0.090713083	0.936839336
Pipelined Linear Regression (poly = 4)	Poverty Rate, Median I	Test	2.438265214	5.24725452	27.53368	-23.45586234
Pipelined Linear Regression (poly = 3)	Poverty Rate, Median I	Train	0.456440357	0.585002097	0.342227454	0.761717797
Pipelined Linear Regression (poly = 3)	Poverty Rate, Median I	Test	0.63739309	0.878663194	0.772049009	0.314253515

Fig. 5. Mean absolute error, root mean squared error, mean squared error, and R squared were recorded for each test and training model.

To see all code and output, view the Jupyter Notebook on GitHub
To see all scores for each model, view the Excel file on GitHub

4 Conclusions

Poverty Rate, Median Income, and Violent Crime Rate can explain a portion of the infant mortality rate. It stands to reason those with higher incomes would be able to afford better health care and have more access to prenatal care than those in poverty. It also shows that there are many more factors that can add to the infant mortality rate that can be explored. The analysis showed that using multiple linear regression with all features did not significantly increase the predictive power of the model. The recommendation from this analysis would be to focus efforts on increasing median income and reducing the poverty rate in order to decrease the infant mortality rate from state to state. States could fund programs for low-income mothers that help them to see doctors and receive postpartum care for both the mother and baby.

5 Limitations

With the data sources being disparate across many sites and sources, this project was limited by the amount of time and effort that could be used to combine multiple sources of data. This led to a relatively small amount of data that could be gathered and required the data to be grouped at the state level.

6 Future Work

In the future, a more granular analysis could provide a more accurate picture of the conditions that cause infant mortality rate. States have a wide range of

populations that may be concentrated in different areas and analysis at the zip code or county level could lead to a more pointed effort in the reduction of infant mortality rate. With more time, more features could also be added to the analysis, including birth weight, percent of the population that smokes, maternal demographics, and distance to doctors or hospitals.

References

1. Demographics and the economy, bluehttps://www.kff.org/state-category/demographics-and-the-economy/
2. Federal bureau of investigation crime data explorer, bluehttps://cde.ucr.cjis.gov/LATEST/webapp/#
3. Health insurance historical tables - hic acs (2008-2021), bluehttps://www.census.gov/data/tables/time-series/demo/health-insurance/historical-series/hic.html
4. Infant mortality rate (between birth and 11 months per 1000 live births), bluehttps://www.who.int/data/gho/indicator-metadata-registry/imr-details/1
5. Infant mortality rates by state, bluehttps://www.cdc.gov/nchs/pressroom/sosmap/infant_mortality_rates/infant_mortality.htm
6. Public high school 4-year adjusted cohort graduation rate (acgr), by selected student characteristics and state: 2010-11 through 2018-19, bluehttps://nces.ed.gov/programs/digest/d20/tables/dt20_219.46.asp
7. Social determinants of health, bluehttps://health.gov/healthypeople/priority-areas/social-determinants-health
8. Table h-8. median household income by state, bluehttps://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html
9. Eggemeyer, D.: Data analytics capstone project, bluehttps://github.com/dylanegg/data-analytics-capstone