California Registered Nurse Shortages: Does Population Predict Gap Percentage?*

Andrew Young

2025-09-23

Abstract

In this paper, I tested whether population size predicts the projected registered nurse (RN) shortage "gap percent" across California using the 2025 Registered Nurse Shortage Areas data set provided by CA.gov and the Department of Health Care Access and Information. The simple OLS regression of gap percent on estimated population I rna found a positive association, for each additional increase in 1,000,000 residents in a county there was about a +10.4 percentage point improvement in the gap percent of RNs, while the model explained about 8% of the variation. My post regression diagnostics found a roughly linear relationship with heavy tails. Because the data I am working with are projections and the effect is modest, I concluded that population is a weakly positively correlated to gap percentage, and that planning for this issue could use additional drivers such as wages, hospital capacity, and education, rather than just population alone.

Introduction

Nurses are a critically important part of California's health-care workforce, and staffing shortages have gotten increased attention from the public since the COVID-19 pandemic. In an effort to guide planning, California has identified something they call Registered Nurse Shortage Areas (RNSAs) and reports a projected "gap percentage" that compares nurse supply to demand.

My paper aims to better understand this information, and asks the question, "Can population size predict the projected RN shortage gap percentage across California's counties?" Communities with greater population numbers might benefit from a larger pool of people from training programs, bigger hospital systems, and deeper labor markets while smaller or rural areas might face recruitment and retention challenges.

To address this, I analyzed the 2025 California RNSA dataset provided by CA.gov and the HCAI and fit a simple linear regression with RN shortage gap percentage as the outcome and estimated population as the predictor. The results show a positive association between population and the

*Project repository available at: https://github.com/AndrewBruceYoung/MATH261A-Project-1-Nurse-Shortage.

gap percentage, but the effect is modest and leaves substantial unexplained variation, indicating other factors are likely more influential.

My analysis found a light positive association: for every additional one million residents, the gap increases by about +10.4 percentage points, indicating that more populous areas tend to have less severe shortages on average. However, the model explains only about 8% of the variation, so population by itself seems to be an incomplete predictor of local RN needs. This matters for policy because it suggests that interventions for this problem can look to population as a lightly guiding metric, there are more variables that can be taking effect on this gap. The rest of the paper proceeds as follows: the Data section describes variables and cleaning, Methods states the regression model and assumptions, Results presents estimates and the fitted plot, Discussion interprets findings and limitations, and Conclusion summarizes implications and next steps.

Data

In the data set that I working with, the observational units are California counties. The raw 2025 Registered Nurse Shortage Areas file contains 66 rows, and after some basic cleaning and removal of one record with missing values, the data set ended up with 65 usable rows. Each row represents one county area in 2025.

Variables used:

Est. Population (renamed to est_population): estimated 2025 population in county.

Gap % (renamed to gap_pct): the RN shortage gap percentage, defined by this equation:

$((supply - demand)/demand) * 100$

Negative values mean there is a shortage, positive values mean there is a surplus.

Supply: numerical amount designating how many RNs a county has, designated by the HCAI

Demand: numerical amount designating how much need there are for RNs in a county, designated by the HCAI

County: name of county (or LA Service Planning Area (SPA)) where information was gathered from

Only est_population and gap_pct enter the regression, the others variables listed here provide context and possible checks.

Collection and transformations:

As to how this data set was collected, it was not, this data set is a state projection for 2025, produced from an RN supply–demand model by the Department of Health Care Access and Information (HCAI). This means that my results speak on what is projected by the HCAI to occur this year, and because of this there might be some discrepancies to what happens in reality. I displayed my work in transforming the raw data into something usable for analysis in the steps below:

1. Standardized column names so they were easier to work with

2. Removed empty extra columns present in the CSV file
3. Changed the strings that were present in the gap_percent column, such as "-15.1%", to numeric

Key characteristics:

The estimated population of each county varies greatly, from some with about 1000 residents to several million in the largest. The Gap_pct variable is similar, spanning from -80% in some counties to +100% in others. This wide spread explains why a single predictor may only partially account for variation in shortage severity.

Below I attached a simple table containing some summary statistics of estimated population and gap percentage.

Table 1: Sample size and ranges for population and RN gap percentage.

n_areas	pop_min	pop_median	pop_max	gap_min_pct	gap_median	gap_max_pct
65	1163	253823	3323902	-81.66	-16.6	95.3

Methods

I studied the association between population size and projected RN shortage severity using an ordinary least squares simple linear regression. Let:

Y_i = RN shortage gap percentage for area i

X_i = estimated 2025 population for area i

β_0 = intercept

β_1 = slope (expected change in gap percentage for a one person increase in population)

ϵ_i = random error for area i

With all of these, the model would be:

$Y_i = \beta_0 + \beta_1(X_i) + \epsilon_i$

Since we are dealing with potentially millions of people, for interpretation I also report β_1 per 1,000,000 people by multiplying the estimated slope by 10^6.

Justification and modeling decisions:

I felt that a bivariate OLS model would be appropriate for this projects goal of demonstrating and interpreting a simple linear relationship. I used population as the explanatory variable because it is an incredibly policy relevant factor that could possibly relate to nurse availability and demand. Because the population spans such different and large values, I presented the slope per 1,000,000 people for readability. I chose not to remove observations that were deemed outliers as I felt that this preserves reproducibility and avoids some post-hoc problems.

Assumptions:

All the standard linear model assumptions apply here: Linearity, independence, homoskedasticity, and normality of errors. On top of those standards, I assume that any errors that come from the fact that the data is projected is small and roughly unbiased. I also included LA SPAs as their own areas due to their population size, and by doing this I am assuming the areas are independent from each other.

Diagnostics and validation steps:

After making my model, I assessed assumptions using two diagnostic plots: Residuals vs Fitted (which examines linearity and equal variance) and Normal Q-Q (which examines normality) in an effort to look deeper into how this model holds up with the assumptions of simple linear regression that we want to abide by.

Software and implementation.

All analyses were performed in R using readr and janitor for import/cleaning, dplyr for wrangling, ggplot2 for the scatter plot with a fitted line, broom for tidy model summaries, and scales for axis labels.

Results and Discussion

Model fit and key estimates:

After fitting an OLS regression of RN shortage gap percentage on population, I ended up with this plot and regression line:

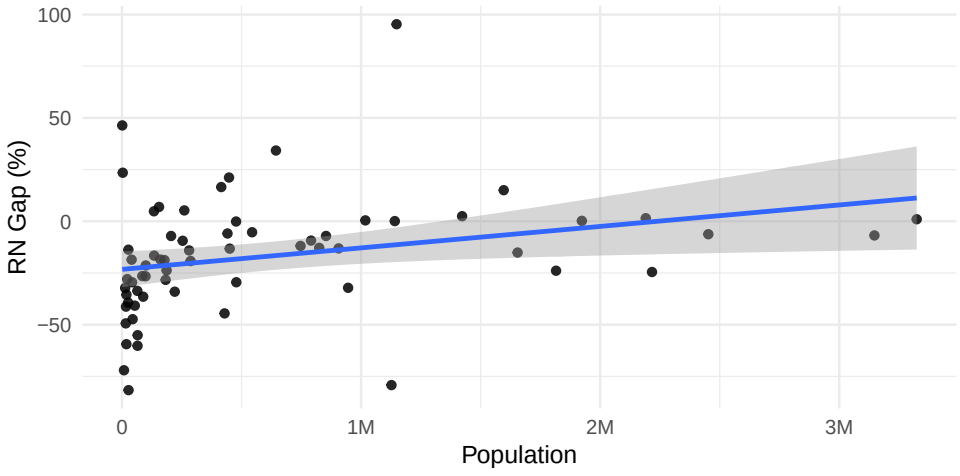

Figure 1: RN Shortage Gap vs Population with OLS fit and 95% band

The estimated slope ended up being statistically different from zero at the 5% level, with a p-value = .0216, with +10.4 percentage points per +1,000,000 people. The intercept is about -23.3 and is not too meaningful in this case as it relates to a population of 0. The model fit was modest

with an R^2 of 0.081 and an Adj. R^2 of 0.06, and the residual standard error is roughly 27 percent. The fitted scatterplot shows a slight upward trend with wide spread that increases the larger the population

Plain-English interpretation:

Essentially, my results found modest evidence that areas with larger populations tend to have less severe RN shortages on average. According to my output, increasing population by one million residents is associated with an improvement of about 10 percentage points in the gap. However, the overall effect is pretty small and leaves most of the variation unexplained, so population by itself should not be used to predict local RN needs.

Connection to the research question:

My original question was whether population size predicts the projected RN shortage gap. While the answer that i got was yes, it is a weak yes. Population alone seems to be correlated, however likely is not a sufficient driver for this output. This implies that planning and policy should not rely on solely population alone, there are probably plenty of other factors that matter in planning for this issue.

Diagnostics:

As stated before, two diagnostic checks were used. My Residuals vs Fitted plot ended up showing no strong curvature, suggesting that using a linear form to represent their relationship is reasonable. However, there is a slight increase in residual spread at higher fitted values. This mild shape that is being formed points to light heteroskedasticity, which mostly affects standard errors rather than the slope itself. You can see this plot below:

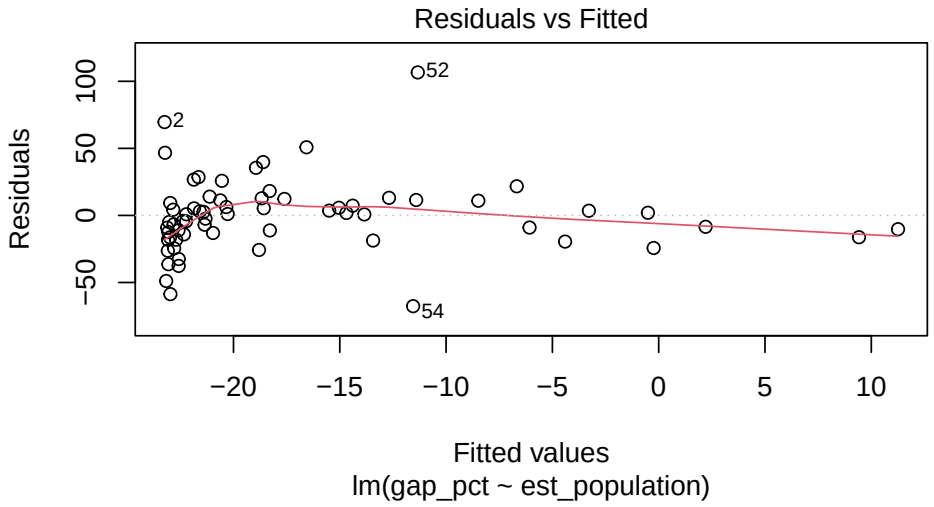

Figure 2: Figure 2. Diagnostics: Residuals vs Fitted.

My Normal Q–Q plot is close to the reference line with deviations at both tails. Overall though, these diagnostics are displaying results that indicate they are adequate support for a simple linear model. The departures from the line at both tails likely reflects that there are a few areas with unusually large positive or negative residuals. You can see this plot below:

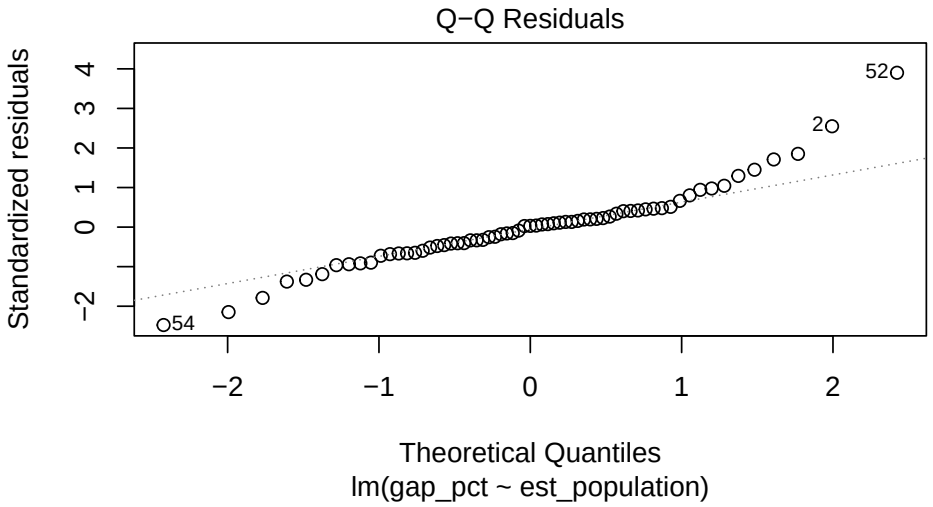

Figure 3: Figure 3. Diagnostics: Normal Q–Q.

Potential limitations and pitfalls.

There was a couple limitations that I experienced while working with this data set, with the initial one I faced being that the data are projections for 2025, not actually collected counts. Another limitation is that for an issue like this, a single predictor almost certainly cannot fully explain what goes on behind these scenes for supply and demand of RNs, as there are many other factors that can go into affecting these needs (wages, amount of nearby hospitals, education). Finally, the other limitation this project is that if for whatever reason the LA SPAs are not independent from one another, this could have affected my results and output.

Possible extensions and improvements:

To further examine robustness and improve fit, I decided on two possible extensions that could be done:

1. Re-fit with log(population) as the predictor to address possible curvature that may come from very large areas
2. A counties-only model that excludes LA SPAs to examine the effect they may have when treated as a single county
3. Run these same tests again at the end of 2025 once these data points have actually been collected and compare to these results from the predicted values.

References

Department of Health Care Access and Information (HCAI). (2025). Registered Nurse Shortage Areas in California. California Open Data Portal. https://data.ca.gov/dataset/registered-nurse-

shortage-areas-in-california

RDocumentation. (2025). RDocumentation. https://www.rdocumentation.org/

Stack Overflow. (2025). Stack Overflow — Questions. https://stackoverflow.com/questions

R Core Team (2025). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.