



Improvements in Using Distance to Predict College Enrollment

Taylor Delaney

University of Oklahoma

Dave E. Marcotte

American University

Geographic proximity to college is a widely used instrument for estimating returns to postsecondary education. Most applications rely on geodesic ("crow-flies") distance, which is easy to calculate but a poor proxy for actual travel costs when terrain, water, or congestion intervene. We show that replacing geodesic distance with driving distance or driving time substantially strengthens instrument performance. Using administrative data on Maryland public high school graduates linked to postsecondary records, we find that driving time raises the first-stage F-statistic for years of postsecondary education from 13 to 266, well above the $F > 100$ threshold recommended by Lee et al. (2022) for reliable 2SLS inference. To facilitate adoption, we introduce *drivingtime*, an R package that queries the Google Maps Distance Matrix API to generate driving distances and travel times for researcher-defined origin-destination pairs.

VERSION: October 2026

Suggested citation: Delaney, Taylor, and Dave E. Marcotte. (2026). Improvements in Using Distance to Predict College Enrollment. (EdWorkingPaper: 26-1608). Retrieved from Annenberg Institute at Brown University: <https://doi.org/10.26300/w1c0-8f83>

Improvements in Using Distance to Predict College Enrollment

Taylor Delaney
University of Oklahoma
delaney@ou.edu
&

Dave E. Marcotte
School of Public Affairs
American University
marcotte@american.edu

September 2026

Abstract

Geographic proximity to college is a widely used instrument for estimating returns to postsecondary education. Most applications rely on geodesic ("crow-flies") distance, which is easy to calculate but a poor proxy for actual travel costs when terrain, water, or congestion intervene. We show that replacing geodesic distance with driving distance or driving time substantially strengthens instrument performance. Using administrative data on Maryland public high school graduates linked to postsecondary records, we find that driving time raises the first-stage F-statistic for years of postsecondary education from 13 to 266, well above the $F > 100$ threshold recommended by Lee et al. (2022) for reliable 2SLS inference. To facilitate adoption, we introduce *drivingtime*, an R package that queries the Google Maps Distance Matrix API to generate driving distances and travel times for researcher-defined origin-destination pairs.

Keywords: College access; Proximity; Instrumental variables

This research was conducted using data from the Maryland Longitudinal Data System (MLDS) and with the technical support of the MLDS Center staff. The conclusions of this research do not reflect the opinion of the State of Maryland, MLDS Center, the MLDS Governing Board, or its partner agencies.

The role of college proximity in postsecondary enrollment choices has long been a topic of study. Since at least the mid-20th century, sociologists and education researchers have focused on geography as part of the broader community opportunity structure shaping educational transitions (Hollingshead, 1949). Much of this descriptive work focused on disparities in educational aspirations, college attendance, and degree completion between rural and urban areas (Blau & Duncan, 1967; Sewell et al., 1969). The role of distance to college campuses has been highlighted as a limiting factor in educational attainment for rural youth, from regular P-20 Population Characteristic reports by the U.S. Census Bureau (U.S. Bureau of the Census, 1969) to recent reports on the geography of opportunity and education deserts (Hillman, 2025; Hillman & Weichman, 2016).

By the 1980s and early 1990s, economists and quantitative researchers began formalizing distance as a measurable predictor of college enrollment. This includes work that uses distance to the nearest college as an instrument to estimate the causal effect of education on earnings. Card (1993) pioneered the use of geographic proximity to college as an instrument for years of college education. This approach has been adapted and remains widely used (e.g., Flanagan & Doyle, 2024; Kane & Rouse, 1995; Kling, 2001; Long, 2004; Long & Kurlaender, 2009). While the instrument has faced scrutiny over the exclusion restriction (e.g., Carneiro & Heckman, 2002; Carneiro et al., 2011), it remains a standard and widely accepted strategy in the literature on returns to schooling, particularly when supplemented with controls for local labor market conditions and family background. Indeed, it is used as the canonical example of an instrumental variable in prominent econometrics textbooks (Angrist & Pischke, 2009; Stock & Watson, 2020; Wooldridge, 2010).

The growing use of distance as a predictor of college enrollment has been made possible by advancements in computing power and geographic information systems (GIS) software that enables the measurement of distances between students completing high school and colleges nearby (and far away). Most papers on the topic have used GIS packages to calculate geodesic (aka “as the crow flies”) distances between the coordinates of students’ high schools and the nearest college. Or, researchers might calculate geodesic distances between a county centroid and the nearest 2-year or 4-year college. Statistical packages like “geodist” in Stata and R have made calculating geodesic distances between points easy. Practically, researchers have used exact latitude and longitude between institutions or centroids of a geographic space (e.g., a county) to calculate these distances.

Recent advances in geographic/mapping services have provided additional options for researchers to use information on distance in education research. [Open Route Services](#) and [QGIS](#) provide accessible geographic information systems to enable researchers to calculate driving distance and driving time between two points, rather than the crow-flies distance. Acton et al. (2025) use these to measure driving distance and time to the nearest college in Texas and find that living more than 30 minutes from a community college increases the likelihood of attending a four-year college for White and higher-income students but decreases the likelihood of any college attendance for low-income and Hispanic students.

These tools require GIS software and, for QGIS in particular, routing infrastructure and spatial data management that most education economists do not use. Applying them to administrative data with thousands of origin-destination pairs is not straightforward at the scale most researchers using distance instruments actually use. Google’s Distance Matrix API is a more familiar and widely used resource, but nothing packages it for the specific task researchers

using distance instruments face: reading in a research dataset of student and institution coordinates, batch-querying driving distance and time for every pair, and merging the results back into the dataset in a format ready for estimation.

In this paper, we close that gap. We develop an easy to use R package called [drivingtime](#) that allows users to generate distances between a series of points of origin and destination, as well as the time required to make the journey. The package allows users to input departure times, optimistic or pessimistic estimates of traffic conditions, and the mode of transportation to calculate travel time. The drivingtime package requires users to obtain an application programming interface (API) key and enable the Distance Matrix API in Google Maps.¹ Once a user provides the API key, drivingtime reads in data files with latitude/longitude for high school/college pairs and reads out the file augmented with driving distances and times.

Setting

To illustrate the utility of using data called directly from Google Maps to estimate travel distance and time rather than relying on geodesic distance, we use data from the Maryland Longitudinal Data System (MLDS). MLDS links administrative data on all public high school students in the state to postsecondary data for public and private non-profit colleges in the state. While Maryland is a small state, it is populous and has geographic features that make geodesic distance a poor measure of travel time. The small size of the state and the dispersal of community colleges means that students at only four high schools (out of 216) would need to drive more than 30 minutes to attend the nearest community college. This compares to the 9.4 percent of Texas high school students who would need to drive at least 30 minutes to the nearest community college (Acton et al., 2025).

¹ The authors used Claude to translate sections of code from Stata to R, and to test and build documentation for the completed package.

What makes Maryland an interesting case is not scale, but the diversity of terrain. The state is divided into three distinct sections, the Eastern Shore of the Chesapeake Bay, Central Maryland, and the western part of the state in the Allegheny Mountains. So, two points that are close as the crow flies can require long drives. For example, it is 33 miles from Chestertown High School to Towson University, as the crow flies. But the Chesapeake Bay complicates the drive, making it 86 miles or 1 hour and 46 minutes by road. Further, in the central part of the state, sprawl from the Washington and Baltimore metropolitan areas can lead to divergence between driving distance and driving time. With adaptable API calls to sites like Google Maps, researchers can readily use the measure most sensible to their project.

Implementation

We use three different measures of the distances Maryland high school students would need to travel to attend the nearest public four-year college and in-district community college to predict college enrollment decisions. We model enrollment decision as measured by total years of postsecondary study and by whether a student enrolls in a 4-year college, two of the most common use cases for empirical researchers using distance to instrument for postsecondary enrollment.

We assess the utility of distance measures as instruments in the context of “first-stage” regressions of respective outcomes for the population of all public high school graduates in Maryland between the 2007/08 and 2018/19 school years. All models control for student demographic characteristics, academic achievement as measured by reading/verbal scores on the PSAT, SAT, or ACT (if available), free/reduced price lunch eligibility, whether a student speaks English as a second language, and special education status.

Results

Table 1 presents descriptive statistics for the analytic sample of 394,576 students. The sample is 48% white, 32% Black, 9% Hispanic, and 7% Asian; 51% are female. One-third qualified for free or reduced-price meals (FARMS) in high school, and 7% received special education services. Among all public high school graduates, 47% enrolled in college, 24% in four-year institutions, and 19% earned a bachelor's degree within five years. Students accumulated an average of 1.6 years of postsecondary enrollment.

In Figures 1-3, we illustrate differences in driving distance and time for students whose high schools are approximately 25, 50, and 150 km from four-year colleges, as the crow flies. The first is a distance that a commuting student might reasonably face. The second is a distance that would be an unusually long commuting distance. The third is a distance that would require the student to relocate to attend. In each case, the driving distance exceeds the crow-flies distance. Each figure presents the cumulative density of drive distance and time for those distances. For the nearest schools, the median driving distance is 43 km, and nearly all students would drive less than 50 km to get to a campus 25 km as the crow flies. The distribution of driving times has a relatively long upper tail. Driving distance is a much tighter function of crow-flies distance than is driving time. Within this commuting zone band, 95% of driving distance are likely to be under 50 km, whereas driving time reaches a comparable threshold much more slowly. The steep CDF of driving time illustrates a more concentrated (lower-variance) distribution, compared to the shallow driving time CDF, reflecting a more dispersed (higher-variance) distribution. The same pattern is apparent for the farthest distance (schools about 150 km away), but somewhat less so for schools at the outer reach of the commuting zone (50 km away as the crow flies).

In Table 2, we summarize the strength of measures of distance high school graduates must travel to in-state colleges and universities as predictors of postsecondary enrollment. Each cell presents the F-statistic for tests of significance for distance measures as a predictor of the enrollment measures. The first set reports results for years of postsecondary education, the variable for which distance was first used as an instrument. In each row of the set, we present F-statistics for models using 1) Distance to closest public 4-year college; 2) Distance to closest public 4-year and distance to closest public 2-year college, and 3) Distance to closest public 4-year, distance to closest public 2-year college, and distance to closest public selective or more selective 4-year college, as defined by having an admission rate less than 75%.

In the columns, we present results from models that measure distance using geodesic/“crow-flies” distances to colleges, readily calculable directly from Stata or R packages. Then we report results from models where distance is measured using the drivingtime package that queries Google Maps for driving distances (Column 2) and driving time (Column 3).

For the task of predicting years of postsecondary education, driving distance and time clearly dominate geodesic distance. The first row and column presents results for the original use-case of proximity to 4-year colleges as an instrument for college education (Card, 1993). The F-statistic in this case is 13.03, exceeding the conventional Staiger–Stock (1997) threshold of 10, alleviating some concerns about weak instruments. At this level, any finite-sample bias in 2SLS would be modest and small compared to bias resulting from using college attainment as a regressor in OLS.

However, using travel-relevant measures of distance substantially strengthens first stage predictive power, increasing the F-statistic to 126.45 using driving distance and 265.94 using travel time. These magnitudes far exceed conventional weak-instrument thresholds and meet

more stringent criteria for reliable inference. Lee et al. (2022) show that while the conventional $F > 10$ threshold might be sufficient to limit bias, substantially stronger instruments are required for valid hypothesis testing. They show that standard *t-tests* in 2SLS can exhibit severe size distortions when instruments are only moderately strong. So, applied researchers use $F > 100$ as a second “rule-of-thumb” of instrument strength.

We also present tests of instrument strength for predicting college enrollment in the bottom panel of Table 2. While the F-statistics for geodesic distance measures are stronger for predicting enrollment than for years of postsecondary education, the same pattern holds. Crow-flies distance passes the traditional ($F > 10$) standard, but not the $F > 100$ standard needed to limit threats to inference. Both travel-based distance measures perform much better on this metric.

Conclusions

Distance to college has long served as a workhorse instrument for estimating the returns to postsecondary education. The dominant approach (geodesic distance calculated from GIS packages) is easy to implement but leaves predictive power on the table. Geodesic distance is a poor proxy for the actual cost of getting to campus when terrain, water, or congestion intervenes between a student’s home and the nearest institution. Travel time and driving distance, which reflect the actual access constraints students face, are stronger predictors of enrollment decisions and, accordingly, stronger instruments.

We demonstrate this using administrative data on Maryland public high school graduates linked to postsecondary enrollment records. Maryland is a useful illustration not because it is extreme, but because it is a mid-sized state where moderate geographic complexity is enough to make crow-flies distance a misleading measure of access. Using our drivingtime R package,

which queries the Google Maps Distance Matrix API, driving distance raises the first-stage F-statistic for years of postsecondary education from 13 to 126; driving time raises it to 266. For college enrollment, the improvement is also pronounced. The gains exceed what is needed to satisfy conventional weak-instrument thresholds and cross the $F > 100$ threshold that Lee et al. (2022) recommend for reliable inference from 2SLS t-tests.

The *drivingtime* package provides details and instructions, along with example uses. It is designed to be easy to use: researchers supply latitude and longitude for origin-destination pairs, provide a Google Maps API key, and receive driving distances and travel times in return. Departure times, traffic assumptions, and mode of transportation are all configurable. Researchers who prefer other mapping services can readily build equivalent tools. The key insight is not *drivingtime* itself, but the principle that travel-relevant distance measures dominate geodesic ones wherever geography complicates the path between students and campuses. That is true in most populated regions of the United States and beyond. Researchers using distance as an instrument have a straightforward opportunity to strengthen their designs.

References

- Acton, R., Cortes, K. E., Miller, L., & Morales, C. (2025). Distance to degrees: How college proximity shapes students' enrollment choices and attainment across race-ethnicity and socioeconomic status. *Economics of Education Review*.
<https://doi.org/10.1016/j.econedurev.2025.102724>
- Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Blau, P. M., & Duncan, O. D. (1967). *The American occupational structure*. Wiley.
- Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. NBER Working Paper 4483.
- Carneiro, P., & Heckman, J. J. (2002). The evidence on credit constraints in post-secondary schooling. *Economic Journal*, 112(482), 705–734.
- Carneiro, P., Heckman, J. & Vytlacil, E. (2011). Estimating marginal returns to education. *American Economic Review*, 101(6), 2754–81.
- Flanagan, C. J., & Doyle, W. R. (2024). “Measuring geographic opportunity for higher education in U.S. metropolitan areas,” *Higher Education Policy*.
<https://doi.org/10.1057/s41307-024-00381-0>
- Hillman, N. (2025). Geography of opportunity: Mapping the availability of broad-access institutions. Institute for College Access and Success.
<https://ticas.org/wp-content/uploads/2025/01/TICAS-Market-Share-report-final.pdf>
- Hillman, N., & Weichman, T. (2016). Education Deserts: The continued significance of “Place” in the Twenty-First Century. Viewpoints: Voices from the Field. Washington, DC: American Council on Education.

- Hollingshead, A. B. (1949). *Elmtown's youth: Structure and change in a small town*. Wiley.
- Kane, T. J., & Rouse, C. E. (1995). Labor-market returns to two- and four-year college. *American Economic Review*, 85(3), 600–614.
- Kling, J. R. (2001). Interpreting instrumental variables estimates of the returns to schooling. *Journal of Business & Economic Statistics*, 19(3), 358–364.
- Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter. (2022). "Valid t-Ratio Inference for IV." *American Economic Review*, 112(10), 3260–90.
- Long, B.T. (2004). How have college decisions changed over time? An application of the conditional logistic choice model. *Journal of Econometrics*, 121, 271-296
- Long, B. T., & Kurlaender, M. (2009). Do Community Colleges Provide a Viable Pathway to a Baccalaureate Degree? *Educational Evaluation and Policy Analysis*, 31(1), 30–53
- Rouse, C.E. (1995). Democratization or diversion? The effect of community colleges on educational attainment. *Journal of Business & Economic Statistics*, 13(2), 217–224.
- Sewell, W. H., Haller, A. O., & Portes, A. (1969). The educational and early occupational attainment process. *American Sociological Review*, 34(1), 82–92.
- Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3), 557–586.
- Stock, J. H., & Watson, M. W. (2020). *Introduction to Econometrics* (4th ed.). Pearson.
- U.S. Bureau of the Census. (1969). Current Population Reports: Series P-20, no. 190. "School Enrollment: October 1968 and 1967." Gov. Printing Office: Washington, DC.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. (2nd ed.). MIT Press.

Table 1 Descriptive Statistics

	(1)	(2)
Variable	Mean	Std. Dev.
Female	0.51	0.50
White	0.48	0.50
Black	0.32	0.47
Hispanic	0.09	0.28
Asian	0.07	0.26
FARMS status	0.33	0.47
Special education status	0.07	0.25
Enrolled in college	0.47	0.50
Enroll 4-year	0.24	0.43
Bachelor's degree in 5yrs	0.19	0.39
Total years in college	1.58	2.10
Observations	394,576	

Table 2**F-Statistics on Distance Measures as Predictors of Post-Secondary Enrollment Decisions**

Postsecondary Outcome				
	Predictors	Geodesic Distance	Driving Distance	Driving Time
Years of Postsecondary Education				
	<i>Closest 4-year college</i>	13.03***	126.45***	265.94***
	<i>Closest 4-year and 2-year colleges</i>	46.65***	100.09***	156.38***
	<i>Fully interacted distance and student chars.</i>	140.17***	412.50***	388.40***
Enroll in Four-Year College? (0/1)				
	<i>Closest 4-year college</i>	92.76***	731.33***	1052.79***
	<i>Closest 4-year and 2-year colleges</i>	81.53***	373.57***	529.38***
	<i>Fully interacted distance and student chars.</i>	40.68***	483.81***	484.96***

*** p-value < 0.01

Each cell presents F-statistic and significance of associated p-value for test of joint significance of distance measures for OLS regressions models of outcomes for the population of all public high school graduates in Maryland between 2000/08 and 2018/19. All models control for student demographic characteristics, academic achievement as measured by scores on the PSAT, SAT or ACT (if available), free/reduced price lunch eligibility, whether a student speaks English as a second language, and special education status. See text for description of distance measures.

Figure 1

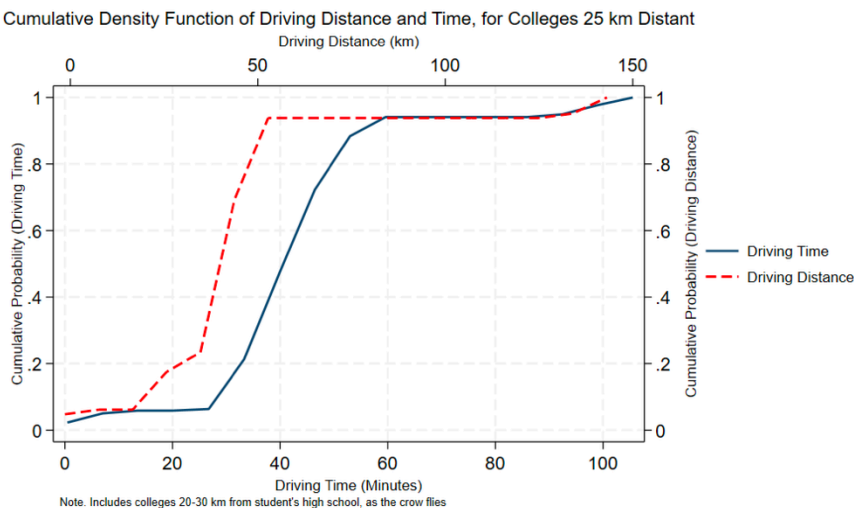


Figure 2

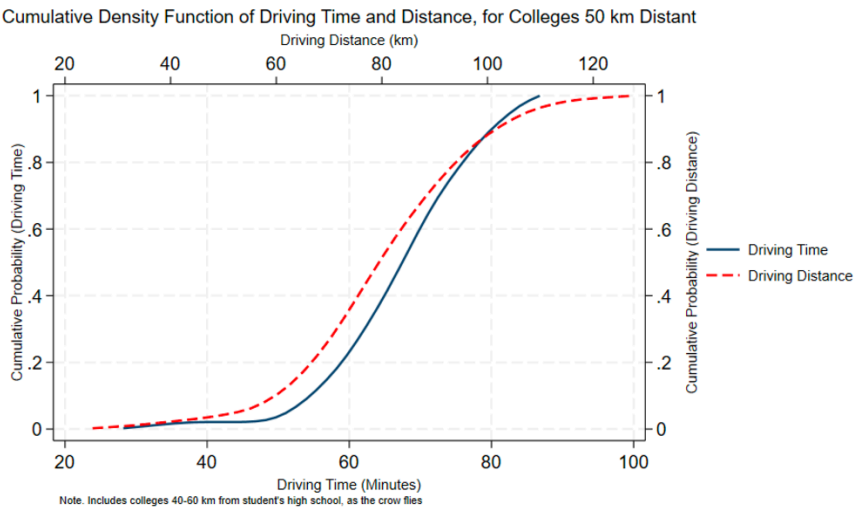


Figure 3

