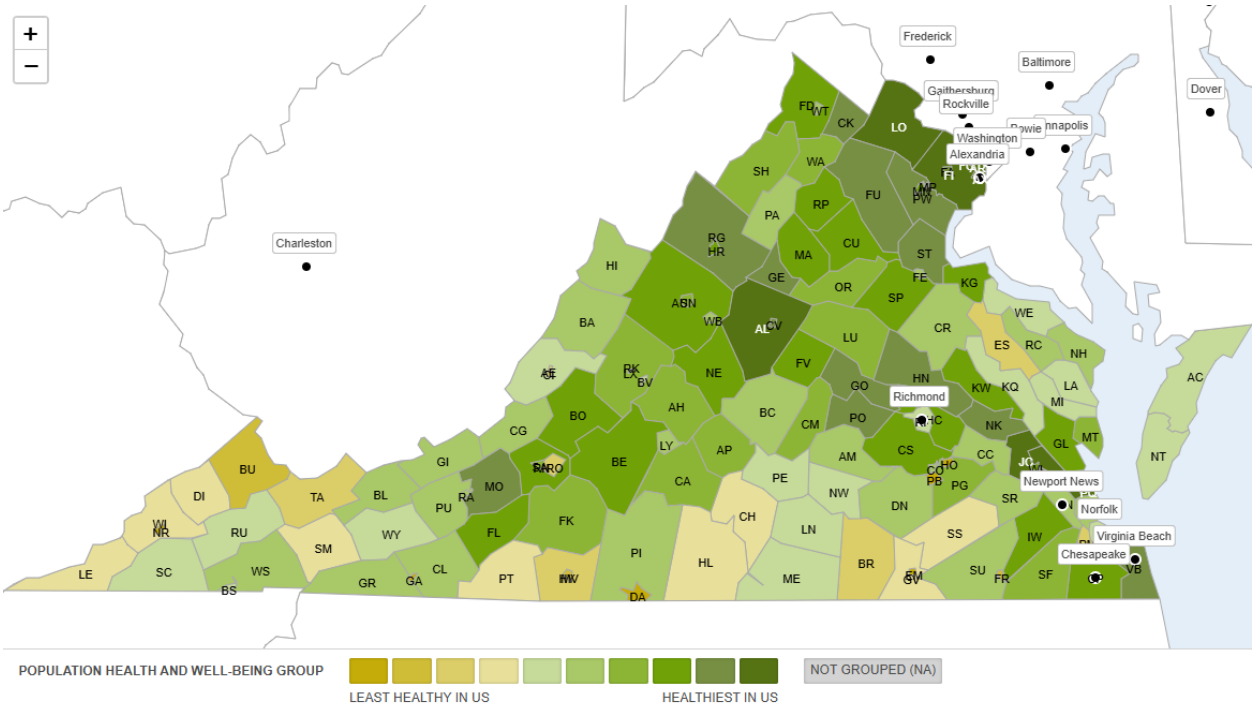
Virginia's Longevity Divide: Income, Obesity, and Exercise Access Across Counties

Team REGAN



Introduction

Despite the fact that life expectancy is frequently thought of as a succinct measure of a community's general well-being, significant differences still occur even within a single state. The *2025 County Health Rankings report*, which highlights differences in the state's progress toward public health goals, found that the average life expectancy in Virginia's **133 counties and independent cities** is just over **seventy-seven years** (County Health Rankings & Roadmaps, 2025). Decades of epidemiological research, such as the *Centers for Disease Control and Prevention's* findings on social determinants of health (Hacker, 2022) and the *World Health Organization's* analyses of obesity-related mortality (WHO, 2025), identified three recurring factors that influence longevity: household economic resources, the burden of chronic diseases (obesity), and the characteristics of local environments (such as rural versus urban areas). Previous research has established clear connections between these factors and health outcomes. *Hood* found that socioeconomic factors drive as much as **50%** of health outcomes in their County Health Rankings analysis, demonstrating the critical importance of economic resources in determining population health (Hood et al., 2016). Furthermore, the CDC's work on social determinants of health has identified significant geographic disparities, with rural communities often facing compounded challenges of limited healthcare access, lower median incomes, and higher rates of chronic disease (Hacker, 2022). Knowing how these elements appear in Virginia counties can help direct focused interventions and fair resource distribution. The overarching question guiding this study is:

Which socioeconomic, health-behavior, and environmental factors best explain the variation in 2025 life expectancy across Virginia's counties and independent cities?

The average life expectancy in years is the outcome of interest, and we created a dataset using publicly accessible data from the County Health Rankings portal, where each row represents a single jurisdiction.

Three specific research questions are presented to focus the general investigation. First, *is the average life expectancy longer in jurisdictions with greater median household incomes?* This tackles the quantifiable relationship that previous national studies have proposed between longevity and economic prosperity. Second, *is life expectancy lower in counties with the highest tertile of adult obesity rates than in those with the lowest tertile?* This investigates the relationship between mortality risk and a quantifiably modifiable health behavior variable. Third, *is there a difference in mean life expectancy between Virginia's primarily rural and urban jurisdictions?* This question investigates whether geographic context alone confers a longevity advantage or disadvantage by contrasting a qualitative classification of place.

Data Summary

Data Sources

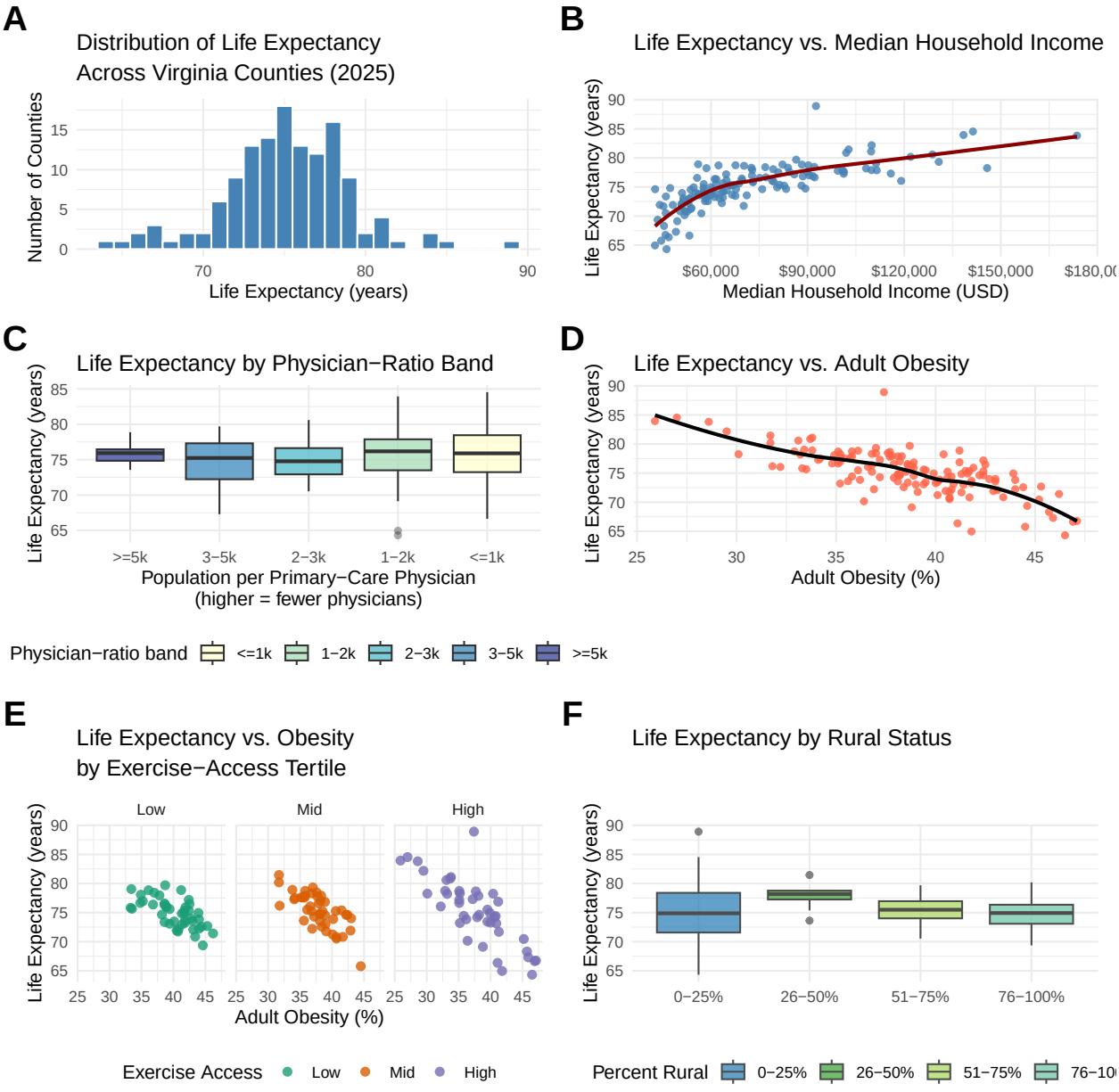
This analysis is based on the *2025 Virginia* excerpt from the *County Health Rankings & Roadmaps program*, which collects county-level health indicators across the country annually through a partnership between the *Robert Wood Johnson Foundation* and the *University of Wisconsin Population Health Institute*. The institute collects each metric directly from authorized federal sources, such as the *CDC WONDER database*, the *American Community Survey*, and the *Bureau of Labor Statistics*, before harmonizing the series to a single 2025 reference year.

For this study, the population is the complete set of **133 Virginia counties and independent cities**. Two companion tables provided by the Rankings, *Select Measure Data* and *Additional Measure Data*, contain (i) core health outcomes and (ii) socioeconomic and health-behavior covariates. These tables were combined on each county's *five-digit FIPS* code to create a single cross-sectional data frame. A continuous response variable—life expectancy at birth (years)—was retained exactly as reported, ensuring comparability with CDC methodology.

Several small, logically motivated changes were necessary. The ratio of primary care physicians was provided as a text string, such as **"2,210:1."** To ensure that lower figures indicate increased provider availability, it was transformed to a straightforward numerical count of *residents per physician*. To draw attention to *non-linear gradients* in built-environment resources, the percentage of people with access to *exercise opportunities* was recoded into an ordered three-level factor (*Low*, *Mid*, and *High*). Ultimately, the rural population percentage was divided into four equal-width groups, referred to as *rural bands*: *0–25%*, *26–50%*, *51–75%*, and *76–100%*. This discretization facilitates the understanding of regional differences while maintaining a monotonic ordering. After removing entries with missing values and one jurisdiction without a county name, **132 observations** were included in the study sample.

Because the County Health Rankings use open documentation of data provenance and imputation procedures and are regularly cited in peer-reviewed health services research, they are highly credible. Three warnings, though, are essential to note. First, *specific data may add to the uncertainty* by displaying suppressed numbers for a small number of counties. Second, since multiple focus areas of data are collected in various years, *there is no set time for retrieval*. Last but not least, measurements are gathered from *numerous trustworthy sources*, each using a different method for collecting data, and then combined into sheets. The overall integrity of the dataset is unaffected by these limitations, which will be examined.

Exploratory Data Analysis



EDA Summary

The response variable, *life expectancy at birth*, has a bell-shaped distribution with a *slight skew* (**–0.04**), a *mean* of **75.15** years, and a *standard deviation* of just **3.89 years**. Since the variable covers more than twenty-four years (**64.3–88.9**), it satisfies the *normality* and *homoscedasticity* assumptions usually needed for *multiple linear regression*; however, most counties cluster firmly between **73 and 78 years**. Therefore, it appears that no transformation is necessary.

A substantial, *positive correlation*, which steadily plateaus above **USD 90,000**, is evident when *life expectancy* is plotted against *median household income*. Together, the *LOESS fit* (**residual SE = 2.33 years**) and the *Pearson correlation* of **0.75** suggest that economic improvement in lower-income areas may result in significant longevity benefits. Extremely wealthy regions, on the other hand, already seem to be close to a ceiling. On the other hand, the *adult obesity rate* exhibits a linear, detrimental effect: a **0.68-year drop** in *life expectancy* is predicted for every percentage point rise, and the *correlation* of **–0.72** accounts for more than half of the variation (R^2**= 0.52**). When obesity is re-examined within *tertiles of access to exercise opportunities*, its relationship with longevity steepens in high-access environments (**r = –0.79**, **slope = –0.84**), suggesting that behavioral choices, rather than structural constraints, exacerbate health disparities in environments with plentiful facilities.

Provider availability also matters, though more subtly. Counties with fewer than **1,000 residents** per primary-care physician record a *median expectancy* of **75.9 years**, while mid-access bands (**2,000–5,000 residents per doctor**) drop to **74.8–75.2 years** before a slight rebound in the scarcest group. This pattern suggests that physician density exerts incremental benefits only up to a threshold, beyond which contextual factors—such as wealth, built environment, or rurality—likely dominate. The geographic context itself is significant. The most extended lifespans (**median = 78.2 years**) are found in jurisdictions with **26–50% rural populations**, outliving both highly *urban counties* (**74.8 years**) and the most *rural groups* (**= 75.5 years**). While larger dispersion within metropolitan areas (**SD = 5.65**) indicates a mix of impoverished inner-city populations and wealthy suburbs, semi-rural regions tend to be more socioeconomically homogeneous and generally healthier.

The *correlation heat-map* clarifies potential collinearity. *Age-adjusted death rate* is nearly a mirror image of the outcome itself (**r = –0.98**); therefore, it will be excluded from formal modelling to avoid redundancy. *Income and obesity* correlate at **–0.63**, a level that may inflate variance inflation factors but not so high as to demand outright removal; instead, diagnostics will verify tolerable *VIF values*. Other predictors—physician ratio, rurality, and

exercise access—display only modest intercorrelations ($|\mathbf{r}| \leq \mathbf{0.45}$), indicating that they each convey largely distinct information about county environments.

All things considered, the exploratory study shows that it is *appropriate to use multiple linear regression with life expectancy* as the continuous response. It also supports a principled variable-screening approach that maintains income, obesity, exercise access, rural band, and physician ratio as core predictors, monitors multicollinearity primarily between income and obesity, and limits the age-adjusted death rate for conceptual and statistical reasons. Because the final model will account for both quantitative gradients and categorical contexts, it will be in a good position to explain why some Virginians live noticeably longer than others.

Methods and Analysis

The multiple-linear-regression concept map's suggested sequence was adhered to by our analytical approach. Using "*na.exclude*" to preserve row indices, we re-fit an *ordinary least-squares (OLS)* model on the **126 complete cases** after exploratory analysis revealed five conceptually relevant predictors: *median household income, adult obesity prevalence, access to exercise opportunities, rurality band (orthogonally coded into three contrasts), and the primary-care-physician ratio.* The overall *F-test* was highly significant ($\mathbf{F(7,\ 118) = 46.7}$, $\mathbf{p < 0.001}$); *income and obesity** were strongly significant ($|\mathbf{t}| > \mathbf{2.9}$), access to exercise met the **0.05 threshold** ($\mathbf{t = -2.02}$, $\mathbf{p = 0.046}$). In contrast, physician density was only marginal ($\mathbf{t = 1.73}$, $\mathbf{p = 0.086}$). We regressed the *absolute OLS residuals* on the *fitted values* to look for possible *heteroscedasticity*; the positive slope showed that *variance rose* as the mean increased. Thus, we estimated a *weighted least-squares (WLS)* model with the precise specification after modeling the residual variance and computing observation-specific weights as the inverse squared fitted values.

The *WLS fit enhanced precision and goodness of fit*: the adjusted R^2 increased from **0.719 to 0.741**, and the *residual standard error* decreased from **1.87 to 1.35 years**. Multicollinearity is minimal, according to variance-inflation diagnostics, which showed that *all VIFs* for the three-df rurality factor were $\leq$ **1.81**, well below the traditional concern threshold of **10**. As an alternative remedy for non-normality and scale issues, we obtained a data-driven *Box–Cox transformation* of the response ($\lambda = \mathbf{0.13}$) and refit an unweighted model on the transformed scale. Although the *Box–Cox model* ($\lambda = \mathbf{0.13}$) preserved the same sign pattern, its fit was markedly worse (**AIC = 1608, BIC = 1633**) than both the *OLS* (**AIC = 525, BIC = 550**) and *WLS* (**AIC = 523, BIC = 548**) models, so it was not pursued further.

The WLS specification's model diagnostics were produced. A *Shapiro–Wilk* test of standard-

ized residuals supports approximate *normality* (**W = 0.985, p = 0.166**), the scale-location plot validates variance stabilization, the residual-versus-fitted panel displays *no systematic pattern*, and the *Q–Q plot* closely aligns with the **45-degree line**. Influence measures reveal eight observations with Cook's distances exceeding the 4/n threshold (= 0.032). The largest Cook's distance is 0.62—slightly above the informal **0.5 flag**—yet leverage values remain moderate, so no single county dominates the estimates, although *two observations* warrant closer substantive review (diagnostic plots are provided in the appendix.)

External validation was implemented in two ways. First, a **70 / 30 hold-out split** yielded a *root-mean-square prediction error* of **1.94 years** on the test set, consistent with in-sample residual dispersion. Second, *ten-fold cross-validation* on the complete, weight-augmented data produced an **average RMSE of 1.94** and an **average R^2 of 0.712**, indicating stable generalization performance.

Comparative evidence from fit statistics, residual behavior, and validation errors all favored the *weighted least-squares specification.* Accordingly, the WLS model was adopted as the final model for subsequent interpretation, while acknowledging the presence of a small cluster of moderately influential counties that merit substantive examination rather than statistical exclusion.

Results

A concise, statistically sound overview of *Virginia's county-level life expectancy* can be obtained by fitting the weighted least-squares (WLS) model to the **126 counties** with complete data. The model explains approximately three-quarters of the between-county variation while keeping the unexplained error *well below* the **3.9-year sample standard deviation**, with an *adjusted* R^2 **= 0.741** and a residual standard error of **1.35 years**. With an *average RMSE* of **1.944 years**, which corresponds to the hold-out split, ten-fold cross-validation confirms that predictive accuracy generalizes and stays small in comparison to public-health planning horizons.

The estimated weighted equation (weights = $1/\hat{y}^2$) on the original scale is:

$$\widehat{\text{LifeExp}}_i \text{ (years)} = 88.157 + 0.00006 \cdot \text{Income}_i - 0.408 \cdot \text{Obesity}_i - 0.02679 \cdot \text{Exercise}_i - 0.4281 \cdot \text{Rural}_{L_i}$$

$$-0.8321 \cdot \text{Rural}_{Q_i} + 0.9977 \cdot \text{Rural}_{C_i} + 0.000087 \cdot \text{PhysicianRatio}_i \quad (1)$$

where life expectancy is in years, *Income is dollars*, *percentages are entered as whole numbers*, the *three rural-band contrasts are orthogonal polynomial scores*, and *PhysicianRatio is people*

per primary-care physician. All continuous predictors were left on their natural scales.

Interpretation of the statistically important terms:

Median household income remains a *strong positive predictor*: every **$10 000 increase** is associated with an estimated **0.63-year longer lifespan**, holding all else constant (**$p < 0.001$**). Adult obesity exerts the largest adverse effect; each percentage-point reduction is associated with an estimated **0.41-year increase** in life expectancy (**$p < 0.001$**), a magnitude that rivals the income effect over realistic policy ranges. Access to exercise facilities exhibits a statistically significant but counter-intuitive negative association: moving from the *25th to the 75th percentile of access* (**≈20 percentage points**) is associated with a **0.54-year decrease** in *life expectancy* (**$p = 0.03$**).The set of orthogonal rurality contrasts is *jointly significant* (**global $F = 2.9$**, **$p = 0.026$**). The pattern indicates that counties in the mid-rural band (**≈26–50 % rural**) enjoy a modest longevity premium, whereas highly urban and highly rural counties fare worse, echoing the EDA box-plots. Physician availability displays a positive but *non-significant coefficient* (**$p = 0.20$**); its practical influence is therefore *uncertain* once economic and behavioural factors are controlled. Model diagnostics verify that the slight *mean-variance relationship* found in the *OLS residuals* is successfully attenuated by weighting, which is calculated as the inverse squared fitted values from an auxiliary variance model; *normality* is maintained (**Shapiro–Wilk $W = 0.985$**, **$df = 118$**, **$p = 0.166$**), and no single observation has undue leverage (**largest Cook's $D = 0.62$**, **leverage ≤ 0.28**). As a result, the fitted equation is a statistically sound tool for benchmarking at the county level, and coefficient standard errors are trustworthy.

The model's usefulness from a policy standpoint is in measuring achievable longevity gains: for example, a realistic **5% decrease** in adult obesity is statistically associated with roughly a **2-year difference** in *life expectancy*, which is comparable to the difference between the **25th and 75th income percentiles**. Although causal claims should be avoided due to the cross-sectional nature of the model and the potential for residual confounding, the model can guide resource allocation with reasonable precision because prediction errors are usually less than two years.

Conclusions

The majority of *2025 health disparities* can be explained by five modifiable factors: *median household income, adult obesity, access to exercise opportunities, rural makeup (orthogonal contrasts), and (to a lesser extent) physician supply.* Our weighted-least-squares model, fitted to the **126 Virginia counties** with complete data, explains roughly **74%** of the county-

level *variation in life expectancy*. Behavioral and economic levers have similar returns: a **10% increase** in *median family income* predicts around **0.6 extra years of life**, whereas a **1-percentage-point decrease** in *adult obesity* is linked to approximately **0.4 additional years**.

For illustration, in reference to equation (1) in the previous section, a county with a **$65,000 median income**, **30% obesity**, **40% exercise access**, a rural band of **26–50%**, and **1,600 residents per physician** is predicted to live **75.3 years**. Lowering obesity to **25 %** raises the forecast to **77.3 years**, a **2.0-year gain** that exceeds the model's cross-validated **RMSE of 1.94 years**.

Model diagnostics show that the residuals are almost *normal* (**Shapiro–Wilk W = 0.985**, **p = 0.166**), that *inverse-variance weights reduce heteroscedasticity*, and that there are very few influential points (**max Cook's D = 0.62**). This confirms the stability of the coefficients and makes the prediction equation helpful for policy planning. A cross-validated **RMSE of 1.94 years** shows that the forecast error is smaller than the observed county spread. This means that the model can reliably rank counties where targeted obesity reduction or income augmentation will lead to the most significant increases in life expectancy.

There are a few things that make inference less reliable. The cross-sectional design can't tell the *difference between cause and correlation*. Residual confounding (such as the quality of the built environment or unmeasured comorbidities) could *change effect sizes*, and multicollinearity between income and obesity could *hide independent contributions* even with diagnostic tolerances. We still don't know what the impacts of the physician-to-patient ratio are. Future research should examine the impact of specialty mix and telemedicine availability on clinical capacity.

To make the evidence stronger, we suggest *(1) putting together multi-year panel data to see how things change within a county*, which would allow for fixed-effects or difference-in-differences analysis; *(2) linking Medicaid claims and social-determinant datasets to get detailed information on how people use care*; *(3) testing non-linear and spatial models to find geographic spillovers*; and *(4) putting policy evaluation metrics like obesity-reduction campaigns into a quasi-experimental framework*. These processes will turn the results from correlations into useful, causally based advice for fair health-equity interventions.

Appendix A: Data Dictionary

Variable Name	Abbreviated Name	Description
life_expectancy	Life Expectancy	Average life expectancy at birth in each Virginia county, measured in years (response variable)
median_household_income	Median Income	Median annual household income in U.S. dollars (USD); inflation-adjusted to 2023 dollars
percent_adults_with_obesity	Adult Obesity %	Share of adults (age $\geq$ 20) classified as obese (BMI $\geq$ 30), expressed as a percentage of county population
percent_with_access_to_exercise	Exercise Access %	Percentage of residents living within a reasonable distance ($\leq$ 3 miles urban / $\leq$ 10 miles rural) of parks or recreational facilities
exercise_tertile	Exercise Access Tertile	Categorical split of percent with access to exercise opportunities into Low, Mid, High thirds to capture nonlinear effects
primary_care_physicians_ratio	PCP Ratio	Population per one primary-care physician (e.g., a value of 2 500 means 2 500 residents per doctor); lower values imply better access.

Variable Name	Abbreviated Name	Description
phys_ratio_band	PCP Ratio Band	Ordinal grouping of primary_care_physicians_ratio: $\leq 1k$, 1–$2k$, 2–$3k$, 3–$5k$, $\geq 5k$ residents per physician
rural_band	Rurality Band	Percent of county population living in rural census tracts, binned as 0–25 %, 26–50 %, 51–75 %, 76–100 %. Encoded with orthogonal contrasts in the regression
age_adjusted_death_rate	Death Rate	Age-adjusted all-cause mortality rate per 100 000 residents (excluded from final model to avoid redundancy with outcome)
income_ratio	Income Ratio	Ratio of mean household income in the top 20 % of earners to that of the bottom 20 %; unitless indicator of within-county income inequality

Appendix B: Data Rows

```
## # A tibble: 6 x 623
##   fips  state.x  county.x  life_expectancy x95_percent_ci_low_5
##   <chr> <chr>    <chr>               <dbl>                <dbl>
## 1 51000 Virginia <NA>                 77.6                 77.5
## 2 51001 Virginia Accomack             73.8                 72.8
## 3 51003 Virginia Albemarle            81.5                 80.9
## 4 51005 Virginia Alleghany            73.1                 71.5
## 5 51007 Virginia Amelia               73.6                 72.0
## 6 51009 Virginia Amherst              75.5                 74.5
## # i 618 more variables: x95_percent_ci_high_6 <dbl>,
## #   life_expectancy_hispanic_all_races <dbl>,
## #   life_expectancy_hispanic_all_races_95_percent_ci_low <dbl>,
## #   life_expectancy_hispanic_all_races_95_percent_ci_high <dbl>,
## #   life_expectancy_non_hispanic_aian <lgl>,
## #   life_expectancy_non_hispanic_aian_95_percent_ci_low <lgl>,
## #   life_expectancy_non_hispanic_aian_95_percent_ci_high <lgl>, ...
```

Appendix C: Final Model Output and Plots

Table 2: Table C-1 · Weighted least-squares coefficients with 95% CIs

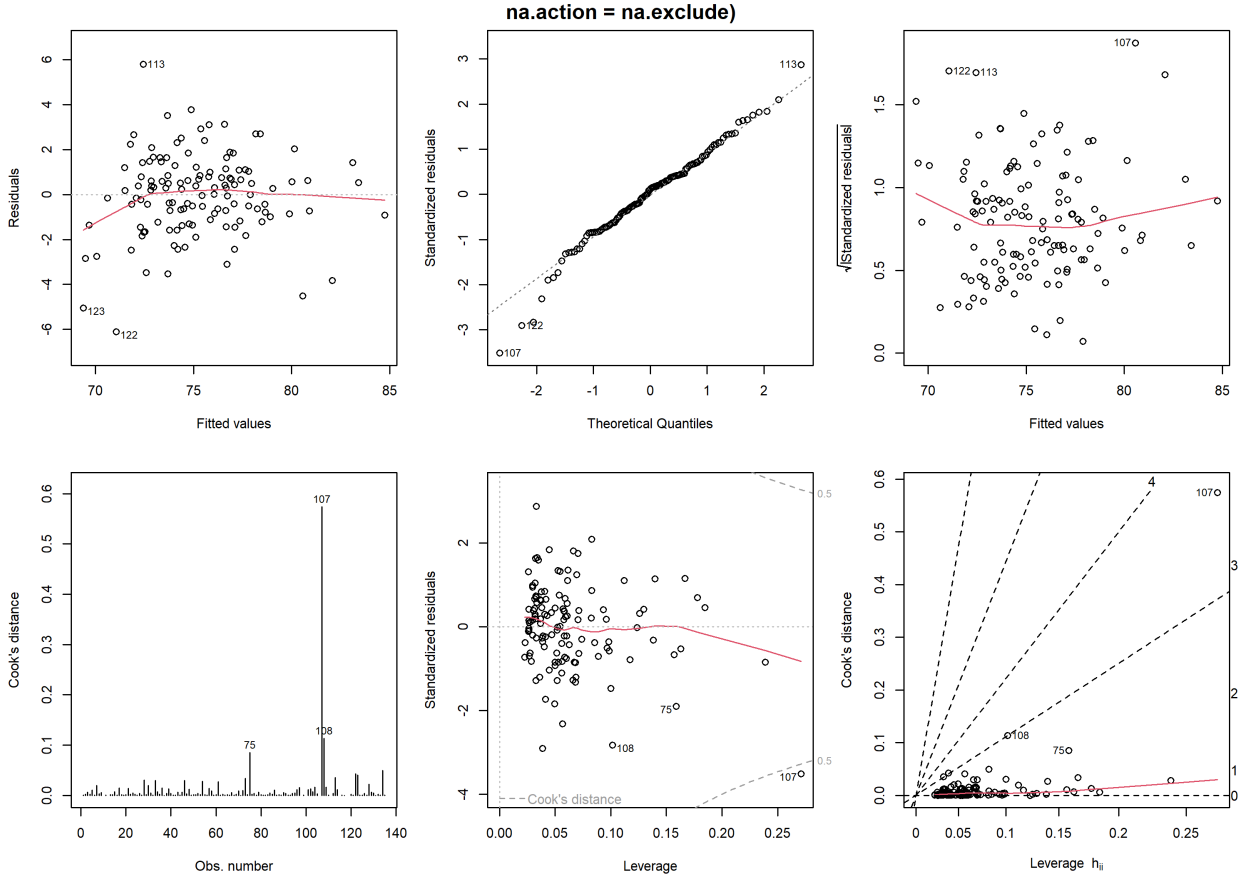
term	estimate	conf.low	conf.high	p.value
(Intercept)	88.15733	81.99210	94.32256	0.00000
median_household_income	0.00006	0.00004	0.00008	0.00000
percent_adults_with_obesity	-0.40795	-0.52856	-0.28735	0.00000
percent_with_access_to_exercise_opportunities	-0.02679	-0.05088	-0.00271	0.02954
rural_band.L	-0.42810	-1.39996	0.54376	0.38481
rural_band.Q	-0.83211	-1.64511	-0.01912	0.04493
rural_band.C	0.99772	0.01629	1.97914	0.04637
primary_care_physicians_ratio	0.00009	-0.00005	0.00022	0.20180

Table 3: Table C-2 · Model-level fit statistics

r.squared	adj.r.squared	AIC	BIC	sigma
0.756	0.741	522.682	548.208	1.345

Table 4: Table C-3 · Variance-inflation factors

Term	VIF
median_household_income	1.60
percent_adults_with_obesity	1.62
percent_with_access_to_exercise_opportunities	1.81
rural_band	1.21
primary_care_physicians_ratio	1.15


lm(formula = life_expectancy ~ median_household_income +
percent_adults_with_obesity +
percent_with_access_to_exercise_opportunities + rural_band +
primary_care_physicians_ratio, data = full_df, weights = w,
na.action = na.exclude)
Residuals
Fitted values
Standardized residuals
Theoretical Quantiles
√|Standardized residuals|
Fitted values
Cook's distance
Obs. number
Standardized residuals
Leverage
Cook's distance
Cook's distance
Leverage h_{ii}

Appendix D: References

Background

Cm, H., Kp, G., Gr, S., & Bb, C. (2016). County Health Rankings: Relationships Between Determinant Factors and Health Outcomes. *PubMed*. Retrieved July 11, 2025, from https://pubmed.ncbi.nlm.nih.gov/26526164/

Hacker, K., Auerbach, J., Ikeda, R., Philip, C., & Houry, D. (2022). Social Determinants of Health—An Approach Taken at CDC. *Journal of Public Health Management and Practice*, 28(6), 589–594. https://doi.org/10.1097/PHH.0000000000001626

Obesity and overweight. (2025, May). WHO. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Data Sources

CDC WONDER. (2024). Retrieved July 10, 2025, from https://wonder.cdc.gov/

Census Bureau Data. (2024). Retrieved July 10, 2025, from https://data.census.gov/

U.S. Bureau of Labor Statistics. (n.d.). Bureau of Labor Statistics. Retrieved July 12, 2025, from https://www.bls.gov/

Virginia. (2025). County Health Rankings & Roadmaps. https://www.countyhealthrankings.org/health-data/virginia?year=2025&measure=Population+Health+and+Well-being&mapView=state

Additional Help

Cran/car. (2025). [R]. cran. https://github.com/cran/car (Original work published 2014)

Cran/e1071. (2025). [R]. cran. https://github.com/cran/e1071 (Original work published 2014)

Cran/MASS. (2025). [R]. cran. https://github.com/cran/MASS (Original work published 2014)

Cran/RColorBrewer. (2025). [R]. cran. https://github.com/cran/RColorBrewer (Original work published 2014)

Kuhn, M. (2025). *Topepo/caret* [R]. https://github.com/topepo/caret (Original work published 2014)

Pedersen, T. L. (2025). *Thomasp85/patchwork* [R]. https://github.com/thomasp85/patchwork (Original work published 2017)

R-lib/scales. (2025). [R]. R infrastructure. https://github.com/r-lib/scales (Original work published 2010)

Tidymodels/broom. (2025). [R]. tidymodels. https://github.com/tidymodels/broom (Original work published 2014)

Tidyverse/dplyr. (2025). [R]. tidyverse. https://github.com/tidyverse/dplyr (Original work published 2012)

Tidyverse/ggplot2. (2025). [R]. tidyverse. https://github.com/tidyverse/ggplot2 (Original work published 2008)

Wilkelab/cowplot. (2025). [R]. Wilke Lab. https://github.com/wilkelab/cowplot (Original work published 2014)

willdm1. (2025). *Willdm1/3220-Final-Project* [Computer software]. https://github.com/willdm1/3220-Final-Project (Original work published 2025)

Zhu, H. (2025). *Haozhu233/kableExtra* [R]. https://github.com/haozhu233/kableExtra (Original work published 2015)