Predicting Asthma Emergency Department Visit Rates Across California Census Tracts*

A Multiple Linear Regression Model Incorporating Air Pollution and Socioeconomic Factors

Giovanni Hsieh

December 12, 2025

Asthma is a leading cause of emergency department (ED) visits in the United States, and has been linked to various air pollution and socioeconomic factors. This study uses census tract level data from CalEnviroScreen 4.0 to develop a multiple linear regression model that predicts age-adjusted asthma ED visit rates in California. Predictor variables include fine particulate matter (PM2.5), ozone concentration, diesel particulate matter (PM), traffic density, and poverty percentage Traffic density, diesel PM, and poverty percentage were log-transformed to address skewness. The model's predictive performance was evaluated using 10-fold cross-validation. The root mean squared error (RMSE) was 26.24, roughly half of the mean asthma ED visit rate of 52.06. While multicollinearity and different variable scaling make individual slope interpretations unreliable, the model captures general trends across census tracts in California, providing a good baseline for population level asthma risk prediction.

1 Introduction

According to the CDC (National Center for Health Statistics 2019), asthma is a leading cause of emergency department (ED) visits in the United States, recording almost 1 million visits per year. Asthma can be caused by various triggers such as family history, allergies, respiratory infections, and air pollution. The Asthma and Allergy Foundation of America (Asthma and Allergy Foundation of America 2025) refers to asthma as one of the most common and costly diseases in the United States. This paper investigates different air pollution and socioeconomic indicators as predictors of asthma related ED visits in California.

*Project repository available at: https://github.com/giovannihsieh/math261A-paper2.

A population level prediction model can help identify areas at higher risk for asthma related ED visits. These predictions can guide environmental health policies and community measures in high risk areas. Although studies (Tiotiu et al. 2020) have linked asthma to air pollution indicators, relatively few have focused on developing prediction models. Existing prediction models (Hwang et al. 2023) rely on various air pollution indicators to predict asthma related hospital visits using machine learning and deep learning techniques. However, many of these models use controlled and limited population datasets and exclude socioeconomic control variables. To address this gap, my proposed multiple regression model uses statewide census-tract data and incorporates both air pollution indicators and poverty percentage as a key socioeconomic factor to accurately predict age-adjusted asthma ED visit rates in California.

To address this research question, I analyzed census tract level data from CalEnviroScreen 4.0 using a multiple linear regression model. The outcome variable is the age-adjusted rate of emergency department visits for asthma. The predictor variables include fine particle (PM2.5) concentration, ozone concentration, diesel particulate matter emissions, traffic density, and the poverty percentage. The model is used to estimate the relationship between asthma ED visits and each predictor variable while holding other variables constant. Because the primary goal of the analysis is prediction, model performance is evaluated using root mean squared error (RMSE) estimated via k-fold cross validation (k=10). To address right skewness, traffic density, diesel PM, and poverty were log-transformed prior to model fitting.

Using the multiple linear regression model, estimated coefficients suggest expected associations between asthma ED visit rates and the predictor variables. However, individual slope estimates, especially ozone, are affected by multicollinearity and different variable scaling, making them unreliable for interpretation. The 10-fold cross validated RMSE was roughly half of the mean asthma ED visit rate, suggesting limited predictive model performance. The residual plots revealed some heteroscedasticity and deviations from normality, but these issues are less of a concern for prediction. Overall, the model captures general trends in asthma ED visits across California census tracts and represents a good baseline for population level asthma risk prediction.

Section 2 introduces the data used in this analysis. Section 3 describes the model used. Section 4 discusses the analysis of the results and conclusions that can be drawn. Section 5 discusses potential strengths and weaknesses of the analysis, and some future steps.

2 Data

The `calenviroscreen` dataset (Office of Environmental Health Hazard Assessment (OEHHA) and California Environmental Protection Agency (CalEPA) 2021) contains data of census tracts in California based on potential exposures to pollutants, adverse environmental conditions, socioeconomic factors, and the prevalence of certain health conditions. The observational unit in this dataset are California census tracts. All outcome and predictor variables are measured or modeled at the census tract level. Because no individual level information is included, the analysis is at the population level.

The outcome variable is the age-adjusted rate of emergency department (ED) visits for asthma. Age adjustment is used to control for age since asthma hospitalization risk is strongly correlated with age. The rate of ED visits for asthma was calculated through a multi-step process. First, records for ED visits with a primary diagnosis of asthma were obtained from statewide OSHPD files using codes identifying asthma (ICD-9 and ICD-10). The age-adjusted rate for each ZIP code was calculated using ESRI population estimates standardized to the 2000 US population using five year age groupings (0-4, 5-9, etc.). The rates are per 10,000 residents per year. Spatial modeling was applied to stabilize estimates for ZIP codes with fewer than 12 ED visits. Lastly, census block were assigned the average rate of the ZIP code they intersected using areal apportionment. The rates were then estimated by the population weighted average of the rates of the census blocks it contains. Because some census tracts use spatially modeled values if there was not enough data, modeled values may reduce variability or introduce smoothing in the outcome.

Predictor variables include annual mean fine particle (PM2.5) concentration, daily maximum 8-hour ozone concentration, diesel particulate matter emissions, traffic density, and the percentage of residents living below two times the federal poverty level. Each predictor variable has been organized into it's own section describing how it was collected or transformed, and whether there are any limitations.

PM2.5($\mu g/m^3$): Fine particle pollution (PM2.5) was defined as the annual mean PM2.5 concentration over three years. PM2.5 data was extracted for all monitoring sites in California from CARB's air monitoring network database. This data was combined with satellite-based Aerosol Optical Depth (AOD) measurements. Census tract estimates were computed by averaging 1x1km grid cells overlapping tract boundaries. A potential limitation is that census tracts that were far from a monitoring site had to rely mainly on satellite data for estimates.

Ozone(ppm): Ozone was defined as the mean of summer months (May-October) of the daily maximum 8-hour ozone concentration (ppm), averaged over three years. Ozone data was obtained from California Air Resources Board's (CARB's) air monitoring network database like the data for PM2.5. The mean concentrations from the monitoring sites were used to model ozone concentrations across the state of California using a spatial interpolation method (ordinary kriging) and used to estimate concentrations for each census tract. The kriging model was used to estimate ozone concentrations for the center of each census tract. Census tracts that had centers more than 50km from the nearest monitor used the ozone value of the nearest air monitor instead of estimates from the model. The main limitation is that for remote areas, interpolation assumptions may affect accuracy.

Diesel PM(tons/year): Diesel particulate matter was defined as the spatial distribution of gridded diesel PM emissions from on-road and non-road sources. Diesel PM data was obtained from CARB's EMission FACtors (EMFAC) on-road emission model and the California

Emission Project Analysis Model (CEPAM) for non-road sectors. Like PM2.5, emissions were allocated to 1x1km grid cells, then to census tracts based on the proportion of populated area. The biggest limitation is that the emissions are modeled, not measured concentrations, so modeling assumptions need to be taken into consideration.

Traffic Density(vehicle-km per hour per kilometer of roadway): Traffic density was defined as the sum of traffic volumes adjusted by road segment length divided by total road length within 150 meters of each census tract. Traffic density was calculated by first placing a 150 meter buffer around each census tract, and ArcGIS was used to link traffic volume data from TrafficMetrix. Length-adjusted volume was summed across all roads in the buffer for each census tract, and normalized by total road length. A small limitation is that for more rural areas with fewer roads, less traffic data may affect reliability.

Poverty(%): Poverty was defined as the percent of the population living below two times the federal poverty level. The data was derived from the 2015-2019 American Community Survey, which contained data about the number of individuals below 200 percent of the federal poverty level. To calculate the poverty percentage, the number of individuals below 200% of the poverty level was divided by the total population for whom the poverty status was determined. Since the ACS estimate is from a sample of the population unlike the census, the standard error (SE) and relative standard error (RSE) were calculated to evaluate the reliability of each estimate. Census tracts with large relative standard errors (RSE $>$ 50) were excluded from the data. The limitation for this variable is the sampling error that comes with a survey, and that some census tracts have no poverty estimate due to unreliable data.

One important thing to note is that many of the predictor variables are highly correlated. There has been research linking fine particulate matter (PM2.5) and ozone concentration (Sun et al. 2023) as well as linking PM2.5 and diesel PM (Bosson et al. 2007). There have been studies (Heydari et al. 2020) linking CO2 emission to traffic density, suggesting that traffic density would be highly correlated with diesel PM. However, because the purpose of the model is for prediction rather than inference about individual predictors, the model can still perform well even if some predictors have high correlation. Additionally, I am using cross-validation to evaluate predictive performance. This makes the model more robust for new data despite correlated predictors.

3 Methods

I chose to perform multiple linear regression because the outcome variable is continuous and approximately normally distributed after accounting for predictors. Multiple linear regression also combines both raw and transformed predictor variables into an effective predictive model that can be supported by cross validation. The predictor variables included in the model were selected based on previous research and articles linking air pollution indicators and socioeconomic conditions to asthma related ED visits. According to a research report (Meng et al. 2011), air pollutants such as PM2.5 and ozone can trigger symptoms among asthmatics.

Another study (McConnell et al. 2010) finds that children living in areas with higher traffic density in California suffer significantly increased rates of asthma. An article (Spira-Cohen et al. 2011) found that diesel PM exacerbates asthma symptoms in children with asthma. Lastly, a paper (Wendt et al. 2014) found low income children had higher rates of asthma diagnoses. Poverty percentage was included in the model to control for socioeconomic disparities that can affect asthma rates.

I looked at the histograms for each predictor variable to assess the shape of each the distributions and decide whether any variable transformations were needed. Based on these histograms, I decided that traffic density, diesel PM, and poverty should be log transformed, since they all had right skewness. The histograms below show the distributions before and after the log transformations.

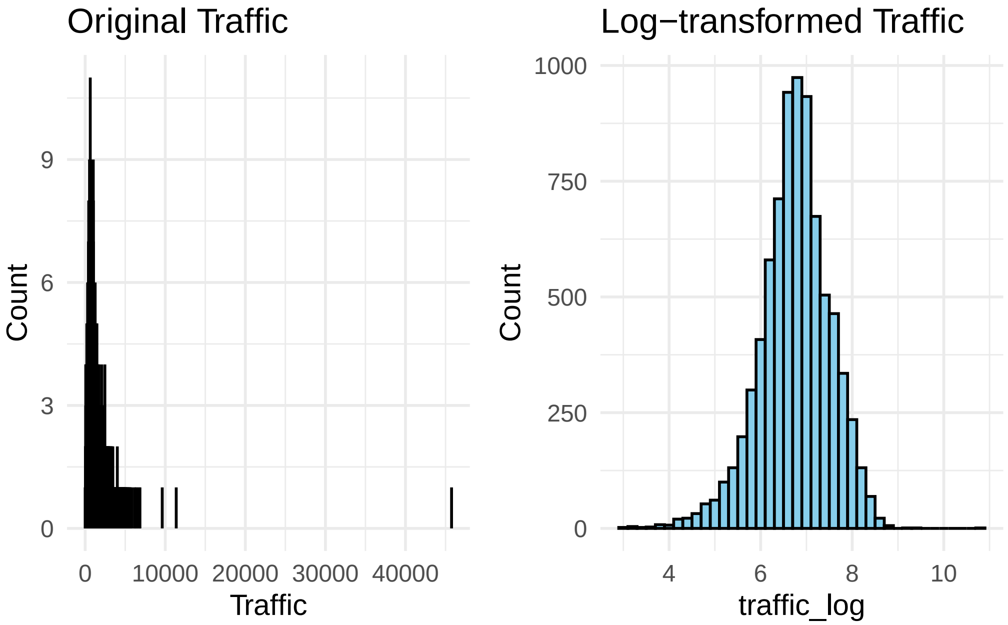

Figure 1: Original and log-transformed histogram of Traffic density.

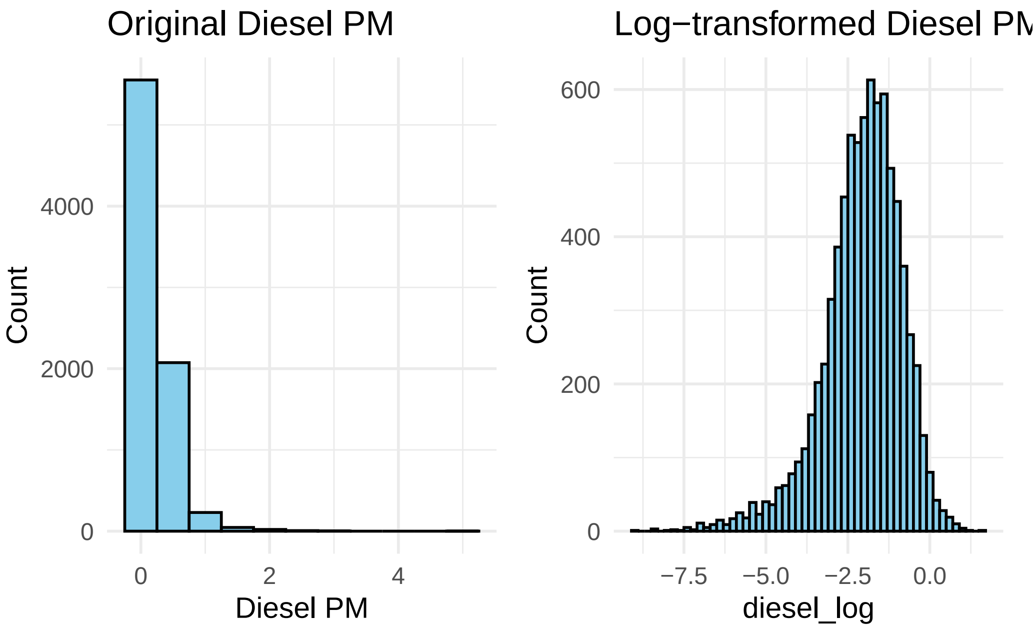

Figure 2: Original and log-transformed histogram of Diesel PM concentration.

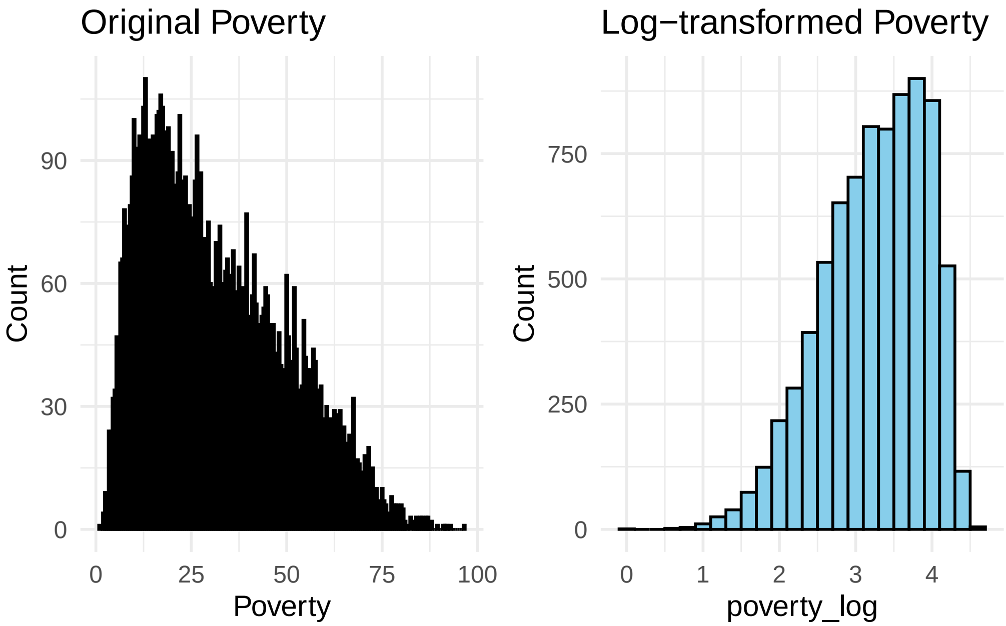

Figure 3: Original and log-transformed histogram of Poverty rate.

I fit the multiple linear regression model.

$$A = \beta_0 + \beta_1(PM_i) + \beta_2(O_i) + \beta_3 \log(D_i) + \beta_4 \log(T_i) + \beta_5 \log(P_i) + \epsilon_i$$

Variable Definitions:

A: Age-adjusted rate of emergency department (ED) visits for asthma in census tract i.

PM_i: Annual mean fine particle pollution (PM2.5) over three years in $\mu g/m^3$.

O_i: Daily maximum 8-hour ozone concentration averaged over three years in ppm.

D_i: Diesel particulate matter in tons/year. Log-transformed ($\log(D_i)$) due to right-skewed distribution.

T_i: Traffic density in vehicle-km per hour per kilometer of roadway. Log-transformed ($\log(T_i)$) due to right-skewed distribution.

P_i: Poverty percentage, or percentage of residents living below two times the federal poverty level. Log-transformed ($\log(P_i)$) due to right-skewed distribution.

ε_i: Error term for census tract i, or the variation in A_i not explained by the predictor variables in the model.

The error terms ε_i are assumed to meet the following criteria:

1. **Homoscedasticity (constant variance) of errors:** $Var[\varepsilon_i] = \sigma^2$, meaning the spread of errors is consistent across all predictor values.
2. **Independence of errors:** Every ε_i is independent from the others, meaning errors for one census tract do not affect errors for another.
3. **Normality of errors:** $\varepsilon_i \sim N(0, \sigma 2)$, normality of errors ensures t-tests and confidence intervals are valid. Mainly important for inference, not prediction.
4. **Linearity**: The expected value of A_i is a linear function of all predictors.

$$E[A_i | PM_i, O_i, \log D_i, \log T_i, \log P_i] = \beta_0 + \beta_1(PM_i) + \beta_2(O_i) + \beta_3 \log(D_i) + \beta_4 \log(T_i) + \beta_5 \log(P_i)$$

I looked at the residual vs predictor plot and a quantile-quantile plot to help validate these assumptions. For linearity, the residual plot should look evenly spread around the residual = 0 line, with no clear patterns. Normality of errors would have points on the quantile-quantile plot along the reference line. Constant error variance would have a residual vs predictor plot that is evenly spread rather than having a cone like shape.

Because the purpose of the paper is prediction rather than inference, I want to evaluate how well the model performs on new data. Even if there is multicollinearity among predictor variables, it does not harm the model's predictive performance as long as the model generalizes well. Therefore, I used k-fold cross validation (k=10), which is able to determine how stable the model's predictions are, whether there is overfitting, and how well the model generalizes for new data. Because p-values or individual coefficients are unreliable if there are highly correlated predictors, cross validation directly evaluates the model's predictive accuracy.

I used 10-fold cross validation by dividing the dataset into 10 equally sized folds. For each fold, the model is trained on the other 9 folds, and predictions are made for the last fold. Prediction error for the last fold is recorded, and averaged with all the other runs to get cross validate root mean squared error (CVRMSE).

I looked into including interaction effects for variables like traffic density and diesel PM since the effect of traffic on asthma ED visit rates makes sense to be stronger in areas with higher diesel emissions. However, there was no significant improvement in the model's predictive performance. Therefore, I decided to only use main effects to prevent increasing model complexity and making the model more prone to overfitting.

I implemented this analysis using the R programming language (R Core Team 2025) using the lm_fit function. The plots were done using the ggplot function.

Some possible extensions would be to include more predictor variables, especially socioeconomic variables such as education level, race/ethnicity, or access to healthcare. Using nonlinear models like machine learning and deep learning models might also lead to higher predictive accuracy, however this would be difficult to interpret.

4 Results

The estimated slope parameter is b_1 = 0.43. In other words, for each one $\mu g/m^3$ increase in annual mean PM2.5, the expected change in the asthma ED visit rate is 0.43, holding all other predictors constant.

The estimated slope parameter is b_2 = -182.74. In other words, for each one ppm increase in daily maximum 8-hour ozone concentration, the expected change in the asthma ED visit rate is -182.74, holding all other predictors constant.

The estimated slope parameter is b_3 = -4.103. In other words, for each one unit increase in log-transformed traffic density, the expected change in the asthma ED visit rate is -4.103, holding all other predictors constant.

The estimated slope parameter is b_4 = 2.719. In other words, for each one unit increase in log-transformed diesel PM, the expected change in the asthma ED visit rate is 2.719, holding all other predictors constant.

The estimated slope parameter is b_5 = 21.7. In other words, for each one unit increase in log-transformed poverty percentage, the expected change in the asthma ED visit rate is 21.7, holding all other predictors constant.

The estimated intercept is $b_0 = 19.464$. This represents the predicted asthma ED visit rate when all predictors are zero. This is not realistic (zero PM2.5, zero traffic, etc.), so it is mainly useful for constructing the regression equation.

Keep in mind that because many of the predictors are highly correlated, these individual coefficient estimates are affected by multicollinearity. This means that interpretations of single variables are less reliable, but the model can still provide accurate predictions for asthma ED visits. Additionally, the scale of predictor variables vary a lot. For example, ozone is measured in ppm, which is much smaller than PM2.5 which is measured in $\mu g/m^3$, further suggesting the interpretation of individual coefficients is unreliable.

The average cross-validated RMSE was 26.24. This represents an estimate of how far, on average, the model's predictions deviate from the observed asthma ED visit rates for new census tracts. Since the mean asthma ED visits was 52.06, the RMSE is relatively high, about half of the average asthma ED visit rate. This indicates that while the model captures some of the overall trends, its predictive accuracy for an individual census tract is limited.

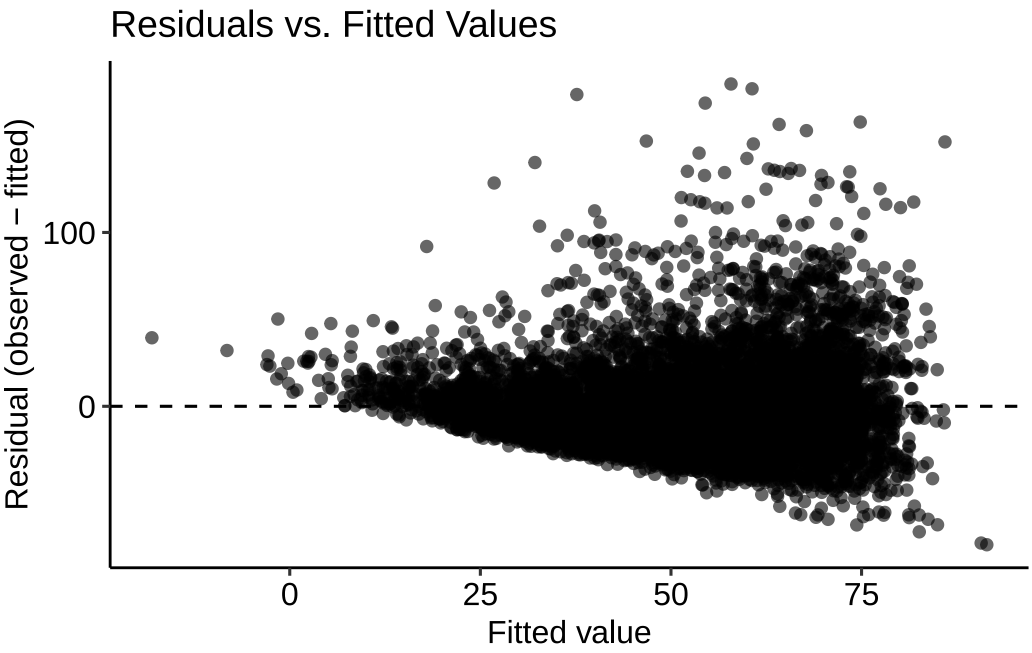

Figure 4: Residuals versus fitted values for the multiple linear regression model predicting asthma ED visit rates.

Figure 4 shows the residuals vs. fitted value plot. The errors seem to be spread out relatively evenly, validating the linearity assumption. However, there is a noticeable cone shape, indicating heteroscedasticity, or that error variance is not constant. Although heteroscedasticity can affect inference, it is less of a concern for prediction. This is because the predictions remain unbiased, and only the predicted variance is overestimated or underestimated for certain fitted values.

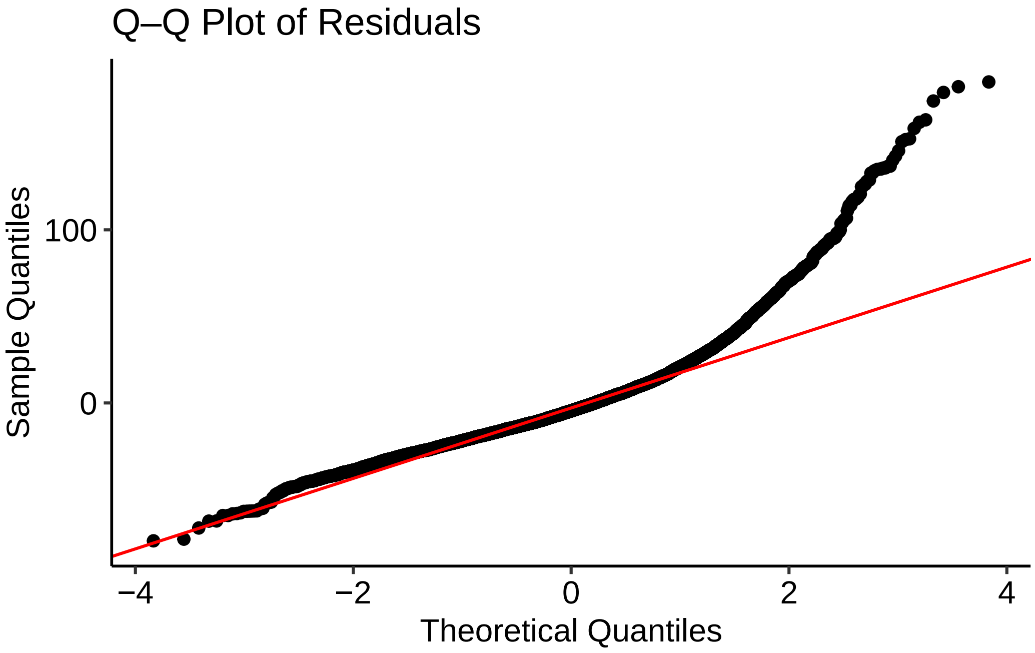

Figure 5: Q–Q plot of residuals from the multiple linear regression model predicting asthma ED visit rates.

Figure 5 shows the quantile-quantile plot. There is some deviation from the reference line shown in red on the right side, indicating unusually high residuals. This suggests the model may underpredict asthma ED visits for those census tracts and that the normality of errors assumption is violated. However, because the goal for the model is prediction, violating the normality of errors assumption is less of an issue, especially since the cross-validated RMSE gives an estimate of predictive performance.

5 Discussion

This study developed a multiple linear regression model to predict age-adjusted asthma emergency department visit rates across California census tracts using air pollution indicators (PM2.5, ozone, diesel PM, traffic density) and socioeconomic factors (poverty percentage). The goal was to predict asthma ED visit rates using publicly available census tract data. Key findings highlight the presence of multicollinearity and differences in variable scaling, making interpretation of individual predictor variables unreliable. The model had moderate predictive performance, with an RMSE roughly half of the mean asthma ED visit rate. Residual analysis suggested some heteroscedasticity and non-normal errors, but these issues are less concerning for this model's goal of prediction.

There are several limitations of the model. First, the analysis relied on census-tract level data, so individual-level risk factors were not considered. Second, there may be key environmental or socioeconomic variables missing from the dataset used, which could affect predictive accuracy. Lastly, some of the variables have uncertainty. PM2.5 and ozone concentrations were partially estimated using satellite and spatial interpolation methods which may reduce accuracy. Diesel PM and traffic density are based on modeled emissions and traffic volume estimates, which might have measurement errors. Poverty percentage was also calculated from the ACS survey, leading to potential sampling error.

Some future improvements could include additional predictors, specifically other socioeconomic predictors like education level, race/ethnicity, or healthcare access. Nonlinear modeling techniques like machine learning or deep learning models could provide better predictive accuracy if interpretation is not prioritized. Overall, this study shows that different environmental and socioeconomic indicators can be used as a solid baseline for asthma risk rate predictions at a population level.

References

Asthma and Allergy Foundation of America. 2025. "Asthma Facts." *Asthma and Allergy Foundation of America*. https://aafa.org/asthma/asthma-facts/.

Bosson, J., J. Pourazar, B. Forsberg, E. Adelroth, T. Sandström, and A. Blomberg. 2007. "Ozone Enhances the Airway Inflammation Initiated by Diesel Exhaust." *Respiratory Medicine* 101 (6): 1140–46. https://doi.org/10.1016/j.rmed.2006.11.010.

Heydari, S., M. Tainio, J. Woodcock, and A. de Nazelle. 2020. "Estimating Traffic Contribution to Particulate Matter Concentration in Urban Areas Using a Multilevel Bayesian Meta-Regression Approach." *Environment International* 141: 105800. https://doi.org/10.1016/j.envint.2020.105800.

Hwang, H., J. H. Jang, E. Lee, H. S. Park, and J. Y. Lee. 2023. "Prediction of the Number of Asthma Patients Using Environmental Factors Based on Deep Learning Algorithms." *Respiratory Research* 24 (1): 302. https://doi.org/10.1186/s12931-023-02616-x.

McConnell, Rob, Tanzina Islam, Ketan Shankardass, Michael Jerrett, Fred Lurmann, Frank Gilliland, et al. 2010. "Childhood Incident Asthma and Traffic-Related Air Pollution at Home and School." *Environmental Health Perspectives* 118 (7): 1021–26. https://doi.org/10.1289/ehp.0901232.

Meng, Ying-Ying, Michelle Wilhelm, Beate Ritz, John R. Balmes, Caitlin Lombardi, Anthony Bueno, et al. 2011. "Is Disparity in Asthma Among Californians Due to Higher Pollutant Exposures, Greater Susceptibility, or Both?" Sacramento, CA: UCLA Center for Health Policy Research.

National Center for Health Statistics. 2019. "National Ambulatory Medical Care Survey: 2019 Summary Tables." *Centers for Disease Control and Prevention*. https://www.cdc.gov/nchs/data/ahcd/namcs_summary/2019-namcs-web-tables-508.pdf.

Office of Environmental Health Hazard Assessment (OEHHA) and California Environmental Protection Agency (CalEPA). 2021. "CalEnviroScreen 4.0 Results." State of California; https://www.arcgis.com/home/item.html?id=be09f14bef6244e8af4da6aeed89ec03.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Spira-Cohen, Ari, Lung Chi Chen, Margaret Kendall, Ranjana Lall, and George D. Thurston. 2011. "Personal Exposures to Traffic-Related Air Pollution and Acute Respiratory Health Among Bronx Schoolchildren with Asthma." *Environmental Health Perspectives* 119 (4): 559–65. https://doi.org/10.1289/ehp.1002564.

Sun, Ke, Liu Zhu, Jelena Tadić, Zong-Liang Yang, and Xubin Zeng. 2023. "Increasing Threat of Extreme Wildfire Smoke to Air Quality in the United States." *Geophysical Research Letters* 50 (23): e2023GL106527. https://doi.org/10.1029/2023GL106527.

Tiotiu, A. I., P. Novakova, D. Nedeva, H. J. Chong-Neto, S. Novakova, P. Steiropoulos, and K. Kowal. 2020. "Impact of Air Pollution on Asthma Outcomes." *International Journal of Environmental Research and Public Health* 17 (17): 6212. https://doi.org/10.3390/ijerph17176212.

Wendt, Joshua K., Elaine Symanski, Thomas H. Stock, Wenyi Chan, and Xianglin L. Du. 2014. "Association of Short-Term Increases in Ambient Air Pollution and Timing of Initial Asthma Diagnosis Among Medicaid-Enrolled Children in a Metropolitan Area." *Environmental Research* 131: 50–58. https://doi.org/10.1016/j.envres.2014.02.007.