Transit and Opportunity: A National Network Analysis

Ivan E. Satsuta & Ishaan S. Prasad

West Valley College & University of California, Los Angeles

Author Note

Correspondence

Ivan E. Satsuta is affiliated with the Department of Economics at West Valley College. Ivan researches the applications of data science as modern econometric methods. To contact him, please email isatsuta@mywvm.wvm.edu.

Ishaan S. Prasad is affiliated with the Luskin School of Public Affairs at the University of California, Los Angeles. Ishaan researches mixed-methods approaches to social welfare and policy research. To contact him, please email iprasad@ucla.edu.

Replication

Replication data and scripts for the quantitative analyses presented in this paper are available with the Harvard Dataverse: https://doi.org/10.7910/DVN/4H71LV.

Distribution and Rights

No funding or grants were accepted in the pursuit of this research. The authors of this paper have designated this paper as an open access publication. This paper may be freely accessed through the Harvard Dataverse.

You are encouraged to cite this paper in accordance with fair use guidelines, ensuring proper citation and credit to the authors. Redistribution of this paper is permitted, provided no fee is charged for access.

Acknowledgements

The authors extend gratitude to those who have made this research possible, especially Reviewer Fatemeh Yarahmadi (fatemeh.yarahmadi@wvm.edu) for her constructive feedback throughout the development of this paper.

The authors also thank the United States Department of Transportation, the United States Census Bureau, and contributing transit agencies publishing General Transit Feed Specification data across the United States for providing data used in this study.

Abstract

Economic opportunity in the United States is closely linked to public transit access, yet disparities in funding and infrastructure create significant inequities in transit robustness, especially when compared across the U.S. While it is generally accepted that strong transit networks support economic mobility, prior research has generally been restrained to qualitative or mixed-method case studies of specific metropolitan areas to demonstrate this effect. This study addresses that gap by executing a nationwide transit network analysis and subsequent regression modeling to quantitatively examine the relationship between transit connectivity and wealth. Using data from the National Transit Map and U.S. Census, the authors model local transit networks and analyze their association with county median income. Results show that wealthier counties tend to have significantly more connected transit systems, reinforcing existing economic disparities. However, this study does not establish causality, as wealthier counties may simply have greater resources to invest in public transit. Other limitations of this study include the unweighted format of the transit network data and occasionally missing county level information, though addressing these factors will likely amplify, rather than weaken, the observed trends. This study serves as a methodological proof-of-concept for modeling transit connectivity nationally and lays the groundwork for future quantitative research on the role of transit in economic mobility.

Keywords: United States, General Transit Feed Specification (GTFS), National Transit Map (NTM), network analysis, degree centrality, closeness centrality, public transit, transit equity, economic mobility.

Introduction

Economic opportunity in the United States is inextricably linked to the structure and efficiency of public transit. While an estimated 69% of workers aged 16 or older report primarily commuting to work alone in a private vehicle, and a further 9% report carpooling, 4% report a reliance on public transit for their commuting needs (U.S. Census Bureau, 2023). It is this last demographic which generally constitutes the most upwardly mobile in society, given the right access. For individuals without private vehicles, transit networks serve as critical conduits to employment, education, and essential services; however, the design and connectivity of these systems vary significantly across the nation.

Nationally, funding for public transit is underprioritized relative to other transportation functions such as the maintenance of the interstate highway system (Mallett, 2024). When national attention does center on public transit, it is nearly always focused on statewide grant programs for large infrastructure projects, such as the electrification of Caltrain's infrastructure in California (Caltrain, 2022). States with robust transportation programs often focus directly on the development of infrastructure projects, delegating the local transit systems—the system for commuters using transit as a means to access gainful emplyoment—to the county level. At this county level, the greatest degree of inequity generally ensues.

Localized transit inequity presents a significant barrier to economic mobility, as the quality and accessibility of public transit systems can vary dramatically, between counties even within the same region. Wealthier counties tend to have extensive and well-maintained transit networks, while neighboring, less affluent counties often face underfunded, infrequent, or nonexistent service. Prior research links robust transit systems to increased access to higher-paying jobs, economic opportunity, and mobility from poverty (Sánchez *et al.*, 2003; Rambaram, 2021). However, these studies are typically limited to the cases of specific metropolitan areas, failing to fully capture the broader, nationwide implications of transit connectivity for economic mobility. For example, these studies fail to account for the potential differences in transit structure and rider behavior in rural counties, limiting the

generalizability of their findings. The authors aim to address this gap in existing research.

The concept of using a mathematical model (i.e. a network) to analyze connectedness, whether in the spread of an epidemic, the development of the World Wide Web, or flights between airports, is not novel. The cardinal work on random networks was done by Erdös and Rényi (1959) in which the degree distribution of nodes follows a poisson. This was further developed by Barabási and Bonabeau (2003) with the concept of a "scale-free," non-random social network, in which the distribution instead follows a power law. Networks are similarly useful in the study of transit connectivity as they reduce complexity into a mathematical representation consisting only of edges and nodes. This method has been used to model and study transit networks in Singapore (Soh *et al.*, 2010), China (Chen *et al.*, 2006), Poland (Sienkiewicz & Hołyst, 2005), Tel-Aviv (Sharav *et al.*, 2018), and other international metropolises (Ferber *et al.*, 2009).

This paper builds upon previous research to affirm the link between income and transit connectivity at the national level. The authors contribute to existing research in two ways. First, this paper employs a purely quantitative network-based methodology, addressing a gap in prior studies that have largely relied on qualitative or mixed methods. Second, this study expands the scope of analysis beyond individual metropolitan areas to examine nationwide trends, including regions with sparse connectivity. This paper promises new insight into the disparities of transit access and their effect on economic mobility, therefore, across diverse geographic contexts.

Research Design

This study utilizes the National Transit Map (NTM) Routes shapefile, published by the Bureau of Transportation Statistics (BTS). The NTM is built using data submissions from local transit agencies across the United States (BTS, 2024). These submissions rely on the General Transit Feed Specification (GTFS), a widely adopted data standard for representing transit schedules and operations. GTFS files provide detailed information on stop locations, route structures, service frequencies, and, in many cases, real-time vehicle locations (GTFS, 2024). Using this data, the NTM Routes shapefile is constructed by

aggregating GTFS information on routes, trips, and shapes across all participating agencies, offering a view of transit network structures nationwide (BTS, 2024).

The NTM Routes shapefile is obtained from BTS; the data is current, at the time of download, to September 18, 2024. The data is first imported into Python (Python Software Foundation, 2024) where the geographical coordinates of each routes' start and end points are retained (Gillies *et al.*, 2024; Jordahl *et al.*, 2020), along with identifying information about the state and county that the route exists in (McKinney, 2010). The data is imported into R (R Core Team, 2024) and formatted as a network edgelist where stop coordinates are replaced with a unique alphanumeric identifier (Hadley *et al.*, 2023; Urbanek & Ts'o, 2024).

The edgelist is transformed into a scale-free network, where each stop becomes a vertex and each route between stops is an edge (Csárdi & Nepusz, 2006; Csárdi *et al.*, 2025). The State and County that a stop exists in are retained as a vertex attribute. The degree and closeness centralities of each stop—given by the number of connections and the inverse of the average shortest path length to all other stops, respectively—are calculated and also assigned as vertex attributes. The network is then aggregated to the county level and summarized, retaining, for each county, the number of stops, the average degree centrality, and the average closeness centrality (Hadley *et al.*, 2023). This county level dataset is then merged with the Small Area Income and Poverty Estimates (SAIPE) data from the 2022 United States Census (U.S. Census Bureau, 2023) to add the median income of each county (Wickham *et al.*, 2023). Figure 1 summarizes the foregoing data cleaning and assembly process.

Table 1 presents the descriptive statistics for the cleaned county level data. At this point, all three variables of interest (the average degree centrality, the average closeness centrality, and the county median income) are heavily skewed. The average degree and closeness centralities follow exponential and reflected exponential distributions, respectively. We normalize the data by applying a weight to these average centralities, given by the number of stops in each county, and then transforming all three variables of interest by applying a natural logarithm. Figure 2 shows that the data follows approximately normal

distributions after adjustment (Wickham, 2016; Auguie, 2017).

Two regression models are specified (Blair *et al.*, 2024). In both, the transformed county median income is the main predictor. The responses are the transformed average degree centrality and transformed average closeness centrality, respectively. To account for unobserved heterogeneity at the state level (e.g. from policy or funding differences), states are estimated as fixed effects; standard errors are also clustered at the state level to account for intrastate error correlation. Lastly, a finite population correction is applied to account for the limited number of total counties in the United States.

Results

The regression results for both models are shown in Table 2. The transformed county median income is a significant predictor of overall transit connectivity, where higher income is associated with greater connectedness. This finding is robust to the operationalization of connectivity, whether as the transformed average degree or transformed average closeness centrality of the county's transit network. Because the main predictors and responses and both regression models were transformed by a natural logarithm. The results can be interpreted as follows: every 1% increase in county median income is associated, on average, with an approximately 2.596% increase in the weighted average degree centrality and an approximately 1.582% increase in the weighted average closeness centrality of a county's transit network.

The diagnostic plots for the first regression model, where connectivity is measured as the transformed average degree centrality, are shown in Figure 3 (Hadley, 2016; Tang *et al.*, 2016; Horikoshi & Tang, 2016). The diagnostic plots for the second regression model, where connectivity is measured as the weighted average closeness centrality, are shown in Figure 4 (Hadley, 2016; Tang *et al.*, 2016; Horikoshi & Tang, 2016). The diagnostic plots for both models are generally unremarkable and indicate that the regression assumptions are reasonably met. In both models, residuals appear to have a random scatter and homoscedastic variance. Similarly, no highly influential outliers biasing the model fit are identified in either model. The observable streaks in the residual plots for both models (Figure 3; Figure 4) are

a byproduct of applying fixed effects and do not indicate a poor model fit.

Notably, the normal quantile-quantile plots for the first and second models show distinct patterns. In both cases, the plots are generally appropriate and demonstrate a strong overall fit (Figure 3; Figure 4). However, the indications in the closeness centrality model are concerning (Figure 4). In social science research, skewed datasets are often transformed to approximate normality, but because the underlying data generation process is not naturally normal, minor dual-tailed leptokursis in the residuals is expected. This effect is observed in the degree centrality model (Figure 3); the residuals curl slightly upwards off of the reference line at the farthest theoretical quantiles, suggesting that the model overpredicts the weighted average degree centrality for most extreme median incomes on both ends (Figure 3). This finding is generally unremarkable.

What is remarkable, however, is that the residuals in the closeness centrality model curl upwards from the reference line only on the left tail, indicating asymmetric leptokursis of the residuals (Figure 4). This suggests that the model underpredicts the weighted average closeness centrality for relatively low median incomes only. The authors attribute the biased behavior of these residuals to the truncated nature of the response variable. The distribution of counties' weighted average closeness centralities approximate normality, but cannot vary below 0, even after log transformation, leading to a poorly specified model. Further, while minor symmetric leptokursis is generally non-problematic, asymmetric leptokursis will bias regression interpretation in only one direction, which is a serious flaw. Consequently, the authors conclude that the transformed average degree centrality of a county is a more suitable operationalization of regional transit connectivity.

Discussion

Both models demonstrate strong explanatory power, even after adjusting for the number of fixed effects (Table 2). However, the authors caution against drawing direct causal conclusions from this result. While the observed relationship may suggest that robust public transit improves the median income of a region, similar to patterns seen in dense urban metropolises (Sánchez *et al.*, 2003; Rambaram, 2021), it could also reflect the capacity of

wealthier counties to invest in more extensive and efficient transit networks through public funding or policy prioritization. Therefore, beyond the typical risks of omitted variable bias in observational research, this study is highly susceptible to misinterpretations due to reverse causality.

Another limitation of this study lies in the data itself. The route network is not weighted by frequency, meaning that a route running once per month is treated the same as one operating every half hour in the network analysis. Additionally, the edgelist considers only the start and end points of routes, without accounting for intermediate stops where passengers can board or deboard, which could enhance connectivity. Despite these limitations, the authors suspect that incorporating frequency data would likely amplify the observed effect, rather than diminish it. Further, it is not clear that incorporating intra-route stops would meaningfully change the model outcomes rather than simply result in a uniform rescaling of county connectivity. While both issues could be addressed through additional data processing, doing so would require integrating multiple GTFS files and possibly decentralizing the data assembly, making the process more complex and costly than using the curated NTM files.

The missing-at-random assumption of the regression models also warrants examination. The data in this study was not randomly sampled in the traditional sense; rather, any county with available data was included. Counties were excluded from the regression models for satisfying at least one of three conditions: (1) no transit agency operates within the county whatsoever; (2) no transit agency within the county published GTFS data; or (3) the Census Bureau did not report a median income estimate for that county. While the impact of missing GTFS data should be considered, the authors believe this problem poses a minimal threat to this study's inference validity.

The absence of GTFS data could indicate that few or no transit agencies operate within a county, or it may simply reflect that agencies are too small to publish such data. Similarly, the Census Bureau typically suppresses median income estimates when the sample size in the American Community Survey (ACS) is too low, as this could lead to a high margin of error or unintentional respondent identification. This sampling bias raises concerns

about the randomness of missing GTFS data at the county level, as counties with weaker transit networks are more likely to be excluded. However, the effect of missing median income data is less clear. There is no strong reason to suspect that a county's ACS response rate is systematically correlated with its median income. Thus, while the missing-at-random assumption merits consideration, there are no clear or critical threats to the validity of the regression results because of the sampling process.

Future researchers may consider additional data processing to obtain high-quality, weighted, and granular transit network data from counties across the United States. They may also explore causal inference methods—such as differences-in-differences, synthetic controls, instrumental variable estimation, or regression discontinuity designs—to establish causality where this study does not. However, the primary goal of this paper was neither to prove a causal link between transit connectivity and regional wealth nor to perfect the national data collection process. Rather, this study was intended to serve as a proof-of-concept, demonstrating a viable method for quantifying and analyzing transit connectivity nationally. This paper enables future researchers to answer causal questions about the importance of transit in the United States. To that end, the authors believe the present study has achieved its purpose.

References

Auguie, B. (2017). *gridExtra: Miscellaneous functions for "Grid" graphics* (R package version 2.3). https://CRAN.R-project.org/package=gridExtra

Barabási, A. L., & Bonabeau, E. (2003). Scale-free networks. *Scientific American*, *288*(5), 60–69. https://doi.org/10.1038/scientificamerican0503-60

Blair, G., Cooper, J., Coppock, A., Humphreys, M., & Sonnet, L. (2024). *estimatr: Fast estimators for design-based inference* (R package version 1.0.4). https://CRAN.R-project.org/package=estimatr

Caltrain. (2022, December 23). *Caltrain to receive $43 million in federal funding for electrification project*. Caltrain. https://www.caltrain.com/news/caltrain-receive-43-million-federal-funding-electrification-project

Chen, Y., Li, N., He D. (2007). A study on some urban bus transport networks, *Science Direct*, *376*(0378-4371), 747-754. https://doi.org/10.1016/j.physa.2006.10.071

Erdös, P. and Rényi, A. (1960) On the Evolution of Random Graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, *5*, 17-61. https://static.renyi.hu/renyi/_cikkek/1960/_renyi/_on/_the/_evolution/_of/_random/_graphs/_17/_61.pdf

Freeman, L. (1978). Centrality in social networks conceptual clarification. *ScienceDirect*, *1*(3), 215-239. https://doi.org/10.1016/0378-8733(78)90021-7

General Transit Feed Specification. (2024). *What is GTFS?* GTFS. https://gtfs.org/getting-started/what-is-GTFS

Gillies, S., & Contributors. (2024). *Shapely* (Version 2.0.6) [Software]. Python Software Foundation. https://shapely.readthedocs.io/

Horikoshi, M. & Tang, Y. (2016). *ggfortify: Data visualization tools for statistical analysis results*. https://CRAN.R-project.org/package=ggfortify

Mallett, W. J. (2024, January 18). *Federal support of public transportation operating expenses* (R47900). Congressional Research Service. https://crsreports.congress.gov/product/pdf/R/R47900

Python Software Foundation. (2024). *Python* (Version 3.13.1) [Software]. https://www.python.org/downloads/

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rambaram, A. M. (2021). *The Role of Transit in the Upward Mobility of Low-Income Indianapolis Residents*. Purdue University Graduate School. https://doi.org/10.25394/PGS.16826254.v1https://doi.org/10.25394/PGS.16826254.v1)

Sanchez, T. W., Stolz, R., & Ma, J. S. (2003). *Moving to equity: Addressing inequitable effects of transportation policies on minorities*. The Civil Rights Project at Harvard University. https://civilrightsproject.ucla.edu/research/metro-and-regional-inequalities/transportation/moving-to-equity-addressing-inequitable-effects-of-transportation-policies-on-minorities/sanchez-moving-to-equity-transportation-policies.pdf

Sharav, N., Bekhor, S. & Shiftan, Y. Network Analysis of the Tel Aviv Mass Transit Plan. *Urban Rail Transit*, *4*, 23–34 (2018). https://doi.org/10.1007/s40864-018-0075-7

Sienkiewicz, J., & Hołyst, J. A. (2005). Statistical analysis of 22 public transport networks in Poland. *Physical Review E*, *72*(4), 046127. https://doi.org/10.1103/PhysRevE.72.046127

Soh, H., Lim, S., Zhang, T., Fu, X., Lee, G., Hung, T., Di, P., Prakasam, S., & Wong, L. (2010). Weighted complex network analysis of travel routes on the Singapore public transportation system. *Science Direct*, *389*(24), 5852–5863. https://doi.org/10.1016/j.physa.2010.08.015

Tang, Y., Horikoshi, M., & Li, W. (2016). `ggfortify`: Unified interface to visualize statistical results of popular R packages. *The R Journal*, *8*(2), 478–489.

The igraph core team. (2024). `igraph` (Version 2.1.0) [Software]. igraph. https://igraph.org/2024/11/06/igraph-0.10.15-c.html

U.S. Census Bureau. (2023). *SAIPE State and County Estimates for 2022*. U.S. Census Bureau. https://www.census.gov/data/datasets/2022/demo/saipe/2022-state-and-

county.html

U.S. Census Bureau. (2023). Means of Transportation to Work by Selected Characteristics [Data set]. *American Community Survey, ACS 1-Year Estimates Subject Tables, Table S0802*. https://data.census.gov/table/ACSST1Y2023.S0802?t=Commuting

U.S. Department of Transportation. (2024). *National Transit Map Routes* [Data set]. Bureau of Transportation Statistics. https://geodata.bts.gov/datasets/usdot::national-transit-map-routes/about

Urbanek, S., & Ts'o, T. (2024). *uuid: Tools for generating and handling of UUIDs* (R package version 1.2-1). https://CRAN.R-project.org/package=uuid

Van den Bossche, J., Jordahl, K., & Contributors. (2024). *GeoPandas* (Version 1.0.0) [Software]. GeoPandas. https://geopandas.org/en/stable/about.html

von Ferber, C., Holovatch, T., Holovatch, Y., & et al. (2009). Public transport networks: Empirical analysis and modeling. *European Physical Journal B*, *68*, 261–275. https://doi.org/10.1140/epjb/e2009-00090-x

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A grammar of data manipulation* (R package version 1.1.4). https://CRAN.R-project.org/package=dplyr

Table 1: Descriptive statistics at the county level.

Statistic	N	Mean	Std. Dev	Min	Max
Num. Stops	952	48.098	94.561	1	1,403
Median Income	884	72,265.950	18,879.550	35,758	167,605
Avg. Degree	952	22.994	107.746	1.000	1,835.889
Avg. Closeness	915	0.734	0.257	0.029	1.000

Table 2: County income is a significant predictor for transit connectivity, with higher income associated with increased weighted average centrality.

Response:	Degree (Avg. Centrality)	Closeness (Avg. Centrality)
Intercept	-24.132***	-15.585***
	(4.402)	(2.386)
Median Income	2.596***	1.582***
	(0.398)	(0.215)
Observations	884	884
Residual SE	1.809	1.021
Residual DF	835	835
R^2	.333	.233
$R^2_{\text{adj.}}$	.295	.189

Significance Codes: * < .1, ** < .05, *** < .01.

Response variables are weighted by the number of vertices (stops) in each county. All variables, including county median income, are transformed by natural logarithm. A finite population correction accounts for the limited number of unique U.S. counties. State fixed effects are included but not shown; standard errors are clustered at the state level.

Figure 1: Data undergoes multiphasic cleaning and assembly process.

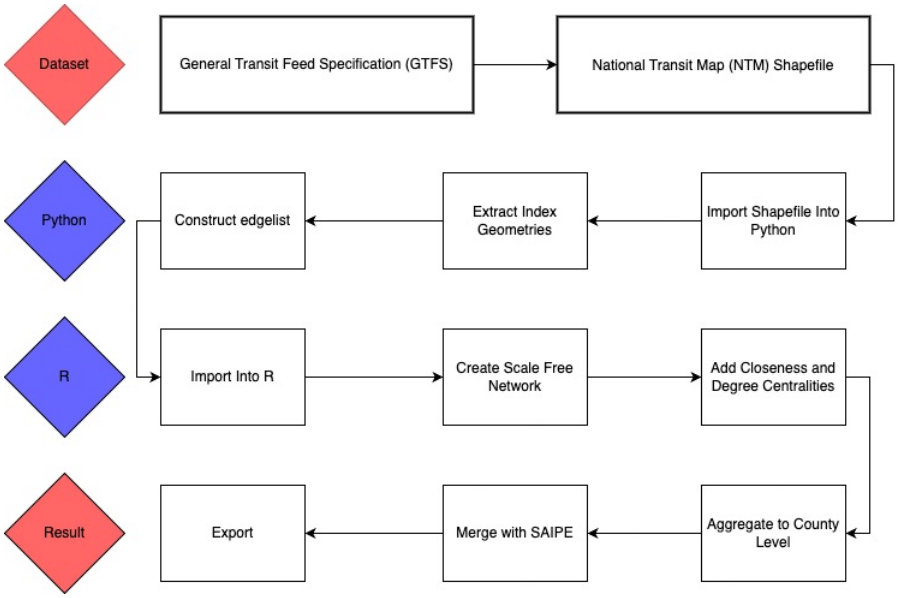
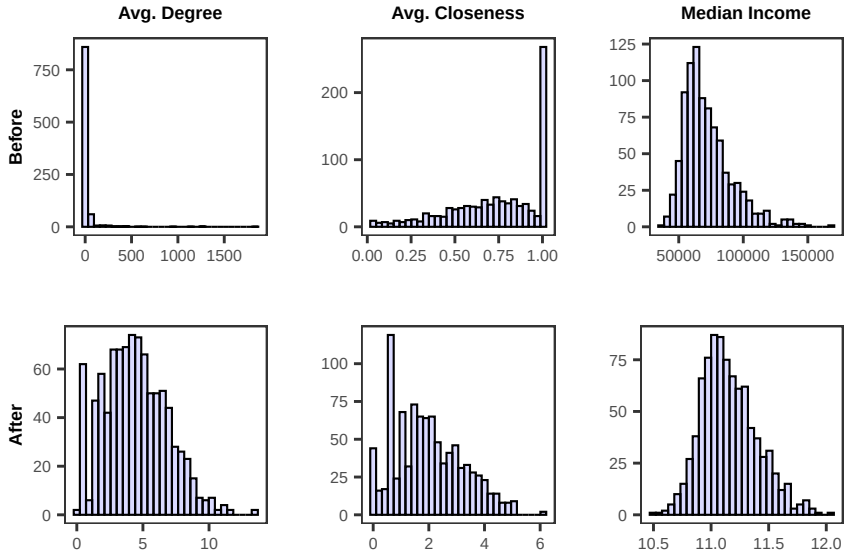
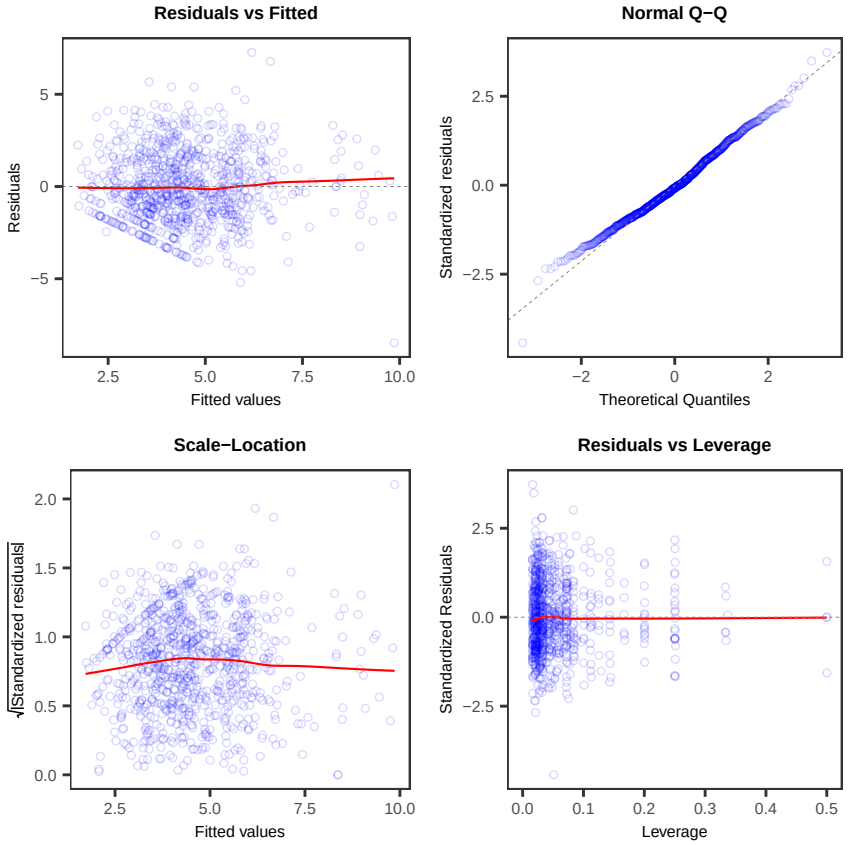


Figure 2: Variable transformations approximately normalize data.



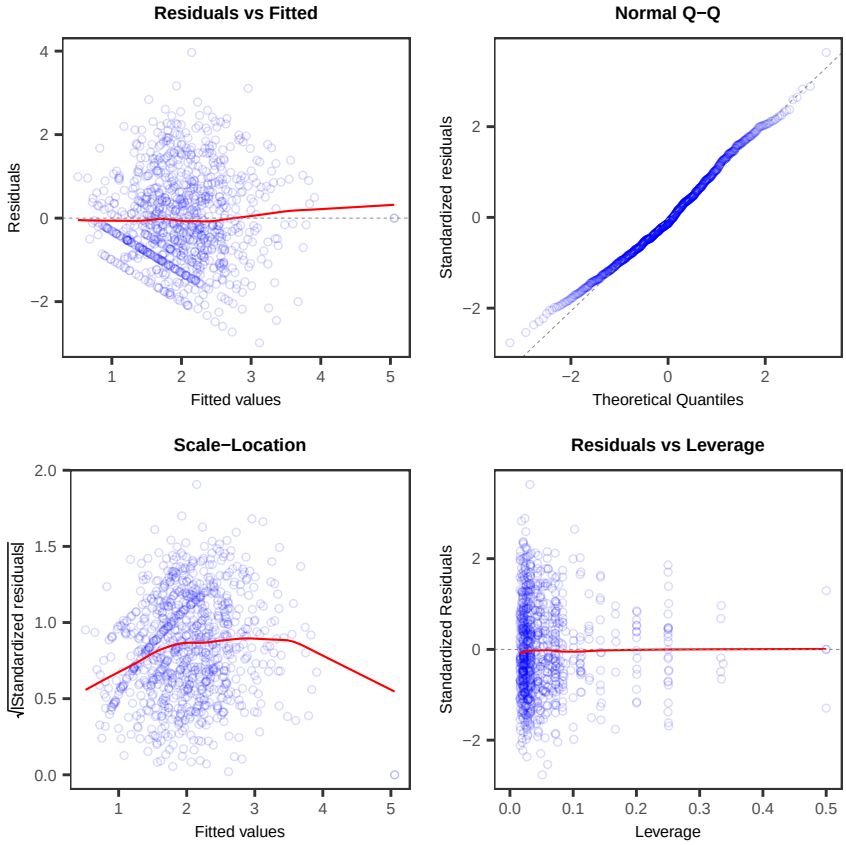
Response variables (average degree and closeness centrality) are weighted by number of vertices (stops) in each county. All variables are transformed by natural logarithm.

Figure 3: Regression assumptions for degree centralities model are reasonably met.



Response variable (average degree centrality) is weighted by the number of vertices (stops) in each county. The response and predictor variables are both transformed by natural logarithm. State fixed effects are included to control for unobserved heterogeneity at the state level.

Figure 4: Closeness centrality regression model is poorly specified.



Response variable (average closeness centrality) is weighted by the number of vertices (stops) in each county. The response and predictor variables are both transformed by natural logarithm. State fixed effects are included to control for unobserved heterogeneity at the state level.