Analyzing Family Earnings and Medical Expenditures

Datasci 203: Building Descriptive Models

Jonathan Ho, Tom Mayer, Eduardo Jose Villasenor

2024-07-22

Contents

1 Importance and Context

In the United States, employment status plays a crucial role in determining citizens' access to private health insurance, making the ability to work an essential aspect of both financial and physical well-being. This paper investigates the financial relationship between family earnings and medical expenditures, aiming to shed light on how earning power affects healthcare affordability and access. Specifically, the analysis addresses the following research question:

> *How may we describe the relationship between a family's total earnings with their total medical expenditure?*

This study contributes to the broader discussion on healthcare affordability and equitable access in America. By examining this relationship, we aim to provide insights that can assist policymakers in addressing the financial challenges families face and in designing interventions that promote both economic stability and health equity. Additionally, future research can expand on this analysis by investigating the effects of various types of employment and insurance coverage on medical expenditures.

2 Data and Methodology

Our analysis utilizes data from the Current Population Survey (CPS), the official source of U.S. Government statistics on employment and unemployment. Specifically, we employed the Annual Social and Economic (ASEC) Supplement, which provides comprehensive monthly labor force data. The CPS sample is based on the civilian non-institutional population of the United States, encompassing approximately 826 sample areas that include 1,328 counties and independent cities, with coverage across all states and the District of Columbia. For this study, we analyzed the cross-sectional survey data for the year 2023.

This dataset includes 65,767 families, each with non-null values for total family earnings and total medical expenditure. We divided the data into an exploratory set comprising approximately 30% of all observations (19,730 families) and a confirmation set with the remaining 70% of observations (46,037 families). The exploratory set was used to investigate the data and develop our model specifications, while the confirmation set was reserved for generating the final plots and tables presented in this report.

We operationalized the concept of a family's earning power using total family earnings, E, which includes income from wages, salaries, farm self-employment, and own business self-employment. While we considered using total family income—which encompasses various other income sources such as disability income and pensions—we decided against it to avoid significant variability and potential confounding factors from non-labor income sources. By focusing on total family earnings, we can more clearly observe the relationship between a family's earnings and medical expenditures without distraction. Negative earnings values were excluded, as they would require separate consideration for special cases such as debt. We also operationalized the concept of a family's total medical expenditure, M, as the total amount of money the family spent on medical services.

Given that our sample size for both the exploration set and confirmation set is larger than 100, we assess the large-sample assumptions for Ordinary Least Squares (OLS) regression.

1: I.I.D data. Based on the CPS sample characteristics, there is potential geographical clustering between families that are close to each other, which might weaken I.I.D. The survey also includes special samples such

as Hispanic households aimed at improving estimates for specific sub-populations. These additions are not random, which might introduce potential biases and dependencies. There were also non-responses, which may not be random and could introduce biases, if unresponsive households have different characteristics than responsive ones. Still, the large sample size and the extensive geographic coverage provide a broad and diverse data set, which can help balance out local dependencies and clustering effects. While the data might not be perfectly I.I.D, we can likely still look at overall trends and patterns.

2: A unique BLP must exist. First, a BLP must exist. This means the covariance between the E and M needs to be defined and finite, with no heavy tails in either distribution. Looking at Figure 1 below, where the variables are not transformed, we note that even there is a clear right skew in both E and M, which indicates that might be a possibility of heavy tails in both. This is also reflected in the outliers in the scatterplot below. Thus, a BLP might exist, but only after transforming the variables to mitigate the heavy tails. Secondly, this BLP must be unique. Since we only have one explanatory variable, there is no perfect collinearity, and the BLP thus is unique. With these two assumptions, we proceed to build our models.

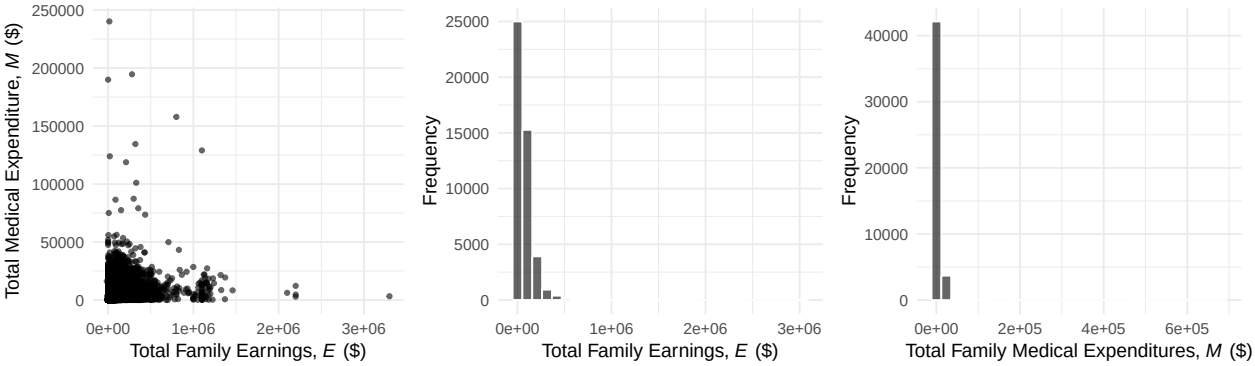

Figure 1: Right skew and potential heavy tails in both Family Earnings E and Medical Expenditures M.

3 Model Results

The two models are seen in Figure 2, and the statistical significance for coefficients and R^2 values are presented in Table 1. Robust standard errors were used for the analysis. For Model 1, $M = \beta_0 + \beta_1 E$, we found a statistically significant overall relationship between earnings and medical expenditures ($p<0.001$), although the adjusted R^2 value is low at 0.050. For Model 2, to address potential heavy tails, we employed a log-log transformation, which also allows us to analyze the relationship in percentage changes. We added one to all earnings and medical expenditures to enable logarithmic transformation, which is a minor adjustment due to typically large values of E and M and still allows us to describe the relationship between both variables. We thus obtain $\ln(M+1) = \beta_0 + \beta_1 \ln(E+1)$, and continued to find strong statistical significance for the coefficient of the explanatory variable ($p<0.001$), with an adjusted R^2 value at 0.087.

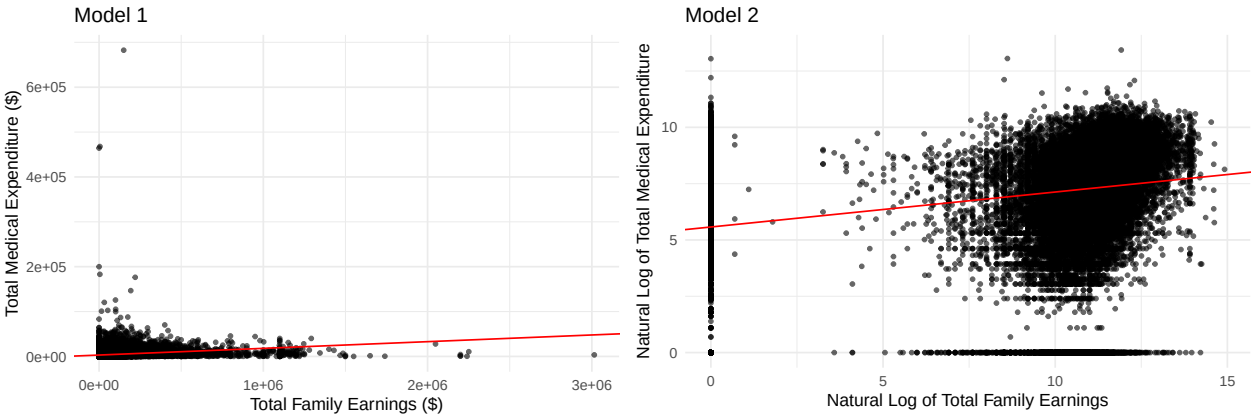

Figure 2: Describing the relationship between Medical Expenditures M and Family Earnings E.

Table 1: Linear Regression Models

	Dependent variable:	
	Total Medical Expenditures ($)	Natural Log of Total Medical Expenditures
	(1)	(2)
Total Family Earnings ($)	0.015^{***} (0.001)	
Natural Log of Total Family Earnings		0.155^{***} (0.003)
Constant	$3,055.884^{***}$ (48.715)	5.576^{***} (0.026)
Observations	46,037	45,995
R^2	0.050	0.087
Adjusted R^2	0.050	0.087
Residual Std. Error	7,337.470 (df = 46035)	2.475 (df = 45993)
F Statistic	$2,410.995^{***}$ (df = 1; 46035)	$4,391.613^{***}$ (df = 1; 45993)
Note:		*p<0.05; **p<0.01; ***p<0.001

We note that Model 2 has the equation $\ln(M+1) = 5.576 + 0.155\ln(E+1)$. Practically, this indicates that for a 1% increase in total family earnings E, we expect to see an increase in total family medical expenditures M of about 0.155%. This relationship suggests that higher earnings correlate with higher medical spending: in a family with $100,000 annual earnings, a 1% increase ($1000) could lead to an approximate $3 increase in medical expenditures, which can be practically significant as the earnings increase and when expenditures are aggregated across millions of families.

4 Discussion

This study found evidence through OLS regression that a family's total earnings and total medical expenditure are positively correlated, with a practically significant effect. When many families earn more, the substantial increase in aggregate medical expenditures benefits healthcare providers and insurers. Policymakers can use these insights to enhance earnings and healthcare access, while insurance companies might adjust coverage options to better meet the needs of families of different income levels.

5 Appendix

1. Link to Data Source: https://www.census.gov/data/datasets/time-series/demo/cps/cps-asec.html

2. List of Model Specifications you Tried:

Equation	Description
$M = \beta_0 + \beta_1 E$	Initial model with untransformed variables to establish a baseline understanding. Heartened to see strong statistical significance in the coefficient.
$\sqrt[3]{M} = \beta_0 + \beta_1 \sqrt[3]{E}$	Tried to take a cube root to spread out the values nearer to the origin and saw some success. However we learned that it is very difficult to explain the results.
$\ln(M+1) = \beta_0 + \beta_1 \ln(E+1)$	Took natural log on both variables to mitigate heavy tails, and picked this model as it also allwoed us to explain our results in a meaningful way, as seen in the report.

Table 2: Summary of Model Specifications

3. A Residuals-vs-Fitted-values Plot