1 Introduction

In recent years, the costs of healthcare have been rising, leading to the increased need of improving the effectiveness of the healthcare system (Bauchner, 2019). In many papers, e.g. in the paper of Shen (2013), the variables affecting insurance coverage, health care utilization, and health expenditures are being investigated. In the field of economics, one could argue that the latter would be an important and valuable factor to determine. In this report, we aim to determine several variables that influence the amount of healthcare expenditures people have. Previous studies have already shown that being insured or not is an important determining factor for healthcare expenditures.

According to Shen (2013), in many papers, insurance is treated as an exogenous variable, whereas there is no evidence that it can be treated as exogenous in the healthcare expenditure equation. Therefore, in this paper, we treat being insured or not as an endogenous variable, and we are going to check whether the conclusion from Garlick & Hyman (2022) that in the case of sample selection, Ordinary Least Squares works just as well, also holds for the Medical Expenditure Panel Survey (MEPS) 2017 data where we analyse health expenditures and being insured or not. To check this, Ordinary Least Squares, the Heckman Selectivity Model, and Polynomial Logit with Newey's selection correction are compared.

The setup of the paper is as follows. Section 2 revolves around the data; section 3 revolves around the relevant models; section 4 explains the results; and in section 5 a conclusion is drawn.

2 Data

This study utilizes data from the Medical Expenditure Panel Survey (MEPS) from 2017, which is conducted annually by the U.S. Department of Health and Human Services to collect information about health insurance, medical expenditures, and health status for a representative sample of the US population. The survey gathers data from individuals, families, and healthcare providers, and is recognized as one of the most thorough sources of healthcare data in the country. For this study, only data from the 2017 surveys of 40 to 59-year-old respondents are analyzed, with the dependent variable being the health expenditures in combination with a dummy for being insured or not. The dataset is composed of over 1800 features classified into four categories: demographic, socioeconomic, health status, and insurance coverage variables.

Shen (2013) formed the basis for selecting most of the features used in this study. Demographic variables such as age, gender, marital status, race, family size, and region were shown to be indicative of the insured equation. Furthermore, socioeconomic variables such as years of education, family income, the number of physician visits, the presence of mental illnesses, and being employed or not were also shown to be indicative by Shen (2013) and are hence also included in our model and in our analysis.

For the health expenditures equation, the number of comorbidities, the presence of mental illnesses, being insured or not, gender, age, family size, the number of physician visits, and family income appeared to be relevant according to Shen (2013).

The dataset contains numerous health-related variables that are highly correlated with each other, as shown in the Appendix in a correlation matrix. To prevent multicollinearity, we defined the variable "number of comorbidities," which tallies the following health issues: high blood pressure, asthma, diabetes, Alzheimer's disease, heart disease, cancer, arthritis, emphysema, osteoarthritis, and stroke. This variable captures the physical health status of individuals and was also used in Shen's (2013) study. Furthermore, the presence of mental illness was used as an explanatory variable in our study, as it led to significant results in Shen's (2013) study.

We have introduced a variable indicating whether or not a person has insurance in order to analyze healthcare costs and their connection with insurance coverage. The likelihood of using healthcare services is higher for insured individuals than for uninsured individuals. Public insurance, private insurance, and insurance, in general, are the three relevant insurance coverage variables in the dataset. As we are interested in examining the association between insurance status and healthcare expenses, we will be using the variable that indicates general insurance coverage for the research. See Table 1 for a summary of statistics of the included variables.

Noteworthy is that we transform health expenditures and family income as follows: $\textbf{health expenditures} = log(\textbf{health expenditures} + 1)$**, and family income** $= log(\textbf{family income} + 1)$. This way, relative increases in the two variables are used in the model, which appears to be of a more explanatory value than without this transformation. Moreover, the addition of the '1' is important since there are 161 households with zero healthcare expenditures, which would result in a negative infinite value after a simple log-transformation without the extra addition. Lastly, the negative values are set to zero by default. In the Appendix, a full description of all

the variables can be found.

TABLE 1: SUMMARY STATISTICS

	Mean	Std	Min	Max
Health expenditures	5,601.474	17,393.926	0.000	499,286.000
Insured	0.870	0.336	0.000	1.000
Age	49.611	5.792	40.000	59.000
Gender	0.541	0.498	0.000	1.000
Married	0.608	0.488	0.000	1.000
Number of comorbidities	0.974	1.140	0.000	7.000
Mental illness	0.088	0.283	0.000	1.000
Employed	0.754	0.431	0.000	1.000
Race (White)	0.470	0.499	0.000	1.000
Years of education	12.900	3.471	0.000	17.000
Family income	82,234.09	71,874.91	-7.621.00	604,068.00
Family size	2.907	1.721	-1.000	13.000
Number of physician visits	3.166	6.444	0.000	232.000
Number of observations	8,010			

3 Methodology

In this report, we estimate health expenditures with OLS, the parametric Heckman selectivity model, and a semi-parametric Newey correction with polynomial logit selection method.

3.1 OLS

When we estimate health expenditures by ordinary least squares we use the following model: $y_i = \beta X_i' + \delta I_i + \varepsilon_i$, where y_i is health expenditure, X_i are the explanatory variables, I_i is a dummy variable that takes value 1 if a person has health insurance and takes value 0 if a person has no health insurance.

3.2 THE HECKMAN SELECTIVITY MODEL

Because the dummy variable of being insured or not is endogenous we also estimate health expenditures with the Heckman two-step procedure. In the first step, the model estimates the probability of a person being insured. We use the probit model to estimate these probabilities. In the second step, we correct for sample selection bias by including the inverse Mills ratio which we have derived from step one. X1 represents the variables of the first-step selection model, X2 represents the variables of second-step model.

The log-likelihood of the **probit model** is as follows:

$$\ell = \sum_{i=1}^{N} \log\left(\Phi\left(\tilde{y}_i x_i \beta_1\right)\right),$$

where $\tilde{y}_i = 2y_i - 1$, and Φ is the cumulative standard normal distribution function.

The Newton-Raphson recursive formula is used to update the $\hat{\beta}_1$ vector. The β_1 vector and the matrix of regressors (X_1) can then be used to compute inverse Mills ratios for privately insured and uninsured individuals in the following way:

$$\begin{aligned}
\lambda_{\text{insured}} &= \frac{\phi\left(X_1\beta_1\right)}{\Phi\left(X_1\beta_1\right)}, \\
\lambda_{\text{not–insured}} &= -\frac{\phi\left(X_1\beta_1\right)}{\left(1-\Phi\left(X_1\beta_1\right)\right)},
\end{aligned}$$

where ϕ, and Φ are the probability density and cumulative standard normal distribution functions, respectively. These inverse Mills ratios can be seen as estimates of the selection bias.

The inverse Mills ratios computed can then be used in the second stage OLS regression of the Heckman model, which is described as:

$$Y_2 = X_2\beta_2 + \delta Y_1 + \theta_1 \lambda_{\text{insured}}\, Y_1 + \theta_2 \lambda_{\text{not-insured}}\, (1 - Y_1) + \varepsilon$$

3.3 NEWEY WITH POLYNOMIAL LOGIT MODEL

We also look at the Newey correction method. In Newey (2009) the following model is considered: Our first model is a series logit model, which uses polynomial expansions in $X1_i$ inside a logistic link function:

$$\Pr\left(Insurance = 1\right) = L\left(\sum_{p=1}^{P}\left(\sum_{j=1}^{K} \psi_k X1_{i,k}\right)^p\right)$$

Second step is :

$$Expenditure_i = X2_i'\beta + Insureced_i\delta + Y1\sum_{q=1}^{Q}\eta_{\text{insured } q}\cdot p_q + (1-Y1)\sum_{q=1}^{Q}\eta_{\text{not-insured } q}\cdot p_q + \xi_i^*$$

with $p_k = [2 * CDF - 1]^k$

In Heckman (1976) the function that determines the selection probability was the probit function and the selection correction function was the inverse Mills ra-

tio. The difference between Newey (2009) and Heckman (1976) is that in Newey (2009) the selection probability function is estimated by a distribution free-method instead than by the probit model. Also, the selection correction function is approximated non-parametric instead of the inverse Mills function.

4 Results

We begin with a simple equation approach using ordinary least squares, which ignores sample selection and endogeneity. The results are below.

Table 2: Output of OLS

	beta
CONSTANT	-0.1232 (0.3368)
Insured	2.3160 (0.0966)*
Age	0.0161 (0.0057)*
Gender	0.6683 (0.0629)*
FAM_SIZE_1and2	0.7144 (0.0887)*
FAM_SIZE_3and4	0.4071 (0.0872)*
Comorbidities	0.7533 (0.0305)*
Mental illness	0.1181 (0.1155)
Number of visits	0.1322 (0.0051)*
Family income	0.1510 (0.0172)*

[1] *significant on a 5% level

Except for mental illness, all other variables show significance. The coefficient of insurance is of interest, and it is interesting to see that there is a significant and relatively big positive effect of whether someone is insured or not on the total healthcare expenditures. However, it is hard to explain given the dependent variable we choose. Total expenditure covers the out-of-pocket payment by patients, as well as the expenditure in insurance, this means people with insurance are very likely to have larger total expenditure by construction. Apart from this, another possible reasons are that people may buy insurance exactly because they know they have a critical illness, such as a genetic disease that requires continuous expenses, or people with insurance go more frequently to the doctor and hence the total health expenditures also go up.

Age has a positive, small but significant effect, it makes sense because older people are more likely to have health issues, which leads to more expenditures,

but only people with age 40 to 59 are considered, so within this sub-population we expect the difference will not be large. However, the results of family size are interesting. Using family size larger than 4 as a reference, people with family size 1 or 2 spend more on health expenditure than people with family size 3 or 4, which means the smaller the family size is, the higher health expenditure is. The reason for this might be that smaller family size makes people have more money spend on their own insurance or other health expenditure, because it is very likely that only a few family members really earn money in a large family. We also see that the number of comorbidities has positive effect on health expenditure, this is in line with our expectations. The same positive effect is observed for the number of visits variable and family income variable, and these two variables are found to have a smaller coefficient than the number of comorbidities, which hints the superior effect of having one comorbidity.

However, exogeneity in this case is a very restrictive assumption, we now estimate the model considering selectivity, we compare the results of parametric selection model (Heckman model) and semi-parametric selection model (Newey with polynomial logistic model). Order of two is used for the polynomial logit model, and order of three is used for the newey model.

We model the probability of having an insurance in the first-step, results can be found in Table 3.

We can see that the signs of the coefficients in the Probit model and the polynomial logit model are the same, but the coefficients are much less significant in the polynomial logit model than in the probit model. Gender, age, marriage status, race,family income,years of education and numbers of visits of doctors have significant effects on the probability of having insurance in both models. We are interested in individual's education level, individuals who have higher education level were found to be more likely to get an insurance as opposed to people with lower education levels. This may be because more educated people tend to be more health and insurance conscious. It is surprising that the number of comorbidities is not significant in the polynomial logit selection model.

The largest effect on the probability of having insurance comes from the race, which is a bit out of our expectation, although literature does suggest that white individuals are the ones who seek private insurance more often. Marriage status comes next. It shows married people are more likely to have an insurance, there are several reasons for it. First, many employers offer health insurance as a benefit to their employees and their families. When a person is married, they are more

Table 3: First-step Results(insurance dummy as dependent variable)

	(1)	(2)
CONSTANT	-1.3174 (0.1088)*	-1.1332 (0.3429)*
Gender	0.1356 (0.0180)*	0.0769 (0.0341)*
Age	0.0134 (0.0016)*	0.0079 (0.0092)
Married	0.1714 (0.0216)*	0.1268 (0.0416)*
White	0.2469 (0.0188)*	0.1974 (0.0476)*
Comorbidities	0.1086 (0.0086)*	0.0320 (0.0180)
Mental illness	0.1099 (0.0335)*	0.0525 (0.0658)
Employed	0.0470 (0.0254)	0.0845 (0.0455)
Family income	0.0631 (0.0065)*	0.0491 (0.0217)*
Years of education	0.1107 (0.0033)*	0.0713 (0.0194)*
FAM_SIZE_1and2	0.1343 (0.0276)*	0.0635 (0.0450)
FAM_SIZE_3and4	0.1117 (0.0258)*	0.0588 (0.0414)
Number of visits	0.0250 (0.0011)*	0.0724 (0.0291)*
RegionNE	0.0051 (0.0269)	0.0083 (0.0537)
RegionMW	-0.1166 (0.0257)*	-0.1058 (0.0548)
RegionS	-0.4529 (0.0229)*	-0.2892 (0.0665)*

[1] Line (1) is coefficients of the first-step probit model of Heckman model, line (2) is coefficients of the polynomial logit selection model.Note that the constructed coefficients of quadratic terms and interactions are not reported.

[2] The bootstrapped standard errors(using 999 replications) are given in parentheses for polynomial logit model

[3] *significant on a 5% level

likely to have access to health insurance through their spouse's employer. Additionally, many insurance companies offer family plans that cover spouses and children, which can make it easier and more cost-effective for married couples to obtain insurance. Furthermore, married couples may have more financial resources and stability, which can make it easier for them to afford health insurance. They may also be more likely to prioritize their health and the health of their spouse and family, which can lead to a greater willingness to pay for health insurance coverage.

In table 4 the estimates of the second-step regression for the Heckman and Newey selectivity model are showed, these two models take endogeneity of insur-

Table 4: Second-step Results(Total expenditure as dependent variable)

	(1)	(2)
CONSTANT	-0.4288	2.2952
	(0.4623)	(0.4031)*
Insurance	4.8838	1.4738
	(0.3788)*	(0.3541)*
Age	0.0102	-0.0005
	(0.0057)	(0.0056)
Gender	0.5608	0.4371
	(0.0676)*	(0.0598)*
FAM_SIZE_1and2	0.4801	0.3556
	(0.0928)*	(0.0896)*
FAM_SIZE_3and4	0.2115	0.0509
	(0.0925)*	(0.0871)
Comorbidities	0.6581	0.6341
	(0.0406)*	(0.0312)*
Mental illness	0.0879	0.1764
	(0.1154)	(0.1141)
Number of visits	0.1325	0.0765
	(0.0244)*	(0.0197)*
Family income	0.0353	-0.0878
	(0.0215)	(0.0224)*
Inv_mills_ratio_insur	-4.2982	
	(0.4992)*	
Inv_mills_ratio_notinsur	-1.1152	
	(0.1876)*	
P_insur		0.5730
		(1.8134)
P_insur^2		1.9634
		(3.7848)
P_insur^3		1.3164
		(2.3454)
P_notinsur		2.9730
		(1.6593)
P_notinsur^2		-5.3139
		(4.5839)
P_notinsur^3		6.6158
		(3.3660)*

[1] Line (1) is coefficients of the second-step of Heckman model, line (2) is coefficients of the Newey correction model.

[2] The bootstrapped standard errors(using 999 replications) are given in parentheses

[3] *significant on a 5% level

ance status into account. There are many interesting finding in the result table. The coefficients of the Heckman model exhibit the same signs as those of the OLS model, which aligns with our initial expectations. However, the age and family income variables in the Newey model display a negative effect on health expenditure, which appears unusual despite the lack of statistical significance in these estimates. The significance of the family income variable is lower in both the Heckman and Newey models compared to the OLS model. The number of comorbidities still is an important variable when estimating the health expenditures, mental illness variable is not.

In the Heckman model, the positive coefficient of insurance variable doubles

in the Heckman model, and the coefficients of two inverse Mills ratio term are significantly large and negative. Hence, it suggests that we do have a self-selection problem here,which will cause the endogeneity. We are not sure how to interpret it, the negative IMR terms imply a negative selection bias, this corresponds to coefficients being downwardly biased if selectivity correction have not been done. The polynomials which estimate the correction term in the Newey model are not significant at the 5% level except the cubic term of not insured, which may suggest the nonlinear correction term. Together with the sign-switched results we do not think Newey with polynomial logit model perform well in this case, in other words, polynomials with order 3 may be not a good estimate for selection correction term. Another thing which is a bit odd is that the coefficient of insurance variable goes down compared with OLS, while the coefficient of cubic term of P_notsur is positively large. This seems to go at the opposite direction compared with Heckman model.

5 Conclusion

For this report, we compare the results of Ordinary Least Squares, the Heckman Selectivity Model, and Polynomial Logit with Newey's selection correction. Given the significant coefficients of selection correction term in the Heckman model, we can infer that selectivity should not be omitted. Heckman and Newey suggest different direction of the selection bias, and the value of the coefficient of insurance variable varies a lot in three models. The sign and significance of some coefficients in the newey model are not in line with other two models and are hard to explain, and most of the polynomial series in the newey model are not significant, so we suspect that Newey perform worst, and Heckman perform best. However, it is total empirical because we don't have a benchmark and a specific metric to compare the models.

In the results, it is shown that gender, age, marriage status, race, the number of comorbidities, family income and education level significantly influence the probabily of having an insurance, insurance, family sizenumber of visits significnatly affect the health expenditue. From the perspective of policy intervention, rising the general education level can lead to higher insurance coverage rate, which may result in lower healthcare costs. However, drawing any other policy-implied conclusion is challenging since the use of total expenditure as the dependent variable makes it difficult to precisely examine the impact of insurance on mitigating health-

care costs.Hence,we can choose a more suitable dependent variable, such as pure out-of-pocket payment, or take utilization of insurance into account in the future study.Moreover, it is worth noting that while both private and public insurance are categorized as 'insured' in the analysis, the reality is that obtaining public insurance may not be a voluntary decision for individuals who meet certain eligibility criteria, such as the elderly, those with disabilities, or low-income individuals. As a result, further research should differentiate between private and public insurance to better understand the impact of private insurance, the role of public insurance in providing healthcare coverage to vulnerable populations, and the potential need for policy interventions to address health disparities among individuals.

In terms of improving our research, there are several model-related factors that we could explore. For instance, we could investigate how changes in the degree of the polynomial, Q, would affect our results. Another option is to combine the different inverse Mills ratios into a single variable and include it in the second stage of the model. Alternatively, we could consider using a multivariate logit model in the first step of the analysis. Additionally, exploring other age groups beyond the specific cohort we focused on would be valuable for future research as it would help provide a more representative view of society as a whole

References

Bauchner, H. (2019). Rationing of health care in the united states: An inevitable consequence of increasing health care costs. *JAMA : the journal of the American Medical Association*, *321*(8), 751-752.

Garlick, R., & Hyman, J. (2022). Quasi-experimental evaluation of alternative sample selection corrections. *Journal of business economic statistics*, *40*(3), 950-964.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4* (pp. 475–492). NBER.

Newey, W. K. (2009). Two-step series estimation of sample selection models. *The econometrics journal*, *12*(s1), S217-S229.

Shen, C. (2013). Determinants of health care decisions: Insurance, utilization, and expenditures. *The review of economics and statistics*, *95*(1), 142-153.

Appendix

DESCRIPTION OF THE VARIABLES AND DESCRIPTIVE STATISTICS

variable	description
Health expenditure	Total expenditure
INSUR	Dummy variable for Health Insurance Coverage Indicator 2017
	1: the respondent is insured (either private or public), 0: the respondent is not)
Number of visits	The number of office-based physician visits in 2017 for each individual
AGE	Age as of 12/31/17
WHITE	Dummy variable for race (1: the respondent is white, 0: the respondent is not)
MARRIED	Dummy variable for marital status (1: the respondent is married, 0: the respondent is not)
regionNE	Dummy variable for the region for the Census Region as of 12/31/17
	(1:the Census Region is Northeast, 0: it is not)
regionMW	Dummy variable for the Census Region as of 12/31/17
	(1:the Census Region is Midwest, 0: it is not)
regionS	Dummy variable for the Census Region as of 12/31/17
	(1: the Census Region is South, 0: it is not)
Years of education	The number of years of education completed
FAM_SIZE_1 and 2	Dummy variable for the family size (1: the family size as of 12/31/17 is 1 or 2, 0: it is not)
FAM_SIZE_3 and 4	Dummy variable for the family size (1: the family size as of 12/31/17 is 3 or 4, 0: it is not)
FAM_INC	Family's Total Income
MENTAL	Dummy variable for the perceived mental health status
	(1: the respondent has a mental illness, 0: the respondent has not)
Comorbidities	The number of comorbidities
Gender	Dummy variable for Gender as of 12/31/17
	(1: female, 0: male)
Employed	Dummy variable for employment status as of 12/31/17
	(1: "currently employed","has a job to return to",or"employed during the reference period", 0: "not employed with no job to return to")

CORRELATION MATRIX