James Greene

Tristan Sahagian

MA517

December 2025

**Math Foundations - MassHealth County Level Per Capita Expenditure Prediction**



## Introduction

The data set we chose is a real world example of historical data collected about county-wide expenditures across the entire United States. The metrics covered in the set are aggregate per capita expenditures, average risk factors (clinical/health information), average demographic risk factors and person-years. For each of these measures, the data set is further stratified into insurance enrollment types (ESRD, DIS, AGDU, ADND).

For this analysis, we chose to focus on the disability (DIS) enrollment type and its relevant data. Another approach could be to aggregate across all enrollment types and sum expenditures. We

chose this approach to investigate how effective these risk factors would be as predictors of per capita expenditure on an individual enrollment group at the county level.

For our model, the dependent vector is the collection of per capita expenditures for each county across the United States.

## Concepts

First we will define what equations, variables, and linear algebraic concepts we utilized through this least squares analysis. See below for the full matrix notation and variable definition for the basic linear regression model we are utilizing.

$$y = X\beta + \epsilon$$

*equation 1: regression equation matrix notation*

$y$ → vector of observed dependent variables (per-capita expenditures)
$X$ → matrix of features (risk scores, demo scores)
$\beta$ → vector of coefficients to be estimated
$\epsilon$ → residual error vector

When applied directly to this data set. This will look like equation 2 below.

$$PER\_CAPITA\_EXP_i = \beta_0 + \beta_1(AVG\_RISK\_SCOREi) + \beta_2(AVG\_DEMOG\_SCOREi) + \varepsilon_i$$

After the model is fitted using the observed feature(s) and observed dependent values (per capita expenditures). The model's efficacy can be judged by measuring the residual values between the predicted y ($\hat{y}$) from the regression line and the observed values.

$$r = y - \hat{y}$$

*equation 2: residual vector equation*

This is another way to represent the residual variable. This shows how the predicted y value is equivalent to the feature matrix multiplied by the minimum beta calculation.

$$r = y - X\hat{\beta}$$

equation 3: alternative form residual vector equation

Using these residual values, the sum of squared errors can be found. From this value, we can calculate the mean-squared error (MSE) and the $R^2$ value. The mean-squared error is exactly what its name implies; the average squared difference between the estimated values and the observed values. It is a metric of determining the quality of an estimator.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2$$

*Equation 8: Mean Squared Error*

The R-squared value ($R^2$) is also known as the *coefficient of determination*. The R-squared is a value between 0 and 1; a percentage. This makes it a valuable tool for

evaluating the strength of the regression's fit because it does not require any additional context for interpretation. The R-squared measures the percentage of variance in the dependent variable that can be explained by the model's predicted independent variable. In general, it measures how well the model fits the data. A value close to 1 would indicate a near-perfect fit and a value near 0 would indicate an extremely poor fit. It can be calculated by subtracting the proportion of the sum of squared error over the total sum of squares from 1 (100%). See equations 4,5,6. A list of many relevant equations are shown below.

$$R^2 = 1 - \frac{SSE}{SST}$$

*Equation 4: R-squared*

The R-squared calculation can be seen in equation 4. The relevant components of this calculation are also seen in equation 5, 6, 7 (sum of squares total and sum of squares regression).

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})$$

*Equation 5: Total Sum of Squares*

---

$$SSE_{OLS} = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

*Equation 6: Sum of Squared Errors (Ordinary Least Square)*

$$SSE_{WLS} = \sum_{i=1}^{n} w_i (y_i - \hat{y})^2$$

*Equation 7: Sum of Squared Errors (Weighted Least Square)*

Another concept from this course that is relevant to our analysis is the orthogonality and how it relates to finding the minimum coefficients in least-squares regression. In linear regression, the optimal coefficient vector $\beta$ is found by minimizing the squared residual. Geometrically, this minimum vector must be orthogonal to the column space of the feature matrix (shown mathematically in equation 9).

$$X^T(y - X\beta) = 0$$

*Equation 9: Orthogonality for Minimum Coefficient Vector*

Contextually, this means that the orthogonality condition guarantees the closest possible linear approximation of the projected per-capita expenses generated by the county-level risk and demographic variables ($X$ matrix). Specifically for the DIS-enrollment group, this means that no additional patterns regarding per-capita expenditures can be explained by a linear combination of the two risk factors we used as features. This ties in well with the concept of *weighted* least squares. This alters the orthogonality condition slightly; factoring in weights in the form of person-years to the feature matrix (equation 9).

$$X^T W(y - X\beta) = 0$$

*Equation 10: Orthogonality for Weighted Minimum Coefficient Vector*

## Analysis Process

### 1. Data Preparation

The first step in the data science process is always data collection. Since we had our data set chosen already, the natural next step was data cleaning. All data preparation, manipulation and analysis was done in a Python notebook using Pandas, NumPy, MatPlotLib and Scikit-Learn. After importing the CMS county-level expenditure data (2020-2024), we cleaned the data with the methods listed below:

- Remove samples with null and missing data
- Remove samples with suppressed values (non-numeric values such as '*')
- Convert all relevant columns to numeric data format
- Filtered by only Disability (DIS) enrollment type - our focus for this analysis
- Values were already normalized so normalization/standardization methods were not needed

### 2. Defining Variables

We then defined our X and y. The features (X; independent variables) are some combination of the average risk score for the disability enrollment type (AVG_RISK_SCORE_DIS) and the average demographic score for the disability enrollment type (AVG_DEMO_SCORE_DIS). Again, these averages are aggregates per county. Our label (y; dependent variable) is the per capita expenditure for the disability enrollment type (PER_CAPITA_DIS). This is the average annual per-capita expenditure for disabled enrollees per county. A third variable type is also utilized only for the WLS aspect of our analysis. Person-years (PERSON_YEARS_DIS) is the column used to define the weights for our model. This value reflects the aggregate enrollee exposure. Essentially, this value gives a normalized eligibility time. It applies more weight to counties with larger populations and more overall coverage time.

### 3. Regression Analysis

We began by running three variants of OLS. These models tested each feature separately as predictors and then in combination.

1. Risk-only
2. Demographic-only
3. Two-factor (Risk + Demographic)

Then, we ran the two-feature regression but with weights (weighted least squares) - using the person years values as weights. For each variation of the model, we used linear regression and fitted it to predict per-capita expenditure. We computed the model parameters; coefficients and intercept ($\beta_0, \beta_1, \beta_2$). We generated the predicted values ($\hat{y}$) and their residuals ($r_i$). Then, the model's performance metrics could be calculated; mean square error and coefficient of determination. Finally, the regression was plotted. First, we displayed a scatter plot for the actual vs. predicted dependent values. Then, we plotted the regression line of best fit.

## Results

The highest performing model was the 2024 2-feature WLS model. Based on the R-squared score, it produced the best fit with a value of 0.53. This means it explained just over half of the variation in county-level per capita expenditures for DIS enrollees. This aligns well with the intuition behind our use of weights: counties with larger populations (higher person-years) should contribute more to the coefficient estimation calculation.

The 2024 OLS version of the 2-feature model performed second best, resulting in a slightly lower R-squared value of 0.43. This drop in regression fit directly enforces our assumptions

about the effect of weighing the coefficient vector. The weights control for smaller, potentially more volatile counties; preventing these counties from skewing the overall fit.

Unsurprisingly, the single-feature models performed worse than the two-feature versions. The 2024 risk-only model reached an R-squared of 0.35, while the demographic-only model explained almost none of the variation with an R-squared of 0.04. This lines up with what we saw in the exploratory analysis: risk scores carry meaningful clinical information, whereas demographic averages vary less across counties and contribute far less on their own. Using both features allows the model to capture multiple dimensions of expenditure variation, which naturally improves the fit compared to relying on a single predictor.

One unexpected result was that the models built on the 2020–2024 merged averages performed noticeably worse than the single-year 2024 models. At first glance, averaging multiple years of data seems like it should reduce noise and lead to a more stable signal. But in retrospect, the lower accuracy makes sense given the structure of the dataset. Earlier years, especially 2020–2021, showed different expenditure patterns and weaker linear relationships between risk and spending. By pooling these years together, the model is forced to fit across shifts in both cost levels and risk distributions that aren't well captured by a simple linear structure. As a result, the combined dataset introduces more variability than it removes, which ultimately reduces the model's overall fit.

### Real-World Context

Although the R-squared values from our models are reasonable, they also highlight the inherent complexity of predicting county-level healthcare expenditures. Per-capita spending for DIS enrollees is influenced by far more than the two features we used. Clinical risk and demographic

averages give us a good starting point, but they cannot capture the full variation across counties. Factors such as differences in healthcare infrastructure, provider availability, socioeconomic conditions, regional pricing variations, and local policy changes all play a role in determining how much a county spends on its disabled population. With so many underlying drivers, it is expected that a simple two-feature linear model will only explain a portion of the variance.

To give some additional context to our results, we can compare our results to the CMS's official model and reports. Their own concurrent analysis (same period predictors and expenditures) for PY2025 reports an R-squared score of 0.4911 and their predictive model (using forecasted predictors) reported an R-squared value of only 0.1245.

https://www.cms.gov/aco-reach-kcc-py2025-risk-adjust-paper?utm_source=chatgpt.com

This information provides valuable insight into how effective these predictors can be. It also illustrates that our results align well with official health care estimates; a prime example of how real-world data is much more chaotic and harder to predict than theoretical data sets.

**Potential Analysis Improvements**

One potential area for improvement would be to expand the feature set by combining this dataset with other publicly available county-level sources. For example, socioeconomic indicators from the Census Bureau, healthcare provider counts from national registries, or cost-of-care indices from CMS would all help the model capture more of the structural variation across counties. Adding additional relevant predictors would likely raise the R-squared values by giving the model more information to explain differences in spending patterns.

Another improvement would be to explore more flexible modeling approaches beyond standard linear regression. Techniques such as regularized regression (ridge or lasso), tree-based models, or gradient boosting could help uncover nonlinear relationships and interactions that a simple linear model cannot represent. These methods may offer better predictive performance, especially when multiple contributing factors interact in ways that are not strictly additive or linear.

From a business and policy standpoint, improving this model has direct implications. A more accurate expenditure model could help agencies better understand which counties face higher cost pressures and why. This would support more targeted resource allocation, more accurate budgeting, and potentially more equitable reimbursement structures. As the model becomes richer and more explanatory, it could also guide future CMS policy adjustments by highlighting which county-level characteristics are most strongly tied to higher or lower expenditures.

## Business and Policy Implications

The findings from this analysis offer several important takeaways for both business decision-makers and policymakers working in healthcare. Even with a limited feature set, the models captured a meaningful amount of variation in county-level DIS expenditures. An important note is we utilized a *real historical data set*, meaning it's less likely to fit cleanly into any type of model. This exercise demonstrates the value of data-driven cost modeling and highlights how structured statistical approaches can help organizations better understand the drivers of healthcare spending. For agencies and insurers responsible for budgeting, forecasting, or allocating funds, having a clear view of which county characteristics matter most—and which contribute very little—provides a more solid basis for planning.

The strong performance of the risk score variable suggests that clinical risk remains one of the most influential predictors of spending differences. This supports continued investment in accurate and up-to-date risk measurement systems, as well as the use of clinical indicators when setting reimbursement rates or evaluating population health needs. In contrast, the demographic variable showed limited predictive value in isolation. For policymakers, this points to a potential need for richer, more granular demographic or socioeconomic metrics if demographic factors are expected to play a role in expenditure models.

Overall, this type of analysis helps organizations and policymakers make more informed decisions about resource allocation, population health strategy, and future data collection efforts. By identifying which variables meaningfully explain expenditure patterns, the model provides a foundation for improving reimbursement structures, guiding targeted interventions, and supporting long-term financial planning across diverse counties.