Exploring Different Model Types for Underweight Births Since 1969

Aiden Kollar, Liam Ralph, Nicolas Ruiz, Shruthi Senthilarasu

The University of Texas at San Antonio

Rowdy Datathon 2022

Abstract

This executive data report provides an analysis of datasets related to underweight newborns per county in Texas from 1969 to 2020. This report also provides an analysis of newborn fatalities per county within the same time period. The primary objective of the analysis is to provide a clear logistic regression model to analyze birth weight compared to different racial groups, provided that socioeconomic conditions remain constant within a selected time interval. Reporting and analysis of the dataset were done using pgAdmin, PostgreSQL, and Python. Python libraries used in this analysis include PANDAS, CSV, Daytime, matplotlib, NumPy, and Sklearn.

Keywords: PostgreSQL, Python, pgAdmin, Google Colab

Logistical Model for Underweight Births Since 1969

To compete in Rowdy Datathon 2022, we decided to attempt the intermediate challenge. The intermediate challenge was to produce a time series regression to project the number of stillbirths and underweight babies to the year 2030.

Background

Fetal death is recognized as death that occurs before, after, or during labor and delivery. As per current regulations, the Texas Department of Health and Safety requires a recorded Certificate of Fetal Death "for every fetal death weighing 350 grams or more, or if the weight is unknown, a fetus aged 20 weeks or more as calculated from the start date of the last normal menstrual period to the date of delivery" (*Handbook on Fetal Death Registration*, 2019, p.3).

Methodology

Rationale for undergoing analysis can include performing a regression on the data and model presented that can be used to predict the annual number of underweight newborns in the future. Further mathematical modeling and analysis could assist in fuelling decisions regarding creating government-funded programs and incentives that protect the health of both the fetus and the mother.

Reporting bias is highly probable within this very large, multi-decade-spanning data set. The omission of certain data entries or intentionally selecting data entries that skew the data in a specific manner. The legal parameters for what constitutes fetal death have also progressively gotten more stringent over time, so the probability that observer bias existed that has incorrectly declared a newborn dead or recorded the wrong weight is also high. Aside from human error, it is important to note that the instruments and parameters used to identify the newborn weight and death of newborns have shifted significantly throughout the fifty-one-year time period. These

inconsistencies might also affect how the data is skewed. Racial, gender and socioeconomic bias also could have affected how the data was recorded, as this data set spans many time periods where there have been many recorded instances of conflict related to those factors.

Data

The data used in this analysis has been provided by three different sources. Firstly, the National Center for Health Statistics (NCHS), has been recording fetal births since 1969 via the Centers for Disease Control. The NCHS provides data on place of birth, level of parent education, place of residence, weight at birth, gestational time period, number of siblings, birth order, and hundreds of other variables of significance. Secondly, the Surveillance, Epidemiology, and End Results Program provided data for unabridged population estimates for five-year age brackets per county. Lastly, SEDAC, the Socioeconomic Data and Applications Center, provided context for the geographic and socioeconomic records relating to infant birth and fatality.

Data

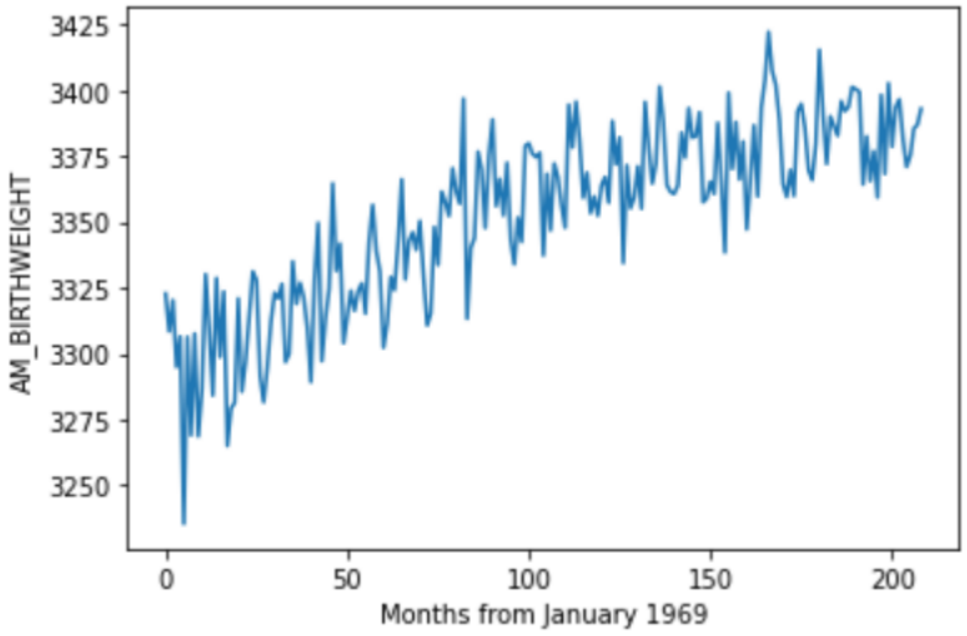

Figure 1. Graph showing the rising increase in underweight births since January 1969. This data was taken from a "sampled dataset" that included 1/100 records from the original 'Natality.csv' dataset. As a result, the trends, but not actual amounts, shown in the graph are accurate.

```
def time_series_graph(variable):
    months = []
    values = []
    for year in range(1969, 1988):
        for month in range(1, 12):
            months.append(str(month) + "/" + str(year))
    for year in range(1969, 1988):
        for month in range(1, 12):
            month_year = str(month) + '-' + str(year)
            values.append(df[df['month_year'] == month_year][variable].mean())
    plt.plot(values)
    plt.xlabel('Months from January 1969')
    plt.ylabel(variable)
    figure(figsize=(8, 6), dpi=80)
    plt.show()
```

Figure 2. Excerpt of code to produce a time series graph grouped by month with any given variable.

```
from sklearn.linear_model import LogisticRegression

x_columns = ['AM_M_AGE15', 'AM_F_AGE11', 'ID_M_EDU6',  'ID_F_EDU14',  'dv_black',
    'dv_white',  'dv_other']
X = df[x_columns]
y = df['dv_birthweight']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=16)
logreg = LogisticRegression(random_state=16)
logreg.fit(X_train, y_train)
y_pred = logreg.predict(X_test)
y_pred_proba = logreg.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)
print(auc)
```

Figure 3. *Code that provided us with a logistic regression function with an auc of approximately ~0.6.*

Importing the Data

The provided datasets will have undergone analysis through PostgreSQL, Python, and pgAdmin. First, using pgAdmin 4, one can restore the provided backup file. Within the database, clicking on Postgre, Tools, and Restore allowed for the restoration of the provided backup file. During this process, we encountered an error stating that *Schema USA* has not been created. In order to fix this issue Parallel to this process, we used Python to attempt to query smaller samples of the data points to retrieve all of the variables for analysis. We found success in splitting the data set into smaller, 100,000 entry files, for a total of 567 files. After we did that, however, we tried just adding the USA *Schema* to the database which fixed the error and we finished restoring from the backup.

Results

While the results of our research are limited due to time constraints, we were able to test a logistic regression model that predicts whether or not a baby would be underweight. If we had more time, one thing we would do is be able to properly query the database using Python to speed up the analysis process. This would allow us to create our charts more easily and increase the reliability of our project. Another thing we would have liked to do is a proper time series regression, something I think we all struggled with because we have Computer Science backgrounds and not much statistics experience. Future mathematical analysis and modeling could include providing a logistic regression using a basis function that allows a path for linear, quadratic, and cubic regression. This challenge allowed for a broader scope of understanding

which variables contribute to low newborn weight, which can provide valuable insight to incentivize government subsidized programs that can assist families and newborns facing these issues.

References

Handbook on Fetal Death Registration. (2019). Texas Department of State Health Services. https://www.dshs.texas.gov/vs/partners/docs/FetalDeathRegistrationHandbook.pdf

SQL Tutorial. (2019). W3schools.com. https://www.w3schools.com/sql/