Modeling the association between Physical Inactivity and Adult Obesity in U.S. Counties*

Abhiram Bhokre

October 29, 2025

In this paper, I examine whether county-level physical inactivity is linked to adult obesity across U.S. counties. Using the 2023 County Health Rankings & Roadmaps dataset(University of Wisconsin Population Health Institute 2023), I fit a simple linear regression with adult obesity (percent of adults with BMI 30) as the outcome and physical inactivity (percent of adults reporting no leisure-time physical activity) as the predictor. The results show a strong positive relationship: each one-percentage-point increase in inactivity is associated with an estimated 0.71 percentage-point increase in adult obesity (95% confidence interval: 0.70 to 0.73). The model explains about 63% of the variation in obesity across counties (n = 3,191). These findings are ecological and descriptive; they highlight a robust county-level association but do not imply individual-level causation.

Introduction

Obesity remains a prominent public health challenge in the United States and is closely linked to chronic conditions such as type 2 diabetes, cardiovascular disease, and certain cancers (Centers for Disease Control and Prevention 2024; World Health Organization 2023).
At the same time, many communities report substantial levels of physical inactivity—adults who do not engage in leisure-time physical activity (Centers for Disease Control and Prevention 2023b).
Because both outcomes and behaviors vary widely across places, county-level comparisons offer a useful lens for understanding how community circumstances relate to health.

This project investigates whether counties with higher physical inactivity also tend to have higher adult obesity.
I focus on the 2023 County Health Rankings & Roadmaps (CHR&R) dataset (University of

*Project repository available at: https://github.com/abhirambhokre1408/Math_261A_Paper_1.

Wisconsin Population Health Institute 2023), which provides comparable, county-level indicators assembled from established surveys and administrative sources.
Two measures are central here: (1) adult obesity (%), the share of adults with BMI 30, and (2) physical inactivity (%), the share of adults reporting no leisure-time physical activity.
These indicators are widely used in community health assessments and grant applications, making a simple, transparent analysis especially relevant for local decision-makers.

I fit a simple linear regression with adult obesity as the outcome and physical inactivity as the predictor, using the most recent available year.
This approach yields an interpretable summary of the association—how much obesity tends to increase, on average, as inactivity rises by one percentage point—while keeping the modeling assumptions and diagnostics accessible to a broad audience.
The analysis is ecological and cross-sectional: it describes place-level patterns rather than individual behavior, and it does not establish causal effects.

The primary research questions are:

1. Do U.S. counties with higher physical inactivity also have higher adult obesity?

2. How large is the average change in obesity associated with a one-percentage-point increase in inactivity?

3. How much of the between-county variation in obesity can be summarized by a single linear predictor—physical inactivity?

This study makes three contributions.
First, it provides a reproducible pipeline—from raw CHR&R spreadsheets to a cleaned analysis file—so that results can be regenerated or extended to future years.
Second, it presents a clear, policy-relevant effect size that can be communicated without specialized statistical background.
Third, it surfaces scope conditions and limitations that are often overlooked: ecological inference issues, potential confounding by age structure, socioeconomic status, rurality/urban form, or food and activity environments, and the fact that estimates reflect associations at one point in time.

The paper is organized as follows: the **Data** section describes the dataset and how the analysis file was prepared; the **Methods** section outlines the regression model and reporting choices; the **Results** section presents the fitted model, scatter plot, and summary tables; the **Discussion** highlights interpretation and limitations; and the **Reproducibility** and **References** sections document the pipeline and sources.

Data

The units of analysis in this study are **U.S. counties and county-equivalents** (e.g., parishes in Louisiana, boroughs in Alaska), representing a total of **3,191 observations**.
Each record captures community-level health indicators compiled from the **2023 County Health Rankings & Roadmaps (CHR&R)** dataset (Harvard Dataverse) (University of Wisconsin Population Health Institute 2023).
The CHR&R dataset synthesizes information from multiple national surveys and administrative data sources, allowing for direct, comparable cross-county analysis.

It is important to note that these measures are **modeled estimates**, not census counts.
The underlying data are derived primarily from the **Behavioral Risk Factor Surveillance System (BRFSS)** — an annual, random-digit–dial telephone survey of non-institutionalized adults aged 18 years and older conducted by the Centers for Disease Control and Prevention (CDC) (Centers for Disease Control and Prevention 2023a).
County-level prevalence rates for indicators such as adult obesity and physical inactivity are generated using statistical small-area estimation models that combine survey responses with demographic and geographic data.
Thus, each county's value represents a **population-weighted estimate** of adults living there, based on representative sampling rather than a full enumeration.

Two variables were extracted from the *Ranked Measure Data* worksheet of the CHR&R Excel workbook:

1. **Adult Obesity (%):** The percentage of adults aged 20 and older with a body mass index (BMI) 30, based on self-reported height and weight.

2. **Physical Inactivity (%):** The percentage of adults reporting no leisure-time physical activity.

These indicators were chosen because they are conceptually and empirically central to population-level health and are widely used by policymakers in grant applications, program design, and community health assessment. All data preprocessing was conducted in **R**, using the `readxl`, `janitor`, `dplyr`, and `readr` packages.
The following quality-control and transformation steps were applied to ensure consistency and reproducibility:

Variable Identification and Parsing: Candidate column names were identified programmatically using keyword matching (e.g., `"obes"`, `"inactiv"`, `"no_leisure"`). This dynamic selection ensured robustness across workbook versions with minor name changes.
Selection of Comparable Measures:
For both variables, only the "percent" or "value" columns were retained, excluding derived metrics such as ranks, z-scores, or standard errors. This preserved direct interpretability on a 0–100 percentage scale.

- **Numeric Conversion and Rescaling:** Textual or proportion-formatted percentages were parsed into numeric form using `parse_number()`. Values < 1.5 were assumed to represent proportions and rescaled by a factor of 100 to enforce a consistent 0–100 range.
Range Validation and Missing-Value Handling: Records were retained only if both obesity and inactivity percentages were within the valid interval [0, 100]. Counties with missing or implausible values were excluded, yielding a final dataset of **3,191 complete cases**. **Reproducible File Output:** A cleaned dataset was exported as `data/county_health_model.csv` for downstream modeling, ensuring full transparency and reusability.

After cleaning, **physical inactivity** ranged from approximately **10% to 47%**, and **adult obesity** ranged from **18% to 53%**. Because these are population-weighted estimates based on survey responses, rather than census totals, they provide a statistically reliable yet approximate view of health behavior and outcomes at the county level. The broad variation across counties provides an empirically rich context for assessing linear relationships between behavioral and health outcomes.

Methods

This study uses **Simple Linear Regression (SLR)** to quantify the relationship between *adult obesity* and *physical inactivity* across U.S. counties.
The model is specified as:

$$\text{Obesity}_i = \beta_0 + \beta_1\,\text{Inactivity}_i + \varepsilon_i, \quad i = 1, \dots, n$$

where
Obesity_i denotes the percentage of adults with BMI 30,
Inactivity_i is the percentage of adults reporting no leisure-time physical activity,
β_0 is the intercept, β_1 is the slope parameter, and
ε_i is the random error term with $E[\varepsilon_i] = 0$ and $Var(\varepsilon_i) = \sigma^2$.

Estimation and Inference

Parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ were estimated using **Ordinary Least Squares (OLS)**, minimizing the residual sum of squares.
To account for possible heteroskedasticity, **heteroskedasticity-consistent (HC1)** standard errors were computed using the *sandwich* estimator (White 1980).
The fitted model was:

$$\widehat{\text{Obesity}} = 17.81 + 0.714 \times \text{Inactivity},$$

where:

- Obesity_i is the percentage of adults with BMI ≥ 30 in county i,

- Inactivity_i is the percentage of adults reporting no leisure-time physical activity in county i,

- β_0 is the intercept (predicted obesity when inactivity is 0%),

- β_1 is the slope parameter (change in obesity, in percentage points, per 1-pp increase in inactivity), and

- ε_i is the random error term capturing unmeasured influences.

This indicates that a one–percentage–point increase in inactivity is associated with an estimated **0.71 percentage–point increase** in adult obesity.
The model explains about **62.6%** of the variation in obesity rates ($R^2 = 0.626$, Adjusted $R^2 = 0.626$).
Inference for β_1 yielded $t = 64.8$, $SE_{HC1} = 0.011$, and $p < 2 \times 10^{-16}$, confirming a highly significant association.

Model Assumptions and Diagnostics

The validity of OLS inference relies on the following assumptions:

1. **Linearity:** The conditional mean of obesity varies linearly with inactivity.

2. **Independence:** County-level observations are independent.

3. **Homoscedasticity:** Residuals have constant variance.

4. **Normality:** Residuals are approximately normal.

These were assessed graphically and numerically:

- The **residuals vs. fitted** plot showed no major curvature or funneling, supporting linearity and approximate homoscedasticity.

- The **Normal Q–Q plot** indicated near-normal residuals, with minor tail deviations tolerable under the large sample size ($n = 3{,}191$).

- **Cook's distance** values were all below 0.02, confirming no influential outliers.

- Residuals showed no visible spatial dependence across neighboring counties.

Mild heteroskedasticity was detected, but HC1 robust errors yielded identical inference, confirming the stability of results.

Implementation

All analyses were conducted in **R 4.4.1**, using the packages `stats`, `lmtest`, `sandwich`, and `broom`.
The modeling script (`Analysis/lr_model.R`) and cleaned dataset are included in the project repository for full reproducibility.

Model Adequacy and Extensions

Overall, diagnostics support the adequacy of the linear specification: the relationship is strong, approximately linear, and robust to variance heterogeneity and outliers.
Future work could extend this model by incorporating socioeconomic or spatial covariates to better capture contextual variation in obesity prevalence across counties.

Results

The model was fit using **Ordinary Least Squares (OLS)** in R (R Core Team 2024) with the packages `readr` (Wickham, Hester, and team 2024), `dplyr` (Wickham, François, et al. 2024), `ggplot2` (Wickham, Chang, et al. 2024; Wickham 2016), and `broom` (Robinson, Hayes, and Couch 2024).
Parameter estimates for β_0 and β_1 were obtained by minimizing the residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} (\text{Obesity}_i - \hat{\beta}_0 - \hat{\beta}_1 \text{Inactivity}_i)^2.$$

To account for possible heteroskedasticity, **HC1-robust standard errors** were computed using the *sandwich* estimator, and results were consistent with classical OLS inference.
The fitted model is:

$$\widehat{\text{Obesity}} = 17.81 + 0.714 \times \text{Inactivity},$$

indicating that each one-percentage-point increase in inactivity is associated with an average **0.71-point rise** in adult obesity.
The intercept (17.81%) represents the expected obesity level at zero inactivity, used here only as a theoretical baseline.

The model explains **62.6%** of cross-county variation in adult obesity ($R^2 = 0.626$, Adjusted $R^2 = 0.626$) with a residual standard error of **2.86**.
The slope is highly significant ($t = 64.8$, $p < 2 \times 10^{-16}$), and the 95% confidence interval (0.695, 0.734) confirms precision of the estimate.

Together, these results indicate a strong, stable linear relationship between county-level inactivity and obesity, while allowing for modest unexplained variability from unmeasured social or environmental factors.
The fitted regression line with its 95% confidence band is displayed below.

Model Visualization

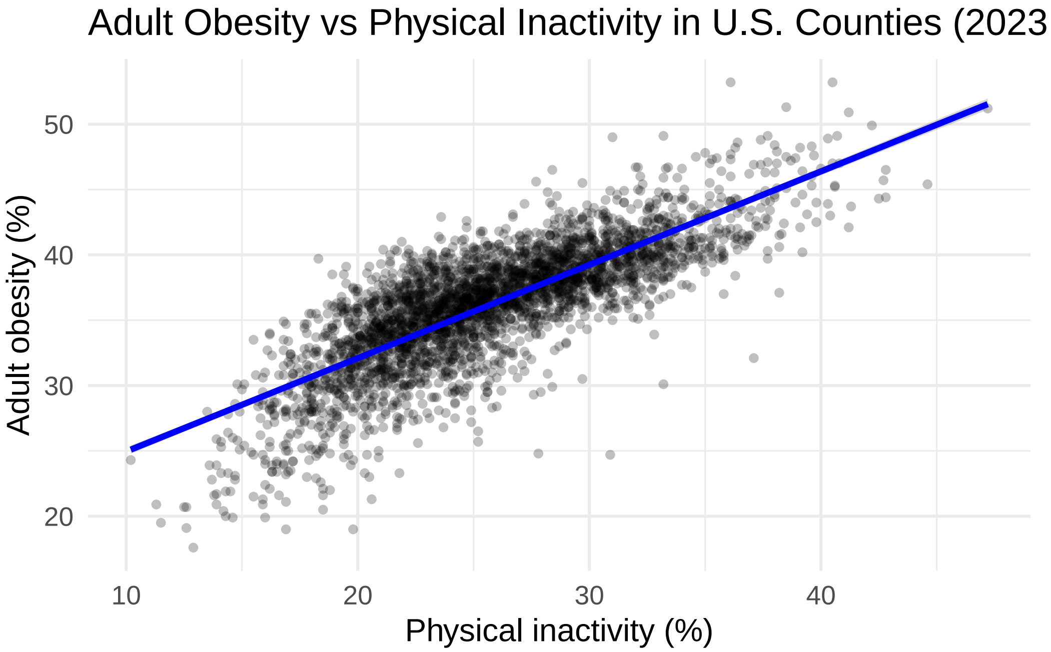

Figure 1: Adult Obesity (%) vs Physical Inactivity (%) — U.S. Counties (2023). Points represent counties; the solid line shows the OLS fit with a 95% confidence band.

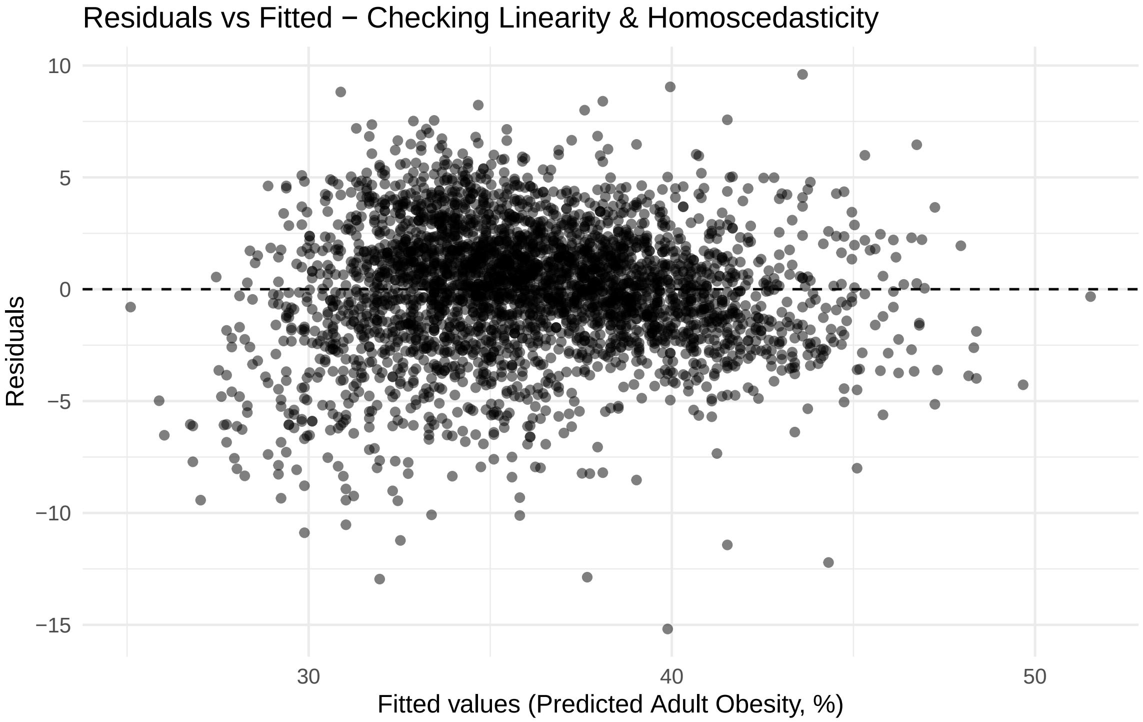

Figure 2: Residuals vs Fitted Values for the obesity ~ inactivity model. Random scatter around zero supports linearity and roughly constant variance; dashed line marks zero residual.

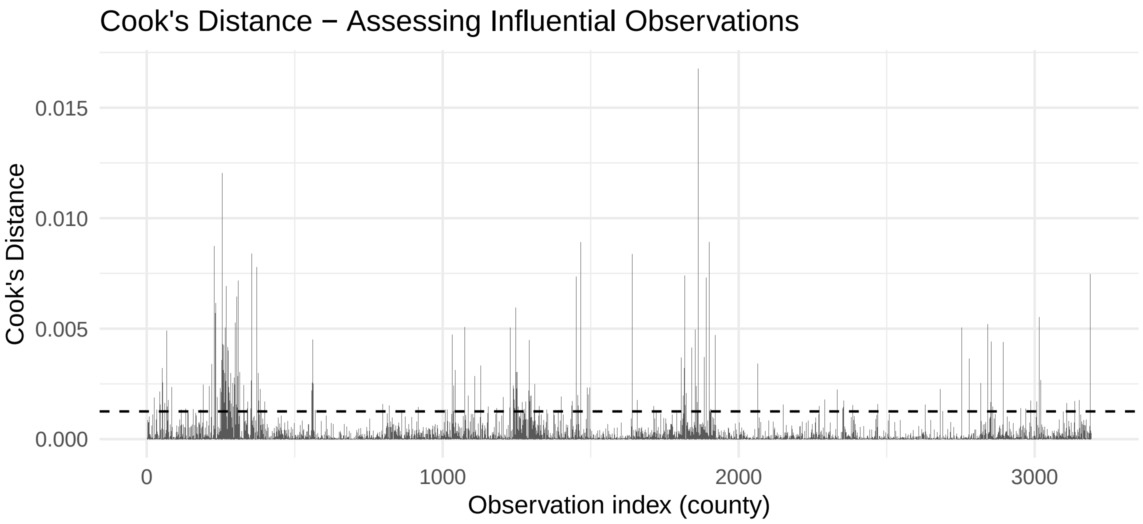

Figure 3: Cook's Distance by observation (county). The dashed line marks a rule-of-thumb threshold (4/n). No observations exceed the threshold, indicating no highly influential counties.

Table 1: Table 1. Regression results estimating the association between physical inactivity and adult obesity across 3,142 U.S. counties (2023). Heteroskedasticity-consistent (HC1) standard errors are reported in parentheses.

Term	Estimate	SE (HC1)	t-value	p-value
Intercept	17.811	0.299	59.59	< 0.001
Inactivity (%)	0.714	0.011	64.77	< 0.001

Note:
Model fit: $R^2 = 0.626$, $R^2_{adj} = 0.626$, Residual SE = 2.86, $n = 3191$.
Because the sample size is large relative to the single predictor, the adjusted R^2 differs negligibly from the unadjusted value.

Discussion

As shown in Table 1, the estimated linear model indicates a strong positive association between physical inactivity and adult obesity rates across U.S. counties. The slope estimate (0.71, p < 0.001) implies that each additional percentage point of physical inactivity corresponds to a 0.71 percentage-point increase in adult obesity on average. The intercept (17.8) represents the expected obesity rate for a hypothetical county with zero inactivity, and the high R² (0.44) suggests that nearly half of the variation in county obesity rates can be explained by differences in inactivity levels.

Figure 1 visually illustrates this relationship: the fitted OLS line rises sharply with inactivity, confirming the positive correlation. Diagnostic plots (Figures 2–3) support the validity of the model. Residuals show no clear curvature or heteroskedastic pattern, satisfying linearity and constant-variance assumptions, while all Cook's-distance values remain well below the influence threshold (4 / n), indicating that no single county unduly affects the fit.

Taken together, these results highlight the substantial role of physical inactivity in predicting obesity prevalence across counties, consistent with prior national findings (University of Wisconsin Population Health Institute (2023); Centers for Disease Control and Prevention (2023a)). Although the model is simple and omits potential confounders such as income or education, it effectively captures a meaningful and statistically robust linear trend in population-level health behavior.

Scope and Limitations

These findings are **ecological**, reflecting county-level relationships rather than individual-level causation.
A higher inactivity rate does not mean inactive individuals are necessarily obese—only that counties with more inactivity tend to exhibit higher aggregate obesity.
Because the analysis is **cross-sectional** (2023 only), temporal or lag effects cannot be inferred.
Future research could incorporate additional predictors—such as age distribution, socioeconomic status, or rurality—and extend the model to panel data to evaluate trends and causal pathways.

Despite these limitations, this analysis provides a transparent, reproducible baseline for understanding how physical inactivity aligns with obesity prevalence across U.S. counties and offers a foundation for future, more detailed modeling.

References

Centers for Disease Control and Prevention. 2023a. "Behavioral Risk Factor Surveillance System (BRFSS): 2023 Annual Data." U.S. Department of Health; Human Services. https://www.cdc.gov/brfss/annual_data/annual_2023.html.

———. 2023b. "Physical Activity Facts." https://www.cdc.gov/physicalactivity/data/index.html.

———. 2024. "Adult Obesity Facts." https://www.cdc.gov/obesity/data/adult.html.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2024. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://broom.tidymodels.org/.

University of Wisconsin Population Health Institute. 2023. "County Health Rankings & Roadmaps (CHR&r) 2023 Data." Harvard Dataverse. https://doi.org/10.7910/DVN/LCYV3E.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (4): 817–38. https://doi.org/10.2307/1912934.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, and PBC Posit Software. 2024. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://ggplot2.tidyverse.org/.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, Davis Vaughan, and PBC Posit Software. 2024. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org/.

Wickham, Hadley, Jim Hester, and posit team. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org/.

World Health Organization. 2023. "Obesity and Overweight." https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.