Predicting Life Expectancy Using Health, Social and Environment Indicators

Amit Badoni, Kriti Agrawal, Vivek Prakash, Vaishnavi Vuyyuru

Introduction

Understanding life expectancy requires examining how health, socioeconomic conditions, and environmental exposures interact to shape outcomes across populations. While previous studies have highlighted the importance of these factors, there remains a need for approaches that combine independent data collection with interpretable statistical modeling. Our work aims to address this gap by identifying the factors that most strongly influence life expectancy and evaluating how accurately these variables can predict life expectancy using regression-based methods.

Literature Review:

Khadke *et al.* [1] used the 2022 Environmental Justice Index (EJI) to demonstrate links between cardiovascular disease, diabetes, obesity, and environmental injustice. Pollution, low walkability, and hazardous sites worsen outcomes, while social stressors like unemployment, high housing costs, limited healthcare, and residential segregation compound inequities disproportionately affecting Black, Hispanic, and younger populations.

Chetty *et al.* [2] analyzed de-identified tax data and Social Security death records (2001–2014), finding higher income associated with greater life expectancy, with variation across geographic areas.

Qiu *et al.* [3] developed a tree-ensemble model to predict all-cause mortality, revealing how feature effects vary across strata. However, this approach looks solely at National Health and Nutrition Examination Survey (NHANES) data, which is aggregated at the national level for public use, with further granularity requiring data access permissions.

Purpose of Study:

Adding to the current research, we wish to validate them by taking an independent approach of collecting data and then using regression analysis for both predictive and prescriptive use-cases. We aim to find the factors which are the highest contributor to life expectancy and check at what accuracy we can predict life expectancy using these.

Goals:

1. Find the top factors which affect life expectancy in the US
2. Apply multiple regression models and compare them on specific metrics
3. Check the assumptions of linear regression with this dataset

Dataset

Description and Source:

1. **The Environmental Justice Index (EJI)** ranks census tracts by health impacts of environmental injustice, drawing from the Census Bureau, EPA, USGS, DOT, and CDC. It integrates social vulnerability, environmental burden, and health vulnerability modules across 36 indicators. Each score is a percentile (0–1).

   Link:
   https://www.atsdr.cdc.gov/place-health/php/eji/eji-data-download.html

   The description of these 36 indicators are present in the table below -

VARIABLE NAME	VARIABLE DESCRIPTION	DATA SOURCE
E_TOTPOP	Population estimate, 2018-2022 ACS	2018-2022 Census Bureau American Community Survey (ACS) 5-year Data - Table S0601https://www.census.gov/data/developers/data-sets/acs-5year.html
E_DAYPOP	Adjunct variable - Estimated daytime population, LandScan 2021	2021 Oak Ridge National Laboratory's LandScan https://www.ornl.gov/project/landscan
E_MINRTY	Percentage of persons who identify as Hispanic or Latino (of any race); Black and African American, Not Hispanic or Latino; American Indian and Alaska	2018-2022 Census Bureau American Community Survey (ACS) 5-year Data - Table DP05https://www.census.gov/data/developers/data-sets/acs-5year.html
E_POV200	Percentage of persons with income below 200% of the federal poverty level	2018-2022 Census Bureau American Community Survey (ACS) 5-year Data - Table S1701 https://www.census.gov/data/developers/data-sets/acs-5year.html
E_NOHSDP	Percentage of persons with no high school diploma (age 25+)	2018-2022 Census Bureau American Community Survey (ACS) 5-year Data - Table B06009 https://www.census.gov/data/developers/data-sets/acs-5year.html
E_UNEMP	Percentage of persons who are unemployed	2018-2022 Census Bureau American Community Survey (ACS) 5-year Data - Table DP03https://www.census.gov/data/developers/data-sets/acs-5year.html
E_RENTER	Percentage of housing units that are renter occupied	2018-2022 Census Bureau American Community Survey (ACS) 5-year Data - Table DP02 https://www.census.gov/data/developers/data-sets/acs-5year.html
E_HOUBDN	Percentage of households that make less than 75,000 who are considered burdened by housing costs (i.e., pay greater than 30% of monthly income on	2018-2022 Census Bureau American Community Survey (ACS) 5-year Data - Table S2503 https://www.census.gov/data/developers/data-sets/acs-5year.html
E_UNINSUR	Percentage of persons who are uninsured (i.e., have no health insurance)	2018-2022 Census Bureau American Community Survey (ACS) 5-year Data - Table S2701 https://www.census.gov/data/dev elopers/data-sets/acs-5year.html
E_NOINT	Percentage of persons without internet	2018-2022 Census Bureau American Community Survey (ACS) 5-year Data - Table S2801 https://www.census.gov/data/developers/data-sets/acs-5year.html
E_DISABL	Percentage of civilian noninstitutionalized persons with a disability	2018-2022 Census Bureau American Community Survey (ACS) 5-year Data - Table DP02https://www.census.gov/data/developers/data-sets/acs-5year.html
E_MOBILE	Percentage of housing units designated as mobile homes	2018-2022 Census Bureau American Community Survey (ACS) 5-year Data - Table DP04https://www.census.gov/data/developers/data-sets/acs-5year.html
E_GROUPQ	Percentage of persons living in group quarters	2018-2022 Census Bureau American Community Survey (ACS) 5-year Data - Table B26001https://www.census.gov/data/developers/data-sets/acs-5year.html
E_OZONE	The annual mean number of days above the regulatory standard for Ozone (O3), averaged over 3-years	The 2018-2020 U.S. EPA Air Quality System (AQS) data, as available through the CDC's National Environmental Public Health Tracking Network https://data.cdc.g
E_PM	The annual mean number of days above the regulatory standard for Particulate Matter 2.5 (PM2.5), averaged over 3-years	The 2018-2020 U.S. EPA Air Quality System (AQS) data, as available through the CDC's National Environmental Public Health Tracking Network https://data.cdc.g
E_DSLPM	Ambient concentrations of diesel particulate matter	The 2019 U.S. EPA AirToxScreenhttps://www.epa.gov/AirToxScreen/2019-airtoxscreen
E_TOTCR	The likelihood of developing cancer from air toxics over the course of a lifetime, assuming continuous exposure	The 2019 U.S. EPA AirToxScreen https://www.epa.gov/AirToxScreen/2019-airtoxscreen
E_NPL	The proportion of the census tract that is within a 1-mile buffer of an EPA National Priority List (NPL) Site	The U.S. EPA Facility Registry Service (FRS)https://www.epa.gov/frs/geospatial-data-download-service
E_TRI	The proportion of the census tract that is within a 1-mi buffer of an EPA Toxic Release Inventory (TRI) Site	The U.S. EPA Facility Registry Service (FRS)https://www.epa.gov/frs/geospatial-data-download-service
E_TSD	The proportion of the census tract that is within a 1-mile buffer of an EPA Treatment, Storage, and Disposal Facility (TSDF)	The U.S. EPA Facility Registry Service (FRS)https://www.epa.gov/frs/geospatial-data-download-service
E_RMP	The proportion of the census tract that is within a 1-mile buffer of an EPA Risk Management Plan (RMP) site	The U.S. EPA Facility Registry Service (FRS)https://www.epa.gov/frs/geospatial-data-download-service
E_COAL	The proportion of the census tract that is within a 1-mile buffer of a coal mine	The U.S. MSHA Mine Data Retrieval System (MDRS) https://www.msha.gov/data-and-reports/mine-data-retrieval-system
E_LEAD	The proportion of the census tract that is within a 1-mile buffer of a lead mine	The U.S. MSHA Mine Data Retrieval System (MDRS) https://www.msha.gov/data-and-reports/mine-data-retrieval-system
E_PARK	The proportion the census tract that is within a 1-mi buffer of a park or greenspace	The USGS PAD-US 4.0 https://www.usgs.gov/programs/gap-analysis-project/science/pad-us-data-download
E_HOUAGE	The percentage of houses built before 1980, to estimate potential exposure to lead	The U.S. Census Bureau American Community Survey (ACS) 2018-2022, census tract level datahttps://api.census.gov/data/2022/acs/acs5/subject/variables/S2
E_WLKIND	The National Walkability Index (NWI) rank, which ranks block groups according to their relative walkability, aggregated to the census tract level	The U.S. EPA National Walkability Index (NWI) https://catalog.data.gov/dataset/walkability-index1
E_ROAD	The proportion of the census tract that is within a 1-mi buffer of a high volume roadway or highway	U.S. DOT National Highway System (NHS) https://www.fhwa.dot.gov/planning/ national_highway_system/nhs_maps/
E_AIRPRT	The proportion of the census tract that is within a 1-mi buffer of an airport	OpenStreetMap (OSM) and the U.S. DOT National Transportation Atlas Database (NTAD) OSM: https://overpass-turbo.eu NTAD: https://www.bts.gov/ntad
E_IMPWTR	The percentage of the census tract that intersects an impaired or impacted watershed at the HUC-12 level	The U.S. EPA Watershed Index Online (WSIO) https://www.epa.gov/wsio
E_NEHD	Number of extreme heat days (days where maximum temperature exceeds 95th percentile of historical maximum temperatures for that tract)	CDC's National Environmental Public Health Tracking Network Data Explorer https://ephtracking.cdc.gov/DataExplorer/
E_SMOKE	Average annualized frequency of smoky days from wildfire smoke	NOAA; OSPO (Office of Satellite and Product Operations)https://www.ospo.noaa.gov/products/land/hms.html#data
E_CFLD	Average annualized frequency of coastal flooding events	FEMA's National Risk Index NRIhttps://hazards.fema.gov/nri/data-resources
E_DRGT	Average annualized frequency of drought events	FEMA's National Risk Index NRIhttps://hazards.fema.gov/nri/data-resources
E_HRCN	Average annualized frequency of hurricane events	FEMA's National Risk Index NRIhttps://hazards.fema.gov/nri/data-resources
E_RFLD	Average annualized frequency of riverine flooding events	FEMA's National Risk Index NRIhttps://hazards.fema.gov/nri/data-resources
E_SWND	Average annualized frequency of strong wind events	FEMA's National Risk Index NRIhttps://hazards.fema.gov/nri/data-resources

Figure 1 – Data Sample

2. **Places**: Local Data for Better Health provides model-based estimates of population health for every U.S. county, place, census tract, and ZIP Code Tabulation Area. Produced by the CDC's Division of Population Health with support from the CDC Foundation and the Robert Wood Johnson Foundation, it includes 36 indicators spanning health outcomes, preventive services, chronic disease risk behaviors, disabilities, and overall health status.

   Link:
   https://catalog.data.gov/dataset/places-local-data-for-better-health-county-data-2023-release

3. **County Health Rankings 2024 Annual Data Release:** This dataset provides county-level measures of health and the social, economic, and environmental conditions that shape health across nearly every U.S. county. Published by the University of Wisconsin Population Health Institute with support from Robert Wood Johnson Foundation, the 2024 release includes over 80 measures organized under core health outcomes (e.g., length of life, quality of life) and "health-factors" (e.g., clinical care access, socioeconomic conditions, physical environment, health behaviors).

Link: https://www.countyhealthrankings.org/sites/default/files/media/document/2024_county_health_release_data_-_v1.xlsx

Data Processing:

1. **EJI data**: The initial data was at a census tract level. In order to bring it to a county level, the following method was used.
    a. Multiply each tract's variable value (e.g., EPL_UNEMP) by that tract's population (E_TOTPOP).
    b. Sum these products across all tracts in the county.
    c. Divide by the total population of the county (sum of E_TOTPOP across its tracts).
2. **CDC PLACES data**: This data was present at a county level. However, it was not arranged in standard X|Y matrix form. Each county had a separate row for each metric (e.g. Current asthma among adults aged >=18 years). To overcome this, standard python script was written to form a pivot against county names. The main snippet is placed here. The entire code file is shared separately.

```
a. df_filtered = df[df["Data_Value_Type"] == "Age-adjusted prevalence"]
b.
c. # Pivot the table — one row per county, one column per metric
d. df_pivot = df_filtered.pivot_table(
e.     index=["StateAbbr", "StateDesc", "LocationName"],
f.     columns="Measure",
g.     values="Data_Value"
h. ).reset_index()
```

Figure 2 – Python Script

3. **CHR Dataset:** This data was processed using excel. Some of the missing values were imputed using means across the whole features. This was a reasonable assumption because the missing data was less than 2% for each feature.

After these operations were performed on these datasets, the final dataset was created by concatenating on state name and county names. The state name was used because some counties across different states had the same name.

EDA

To understand the structure of the combined dataset and evaluate which variables were ready to be analyzed, we conducted an exploratory review of our data before fitting any regression models.The merged county-level dataset contained 91 columns with a wide range of measures from health, socioeconomic, and environmental indicators. We inspected the first few rows of the dataset to verify that the counties and associated measures aligned after the data preparation. Several of the income-related fields were stored as character strings, so we converted them to numeric form to ensure we could use them for any quantitative analysis. We also standardized the column names to be

shorter and more interpretable by using the clean_names function from the janitor package. These steps produced a usable dataset for our future regression analyses.

Metrics Used

1. **VIF Statistics** - VIF quantifies how much the variance of a regression coefficient is "inflated" because of its correlation with other independent variables in the model. A high VIF indicates that a variable is highly correlated with other variables in the model, causing inflated standard errors, unreliable coefficient estimates, and potentially skewed results.
    a. VIF=1: No multicollinearity.
    b. 1<VIF<5: Moderate multicollinearity, generally not a concern.
    c. VIF>5: High multicollinearity, may require further investigation.
    d. VIF>10: Severe multicollinearity, indicating unreliable coefficient estimates.
2. **P-Value** - The p-value assesses the statistical significance of each independent variable's coefficient, indicating the probability that the variable's observed effect is due to chance if there were actually no relationship. A small p-value (conventionally below 0.05) suggests the variable is a significant predictor of the dependent variable, while a large p-value means there is likely no relationship.

Regression Approach

This section outlines the regression techniques used to understand the key determinants of life expectancy across U.S. counties. We begin with simple linear regression to examine individual relationships, then introduce multiple regression to control for confounding factors, followed by model selection using stepwise procedures, and finally test whether interaction effects improve explanatory power.

Simple Regression

Simple regression was used as an initial exploratory tool to evaluate the bivariate relationship between life expectancy and several health, socioeconomic, and environmental predictors.
Across the models, several predictors showed strong and statistically significant associations:

1. Smoking displayed a large negative effect ($\beta \approx -0.63$, $p < 0.001$), with higher adult smoking rates strongly associated with lower life expectancy.
2. Diabetes prevalence ($\beta \approx -1.16$, $p < 0.001$) and food insecurity ($\beta \approx -0.74$, $p < 0.001$) also exhibited substantial negative relationships.
3. Obesity ($\beta \approx -0.50$, $p < 0.001$) and percent Black population ($\beta \approx -8.78$, $p < 0.001$) showed sizable negative slopes.
4. Positive predictors included percent with some college education ($\beta \approx 0.19$, $p < 0.001$) and median income for White households.

These results provided an important first look at which factors are individually influential. However, because predictors are often correlated (e.g., education with income, obesity with smoking), multivariate analysis was essential for isolating their independent effects.

Multiple Regression

A full multiple regression model was fitted using all numeric predictors (87 variables). The model achieved an R^2 of 0.8068 and Adjusted R^2 of 0.8009, indicating that the set of predictors explained a large portion of the variance in county-level life expectancy.

Key findings from the multivariate model:

1. Several simple-regression predictors remained highly significant even after adjustment:
    a. Smoking ($\beta \approx -0.27$, $p < 0.001$)
    b. Food insecurity ($\beta \approx -0.22$, $p < 0.001$)
    c. High blood pressure, obesity, stroke, and multiple environmental risk indicators (e.g., e_drgt, e_totcr, e_coal)
2. Some variables that were significant in isolation (e.g., diabetes, cancer) lost significance in the multivariate context, likely due to shared variance with stronger predictors.
3. Demographic proportions such as percent American Indian/Alaska Native and percent female remained significant, while others (e.g., percent Black or percent Hispanic) were not consistently significant, suggesting confounding with socioeconomic and environmental variables.
4. A Variance Inflation Factor (VIF) analysis revealed substantial multicollinearity for many predictors (e.g., VIF > 100 for some demographic percentages). This motivated the shift to model reduction.

Model Selection and Comparison

Stepwise Regression

A stepwise AIC-based model selection procedure produced a more parsimonious model with 56 predictors, retaining variables that contributed meaningfully after penalizing model complexity. This reduced model achieved:

1. Adjusted $R^2 = 0.8014$, nearly identical to the full model.
2. All included predictors had acceptable VIF values compared to the full model.
3. Key retained predictors included:
    a. Health factors: smoking, obesity, high blood pressure, chronic disease indicators
    b. Socioeconomic factors: food insecurity, education levels, percent female
    c. Environmental burden variables: e_drgt, e_totcr, e_swnd, e_coal, e_noint
    d. Demographics: percent American Indian/Alaska Native, percent Asian
    e. Healthcare access indicators: cervical screening, older adult preventive care
    f. Population-level metrics: total population, high school graduation rate

Further Reduced Model

For conceptual clarity and lower multicollinearity, a further reduced model with 37 predictors was estimated. This model maintained strong performance:

Adjusted $R^2 = 0.7918$

Despite the large reduction in predictors, explanatory power dropped only marginally, demonstrating that many original variables were redundant.

ANOVA Comparison

An ANOVA test comparing the reduced model against the full model showed:

1. Significant difference ($p < 0.001$), but practical improvement in fit was modest relative to the number of parameters saved.

Thus, the reduced model was preferred for its interpretability and efficiency.

Interaction Terms

Smoking × Obesity Interaction: Given their conceptual linkage and strong individual effects, an interaction between smoking and obesity was tested. Both variables were mean-centered before constructing the interaction term.
The interaction model showed:

1. Significant interaction effect: $\beta \approx 0.008$, $p < 0.001$
2. Slight improvement in model fit:
    a. Adjusted R^2 increased from $0.7918 \rightarrow 0.7936$
    b. ANOVA confirmed the interaction improved fit ($p \approx 1.17e\text{-}06$)

Interpretation: The positive interaction suggests that the marginal effect of smoking on life expectancy becomes slightly less negative in counties with higher obesity (or vice versa). While statistically significant, the practical impact is small relative to main effects.

Food Insecurity x Education Interaction: Given the conceptual link between socioeconomic hardship and human capital, an interaction between food insecurity (% Food Insecure) and educational attainment (% Some College) was tested.
The interaction model showed:

1. Significant interaction effect: $\beta \approx 0.00187$, $p \approx 0.018$
2. Improvement in model fit:
    a. Adjusted R^2 increased from $0.7918 \rightarrow 0.7921$
    b. ANOVA confirmed the interaction improved fit: $p \approx 0.01875$

Interpretation: The positive interaction suggests that the negative impact of food insecurity on life expectancy is slightly weaker in counties with higher levels of college education. Although statistically significant, this moderating effect is small in magnitude and does not meaningfully change overall predictions.

Doctor Visits × % Female Interaction: To test whether preventive healthcare use varies by gender composition, an interaction between doctor_visit and pct_female was included.
The interaction model showed:

1. Interaction not statistically significant: $\beta \approx -0.051$, $p \approx 0.135$
2. No improvement in model fit:
    a. Adjusted R^2 remained essentially unchanged ($0.7918 \rightarrow 0.7919$)
    b. ANOVA confirmed no added explanatory power: $p \approx 0.1336$

Interpretation: Despite both variables being meaningful individually, the joint effect does not influence life expectancy beyond their separate contributions. There is no evidence that the relationship between preventive healthcare visits and longevity differs in counties with higher or lower proportions of women.

% Black × Education Interaction: Because structural inequities may interact with educational opportunity, an interaction between % Black population and % Some College was tested.
The interaction model showed:

1. Statistically significant interaction: $\beta \approx 0.00866$, $p \approx 0.006$
2. No improvement in model fit:
    a. Adjusted R^2: 0.7918 → 0.7917 (essentially unchanged)
    b. ANOVA confirmed no meaningful improvement: $p \approx 0.691$

Interpretation:
Although statistically significant, the interaction does not improve model performance. The coefficient suggests that higher education levels slightly reduce the negative association between % Black population and life expectancy, but the effect size is extremely small, and does not materially affect predictions.

Diagnostics

To evaluate whether the assumptions of multiple linear regression were reasonably satisfied, we examined several standard diagnostic plots, including the Residuals vs Fitted plot, Normal Q–Q plot, Scale–Location plot, partial residual plots, and leverage/Cook's distance diagnostics. Overall, the diagnostics indicate that the model performs well for large-scale county-level data, though some assumptions are only approximately met.

Residuals vs Fitted Plot

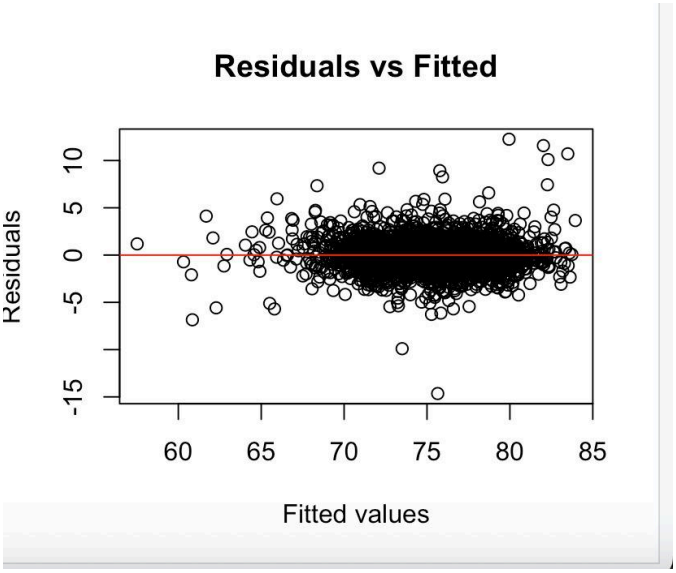

Figure 3 – Residuals vs Fitted

The Residuals vs Fitted plot shows a dense cloud of points centered around zero with no strong curvature. This suggests:

1. Linearity assumption is reasonably satisfied, as no major systematic pattern is visible.
2. No major heteroscedasticity, though the spread of residuals appears slightly wider at some fitted value ranges (around 75–80), indicating mild non-constant variance.

Overall, the pattern is acceptable for a large cross-sectional dataset.

Normal Q–Q Plot

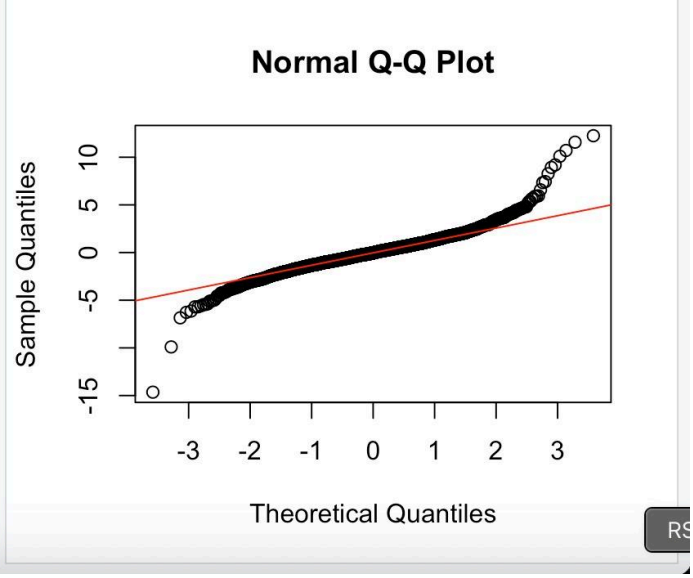

Figure 4 – Q–Q Plot

The Normal Q-Q plot shows that:

1. Residuals follow the theoretical normal line fairly closely in the middle region.
2. Deviations occur in both tails, with extreme points falling away from the line.

This indicates non-normality driven mainly by heavy tails and outliers. Because our dataset has nearly 3,000 observations, the Central Limit Theorem mitigates the effect of non-normal residuals on inference, so this is not a major concern.

Scale-Location Plot

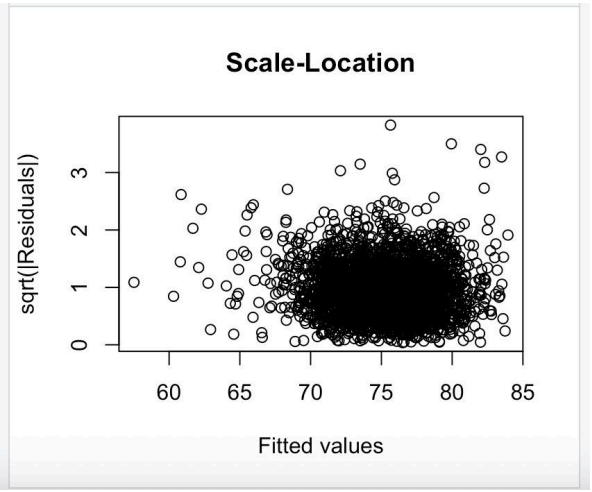

Figure 5 – Scale-Location Plot

The Scale-Location plot (sqrt|residuals| vs fitted):

1. Shows a mostly horizontal band around fitted values.

2. The spread is roughly constant, though there is slightly more variability at mid-to-high fitted values.

This suggests mild heteroscedasticity, but not enough to invalidate the model. Robust standard errors could be considered, but the effect size is likely minimal.

Partial Residual (Component + Residual) Plots

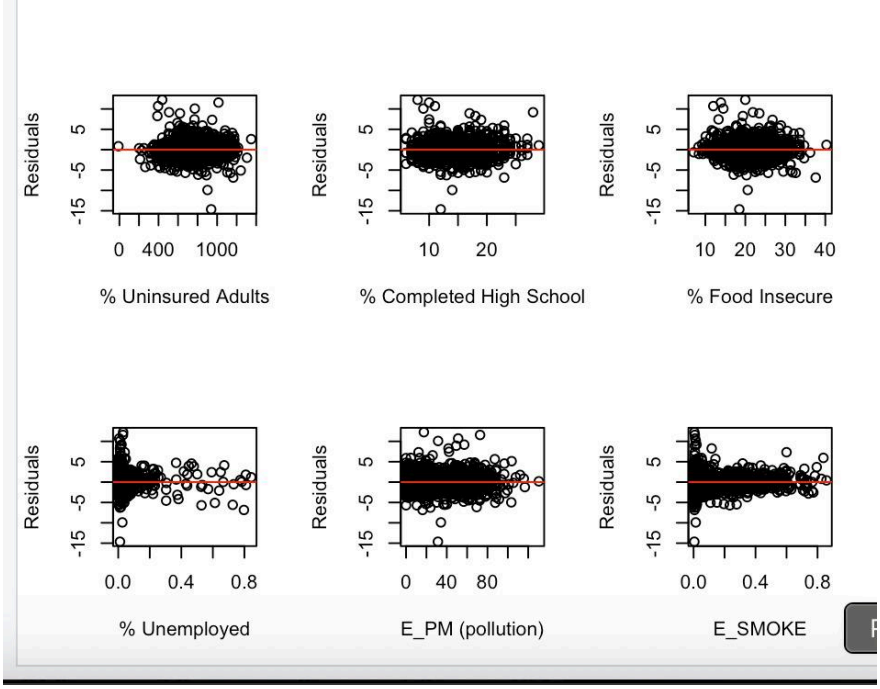

Figure 6 – Partial Residual Plots

We selected those predictors which had the least p-values in the model. The partial residual plots for these key predictors such as % uninsured adults, % completed high school, % food insecure, % unemployed, E_PM (pollution), and E_SMOKE show:

1. No strong nonlinear patterns, supporting the linearity assumption.
2. Residuals remain centered around zero, indicating that the functional form used for these predictors is appropriate.
3. Some predictors show uneven residual spread, which aligns with the mild heteroscedasticity observed earlier.

These plots validate the inclusion of these predictors in the model and confirm that linear effects are adequate.

Leverage and Cook's Distance

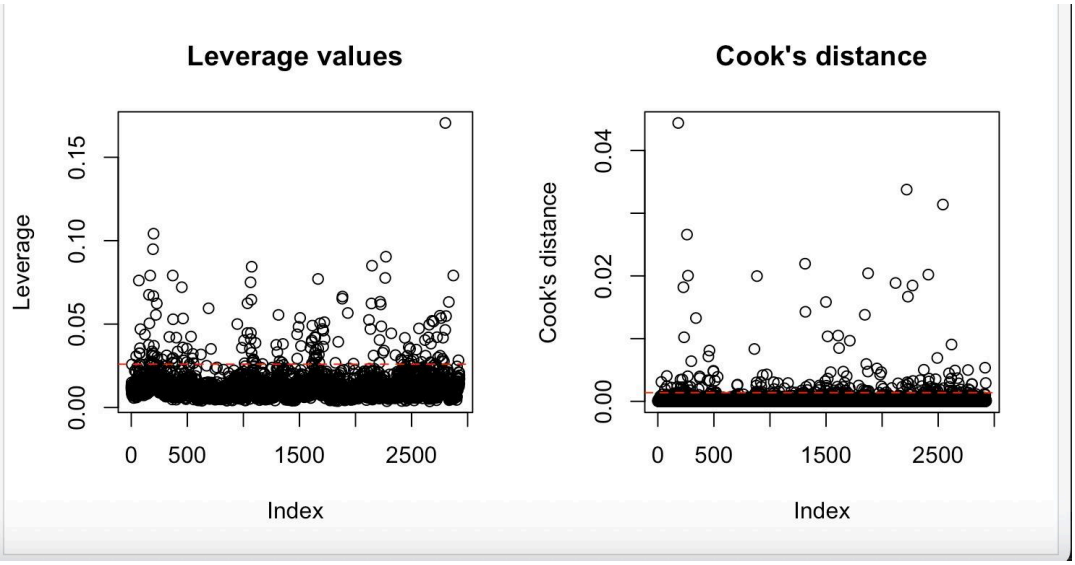

Figure 7 – Cook's Distance

The leverage plot shows:

1. Most observations have low leverage, as expected in a large dataset with many predictors.
2. A handful of points exhibit higher leverage, but none reach problematic levels.

The Cook's distance plot similarly indicates:

1. A small number of observations exert moderately high influence.
2. No points exceed the common cutoff (≈ 4/n), meaning no single county overly drives model estimates.

Thus, the model is not dominated by influential outliers, and the regression is stable.

Summary of Diagnostic Findings

1. Linearity: Reasonably satisfied. No major curvature in residuals.
2. Normality: Minor deviations in the tails, expected with large sample sizes.
3. Heteroscedasticity: Mostly constant variance; slight widening at some fitted values.
4. Influential Observations: A few moderate-leverage points, but none problematic.
5. Model Adequacy: Diagnostics indicate that the reduced model performs well and satisfies regression assumptions sufficiently for inference and interpretation.

Overall, the diagnostics support the validity of the chosen model and suggest that the fitted regression provides a reliable explanation of life expectancy variation across U.S. counties.

Prediction

To show how the reduced model can be used for forecasting, we selected a county in the test set and generated a 95% confidence interval for the mean life expectancy of counties with similar characteristics and a 95% prediction interval for the individual county. The estimated mean life

expectancy was 72.26 years with a confidence interval of [71.86, 72.65], suggesting that the estimated regression function is rather precise with this more narrow interval. The prediction interval of [69.05, 75.47] is wider and highlights the fact that there will be noise in real-world public health data, and that individual predictions will be more uncertain than population level trends.

Conclusions

In this project, we used multiple regression techniques to examine how socioeconomic, behavioral, and environmental factors relate to county-level life expectancy across the United States. By moving from simple regression to multiple regression, then model comparison and interaction effects, we were able to identify which predictors consistently explained meaningful variation in life expectancy. Across all our models, factors like smoking exposure and food insecurity showed strong negative associations with life expectancy while education measures showed strong positive associations. This reinforces public health findings that show how those who experience longer average lifespans have lower clinical risks, more access to prevention, and higher education experiences.

However, our analysis has limitations as many of the variables are correlated in the reduced model, meaning that effects could be the result of related factors. We also do not include all potential influences, like policies or economic factors. Despite limitations, we were able to use regression methods to interpret the patterns in our data and provide insights. Overall, the reduced model offered a strong balance of interpretability and predictive accuracy, and the analysis highlighted the multidimensional nature of life expectancy across U.S. counties.

Contributions

All four members contributed equally to this project. Responsibilities were divided for workflow efficiency, but each member participated in group discussions and reviewed all components of the analysis and report. The breakdown is outlined below:

1. Amit Badoni
    a. Co-led data cleaning and preprocessing across all three datasets.
    b. Contributed to regression model development and interpretation of key predictors.
    c. Helped write and revise the Data Processing and Regression sections.
2. Kriti Agrawal
    a. Co-led dataset restructuring and exploratory data analysis.
    b. Contributed to checking model assumptions and preparing diagnostic visuals.
    c. Helped write and revise the EDA and Diagnostics sections.
3. Vivek Prakash
    a. Co-led modeling work, including simple, multiple, and reduced regression models.
    b. Contributed to interaction term analysis and evaluation of model performance.
    c. Helped write and revise the Model Selection and Interaction Effects sections.
4. Vaishnavi Vuyyuru
    a. Co-led modeling work, diagnostics and validation of model assumptions.
    b. Contributed to interpretation of results and construction of final insights.
    c. Helped write and revise the Prediction, Conclusions, and Literature Review sections.

References

[1] A. Khadke, P. M. Sacco, A. I. Fernandez, J. J. Garcia, A. L. Stewart, and M. A. Carnethon, "Association of Environmental Injustice and Cardiovascular Diseases and Risk Factors in the United States," *Journal of the American Heart Association*, vol. 13, no. 4, Feb. 2024, Art. no. e033428. [Online]. Available: https://doi.org/10.1161/JAHA.123.033428

[2] R. Chetty, M. Stepner, S. Abraham, S. Lin, B. Scuderi, N. Turner, A. Bergeron, and D. Cutler, "The Association Between Income and Life Expectancy in the United States, 2001-2014," *Journal of the American Medical Association*, vol. 315, no. 16, pp. 1750–1766, Apr. 2016. [Online]. Available: https://doi.org/10.1001/jama.2016.4226

[3] W. Qiu, H. Chen, A. B. Dincer, S. Lundberg, M. Kaeberlein, and S.-I. Lee, "Interpretable machine learning prediction of all-cause mortality," *Communications Medicine*, vol. 2, art. no. 125, Oct. 2022. [Online]. Available: https://doi.org/10.1038/s43856-022-00180-x