Solving the endogeneity of the determinant(s) of health expenditure

Jiyeon Kim (12470236), Winston Bartle (12768928), Younjoo Mo (12475440), Zhenning Zhang (12807109)
Course: Empirical Project
Group 17

Abstract

Using US data, we analyzed the determinants of health expenditure and explored a solution to the endogeneity of one of the determinants, namely years of schooling. Our findings revealed that among other significant factors, insurance is the most important determinant of health expenditure. However, the findings of this study may be subjected to omitted variable bias caused by a lack of vaccination data, for example.

Keywords: IV, endogenous, expenditure, schooling

Introduction

A major concern in the United States today is how the country is going to recover from the impact of the COVID-19 pandemic. The US-Dollar is weakening, causing the cost of living to go up. Many are struggling with their physical and mental health, some also finding it hard to make current and future health-related payments. One of the initiatives the government is currently implementing is quantitative easing; however, this stream of dollars only benefit banks and not regular citizens directly. A more practical approach would be to distribute a customized stimulus package for each household. So the key question would be: "What are the determinants of healthcare expenditure?" With this information, the US government can more accurately give out personal stimulus help without spending excessively, which could devalue its currency even more in the future.

Theoretical background and method

Many empirical challenges arise when it comes to modelling people's health expenditures. For example, the simultaneity effect between private insurance coverage and health expenditure explaining the phenomenon of adverse selection and moral hazard as covered by Shen (2013). We however do not strongly believe that insurance coverage is endogenous because it could be that individuals with high health-risks do not get approved for insurance to begin with, for instance. We believe a more appropriate endogenous variable is years of education. Not only is it a product of measurement error since our data is integer when it should be continuous, it may also be correlated to unobserved factors such as doctor's bias. Some doctors may prescribe unnecessary drugs but someone who has received adequate education is more likely to be critical in making health-related payments.

Data

We considered the subsample of adult male ages between 18 to 85 from the Medical Expenditure Panel Survey (MEPS) 2011 Cohort. We focus on the male population because females may have more needs for healthcare such as in cases of pregnancy. We exclude individuals who did not report an answer to any of our variables. The final sample consists of 7325 individuals.

The variables used to explain expenditure are similar to those in Shen's (2013) study, but we used data that are more recent and detailed. We also had education instead of insurance as the endogenous variable (the latter was assumed to be exogenous). We use an indicator for higher-pay jobs (financial operations, professional, office, business) as well as an indicator for private-sector jobs as our instruments for years of education. This is because individuals with private-sector jobs usually receive more years of education and have to remain competitive to secure the job, whereas public-sector jobs are less competitive. Generally, those who have a high-paying job usually are people who have done more years of schooling. Some summary statistics are provided in the Appendix.

Results

Health expenditure (Mean = 2735.86, Median = 295.00, SD = 27225.80) has a very skewed distribution ($\tilde{\mu}_3 = 74.75$). Thus, we introduced a log transformation. Our solution to the mathematical limitations of the $log(0)$ is to use $log(expend + 1)$. We estimated our model via a two stage least square system describing log of expenditure plus one (y_i) and years of schooling (S_i):

$$y_i = X_i^{'}\gamma + \beta S_i + u_i \tag{1}$$

$$S_i = Z_i^{'}\alpha + v_i \tag{2}$$

Our assumption is that the error terms of the equations are normally distributed. Table 1 shows a shorten version of the

coefficient estimates of the OLS and IV method. The OLS estimated marginal effect of schooling is significant at 13.9% ($t = 10.22, p < .001$), which is evidently smaller than when it is IV estimated. Note that this IV is achieved by including two dummy variables. These two variables, namely private sector ($p < .001$) and higher-pay jobs ($p < .001$), were significant in the reduced forms of Equation 2 (see Appendix), suggesting strong endogeneity of schooling with respect to expenditure. This conclusion is confirmed by the Hausman (1978) chi-squared test ($\chi^2 = 7.54, p = .006$). Furthermore, we conducted a test excluding the two potential instruments to check their strengths as proposed by Bound et al. (1993). The test yielded a F statistics of 584.20. Moreover, we conducted a Sargan test (1958) for exogeneity of our instruments which resulted in a p-value of .883. Therefore, we do not reject the null hypothesis of exogenous instruments. Thus, our IV estimated coefficient of schooling shows that for an extra year of schooling, health expenditure increases significantly by 25.0% ($t = 6.13, p < .001$), which is almost twice than when it was estimated by OLS.

Table 1: Coefficients of OLS and 2SLS

	(OLS)	(2SLS)
Independent	log(Hexp+1)	log(Hexp+1)
Schooling Years	0.130 (0.0127)	0.223 (0.0363)
White	0.416 (0.0788)	0.462 (0.0808)
Mental Health	0.348 (0.0452)	0.392 (0.0478)
No. Comorbs	0.827 (0.0283)	0.822 (0.0285)
Public Insured	-0.451 (0.160)	-0.351 (0.167)
Not Insured	-2.030 (0.0913)	-1.892 (0.106)
Family's Income	4.46e-06 (0.000)	3.04e-06 (0.000)
Constant	0.957 (0.389)	-0.417 (0.644)
R-squared	0.337	0.333

Robust standard errors in parentheses

Worth noticing is that OLS has a downward bias, which suggests that the measurement error dominates sources of endogenous bias for years of schooling, since other sources would usually observe an upward bias. Ethnicity-wise, health expenditure for white males is significantly 58.7% higher than that of males of other ethnic groups ($t = 5.72$, $p < .001$), white men spend more on their health than non-white men. As expected, mental health issues and other health conditions significantly affect health expenditure by 48.0% and 127.5% respectively (both $p < .001$). The worse the mental health and the more illnesses a man suffers from, the more he spends on health expenditure. Comparing to men who are privately insured, men who are publicly insured spend 29.6% ($p < .001$) less in health expenditure while men who did not take any insurance spend 84.9% ($p < .001$) less in health expenditure. Last but not least, higher-income families tend to spend more on health expenditure ($p < 0.001$) than lower-income families. However, the marginal effect is negligible as it is too small.

Conclusion

The main goal of this study was to estimate the determinants of health expenditure. Consistent with Shen's (2013) study, our findings suggest that insurance is the most dominant determinant of health expenditure among men in the US, followed by number of illnesses, ethnicity, mental health state and years of schooling, even with more recent and detailed data. Interestingly, the marginal effects of these determinants also appear more pronounced than reported by Shen (2013), perhaps due to inflation over the years. It is especially the case for years of schooling. Furthermore, we showed that it is worth distinguishing between public and private insurance in our analysis (which Shen (2013) consciously left out) as a significant difference was observed, such that men who took out private insurance spent the most on health expenditure, followed by men who took out public insurance and men who did not take out any insurance.

The biggest limitation of the present study is the lack of applicable vaccination data. It is plausible that willingness to be vaccinated could be translated to willingness to spend on health, making it a potentially important determinant of health expenditure. For example, anti-vaxxers who generally do not trust health technologies are likely to be unwilling to spend on healthcare. Since we did not include it in our analysis, it is possible that omitted variable bias will affect our results. Therefore,vaccination data is especially relevant for future research considering that the COVID-19 pandemic has revealed a substantial number of anti-vaxxers in the US.

Additionally, it could be interesting to conduct a tobit regression in future research, as health expenditure is bounded between zero and a limitless positive value. What might be useful for future research is that a Jarque-Bera test (Jarque & Bera, 1980) suggested that the log of positive values of health expenditure follow a normal distribution, which could yield more consistent test statistics.

References

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, *90*(430), 443-450.

Hausman, J. (1978). Specification Tests in Econometrics. *Econometrica*, *46*(6), 1251-1271.

Jarque, C. & Bera, A. (1980). Efficient tests for normality homoscedasticity and serial independence of regression residuals. *Econometric Letters*, *6*, 255–259.

Sargan, J. D. (1958). The Estimation of Economic Relationships Using Instrumental Variables. *Econometrica*, *26*(3), 393–415.

Shen, C. (2013). Determinants of health care decisions: Insurance, utilization, and expenditures. *Review of Economics and Statistics*, *95*(1), 142-153.

Appendix

Table 2: Summary Statistics

	Mean	(Std. dev.)
log(health expenditure + 1)	4.653	(3.460)
Schooling Years	13.1	(3.003)
Private Sector	0.525	(0.499)
Higher-Pay Jobs	0.373	(0.484)
Age	42.2	(13.6)
White	0.740	(0.439)
Family Size	3.23	(1.72)
Mental Health	1.85	(0.778)
No. Comorbs	0.951	(1.33)
Smoker	0.169	(0.375)
Insurance		
Private Insured	0.699	(0.459)
Public Insured	0.061	(0.240)
Not Insured	0.239	(0.427)
Region		
Northeast	0.155	(0.362)
Midwest	0.210	(0.408)
South	0.362	(0.480)
West	0.273	(0.446)
Family's Income	73595.69	(57558.15)
Number of observations	7325	

Table 3: Coefficients of OLS and 2SLS

	(OLS)	(2SLS)
Independent	log(Hexp+1)	log(Hexp+1)
Schooling Years	0.130 (0.0127)	0.223 (0.0363)
Age	0.0376 (0.0144)	0.0351 (0.0145)
Age squared	-0.000226 (0.000160)	-0.000188 (0.000161)
White	0.416 (0.0788)	0.462 (0.0808)
Family Size	-0.184 (0.0211)	-0.149 (0.0242)
Mental Health	0.348 (0.0452)	0.392 (0.0478)
No. Comorbs	0.827 (0.0283)	0.822 (0.0285)
Smoker	-0.203 (0.0913)	-0.163 (0.0960)
Public Insured	-0.451 (0.160)	-0.351 (0.167)
Not Insured	-2.030 (0.0913)	-1.892 (0.106)
Midwest	0.288 (0.111)	0.304 (0.112)
South	-0.237 (0.103)	-0.207 (0.104)
West	-0.182 (0.107)	-0.157 (0.108)
Family's Income	4.46e-06 (6.01e-07)	3.04e-06 (7.85e-07)
Constant	0.957 (0.389)	-0.417 (0.644)
R-squared	0.337	0.333

Robust standard errors in parentheses

Table 4: Reduced form regression

Independent	Schooling Years
Private Sector	-0.350 (0.0587)
Higher-Pay Jobs	1.918 (0.0606)
Age	0.00569 (0.0128)
Age squared	-0.000180 (0.000148)
White	-0.423 (0.0594)
Family Size	-0.290 (0.0204)
Mental Health	-0.389 (0.0398)
No. Comorbs	0.0540 (0.0251)
Smoker	-0.244 (0.0752)
Public Insured	-0.977 (0.154)
Not Insured	-1.200 (0.0846)
Midwest	-0.131 (0.0899)
South	-0.246 (0.0829)
West	-0.271 (0.0894)
Family's Income	1.09e-05 (5.27e-07)
Constant	14.37 (0.287)
R-squared	0.350
F-score (excluding instruments)	584.20

Robust standard errors in parentheses

Table 5: Hausman Test and Sargan-J Test

Independent	e_{OLS}	e_{IV}
Schooling Years	0.0614 (0.0257)	-
Age	-0.00190 (0.0147)	-6.33e-05 (0.015)
Age squared	1.89e-05 (0.000167)	6.05e-07 (0.00)
White	0.00105 (0.0762)	0.00058 (0.079)
Family Size	0.00376 (0.0209)	5.72e-07 (0.021)
Mental Health	0.00253 (0.0445)	-4.79e-05 (0.045)
No. Comorbs	0.000132 (0.0294)	-0.00013 (0.028)
Smoker	0.0149 (0.0904)	0.00030 (0.095)
Public Insured	-0.00324 (0.147)	0.00014 (0.161)
Not Insured	0.0108 (0.0869)	-6.71e-05 (0.089)
Midwest	0.002931 (0.111)	0.00017 (0.112)
South	0.00303 (0.102)	-0.00021 (0.103)
West	-0.00395 (0.105)	-0.0003 (0.108)
Family's Income	-2.59e-07 (6.65e-07)	2.73e-09 (5.88e-07)
Private Sector	-	-0.099 (0.069)
Higher-Pay Jobs	-	-0.004 (0.075)
Residual First Stage	-0.0745 (0.0272)	-
Constant	-0.765 (0.473)	0.0078 (0.345)
Observations	7,325	7,325
nR^2	7.544	0.022
df	1	1
p-value	0.0060	0.8826

Robust standard errors in parentheses