Assignment 3 (Using Dataset 2)

Students should submit a hardcopy that contains:

- Answers in text
- An output file generated by Stata (a log file) or R (a sink file)

OBJECTIVE

The objective of this report is to find the data generating processes for income and health, after controlling for age, inherited health, and working experience. A two-stage least square procedure was conducted to investigate if income and health were simultaneously determined in a system and if this system was well-identified.

METHODS

Data Source, Outcome, and Exposure

Person-level data from "Data2" was used in the analysis. The primary outcome is an individual's income which is measured in thousands of dollars. The main exposure variable is an index of an individual's health status, where higher values correspond to better health.

Endogeneity of Health

Health is an endogenous variable that is used to predict income for two main reasons: First, there is reverse causality. While health may directly influence income, we expect high income to promote better health as well; Second, there is significant confounding due to both observed and unobserved variables that impact both health and income. For instance, older people are more likely to have higher income compared with their younger counterparts while they may also have worse health. In addition, there are factors that are not captured in our dataset, such as education attainment, that can influence both income and health.

Covariates

An individual's current age (years), working experience (years), and inherited health status were captured in the dataset. Inherited health is an index that was calculated as the average lifetime health index of an individual's parents.

A Two-Stage Least-Square Model (2SLS)

Due to the hypothesized simultaneity of income and health, we used a two-stage least-square model where a health equation and an income equation were estimated simultaneously using age, working experience, and inherited health.

Health equation:

$$\text{Health} = \alpha_0 + \alpha_1\text{Income} + \alpha_2\text{f(Inherited Health)} + \alpha_3\text{g(Age)} + \varepsilon_1$$

Income equation:

$$\text{Income} = \beta_0 + \beta_1\text{Health} + \beta_2\text{p(Working Experience)} + \beta_3\text{q(Age)} + \varepsilon_2$$

At the first-stage, health was modelled using income, inherited health, and age. Working experience was excluded as a potential independent predictor of health since we do not expect an individual's number of working years to directly influence health status. In the second-stage income equation, we used health, working experience and age to predict income. Inherited health was not entered into the equation as a predictor of income since parental health in itself does not play a direct role in determining a child's current earning. Hence when we estimate the 2SLS model, an individual's inherited health will be used as an instrumental variable for health status while working experience will be used as an instrumental variable for income.

In order to determine the correct functional form of each predictor to enter the two equations, i.e., to specific the function f(·), g(·), p(·), and q(·), we visually inspected the relationship between each independent predictor and health/income by plotting them in a scatterplot. A clear linear association between the predictor and the corresponding outcome (health or income) would indicate that no transformation was needed for the predictor and we should enter it directly into the equation. A logarithm curve would necessitate a log-transformation of the predictor prior to entering it into the equation. A polynomial curve indicated a need to add one or more polynomial terms of the independent variable. For example, a U- or inverted U-shaped curve required an addition of the quadratic term of the predictor into the regression equation.

Durbin-Wu-Hausman's test was conducted for both the health equation and the income equation to assess the validity of using an individual's inherited health and working experience as an instrumental variable.

Sensitivity Analyses

Model diagnostics were performed to examine the performance of the 2SLS. The Ramsey Regression Equation Specification Error Test (RESET) was conducted to assess the presence of omitted variables. Shapiro-Wilk Test was used to assess for normality of residuals. Heteroscedasticity and the presence of outliers were checked by plotting the residuals and then the fitted values against each predictor of the 2SLS procedure.

RESULTS

Baseline Characteristics

Table 1 reports the summary statistics of individuals in the study cohort. Among a total of 320 individuals, the mean income is $97.85K that ranges from $59.68K to $125.68K. The average health index is 76.17 where the lowest score occurs at 18.33. Inherited health index varies from

50 to 100 where the mean score is at 75.53. Individuals have a mean age of 50.13 years and report to have an average of 18.96 years of working experience.

Table 1. Summary statistics

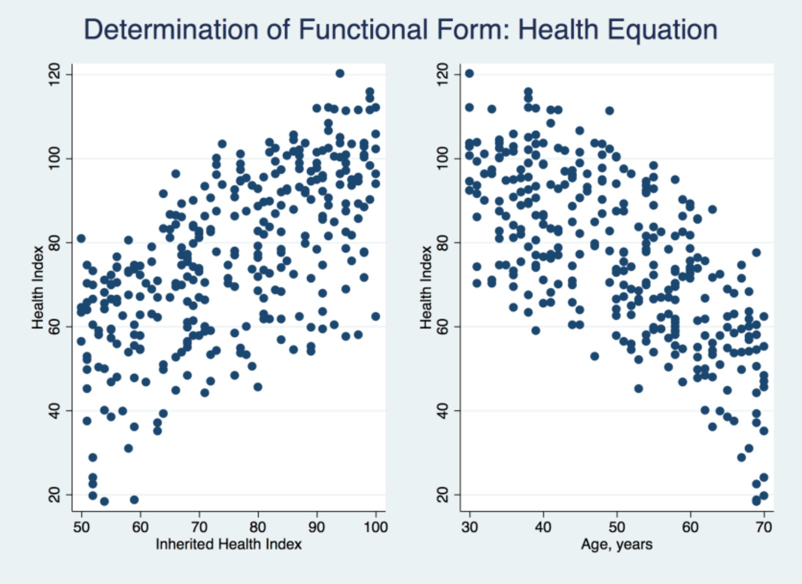
Variable	Mean (SD)	Median (IQR)	Min	Max
Health	76.17 (19.92)	75.58 (31.00)	18.33	120.27
Income, 1 000USD	97.85 (13.74)	98.68 (20.03)	59.68	125.68
Inherited health	75.53 (14.92)	76 (26)	50	100
Age, years	50.13 (11.59)	51 (20.5)	30	70
Working experience, years	18.96 (4.14)	19 (8)	12	25

Specifying the Functional Form of the 2SLS

Health Equation

Figure 1 contains two scatterplots where the observed health index was plotted against an individual's inherited health index (left) and age (right), respectively.

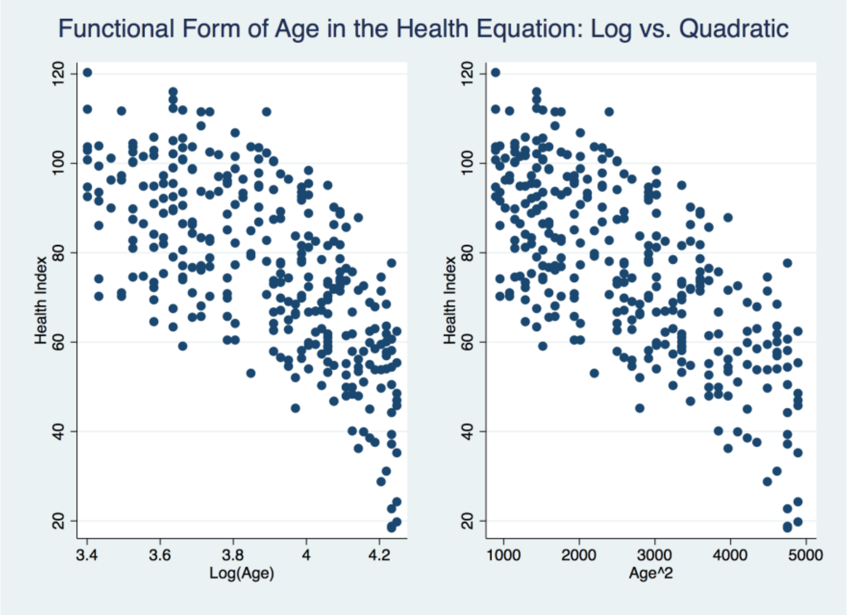
Figure 1. Determination of functional form in the health equation



An individual's health index was roughly linearly associated with the inherited health index, as evident by the linear pattern displayed in the scatterplot on the left. The age-health relationship was not linear.

We presented two scatterplots in Figure 2 below where health index was plotted against the log-transformed age and age^2, respectively:

Figure 2. Functional form of age in the health equation: log vs. quadratic



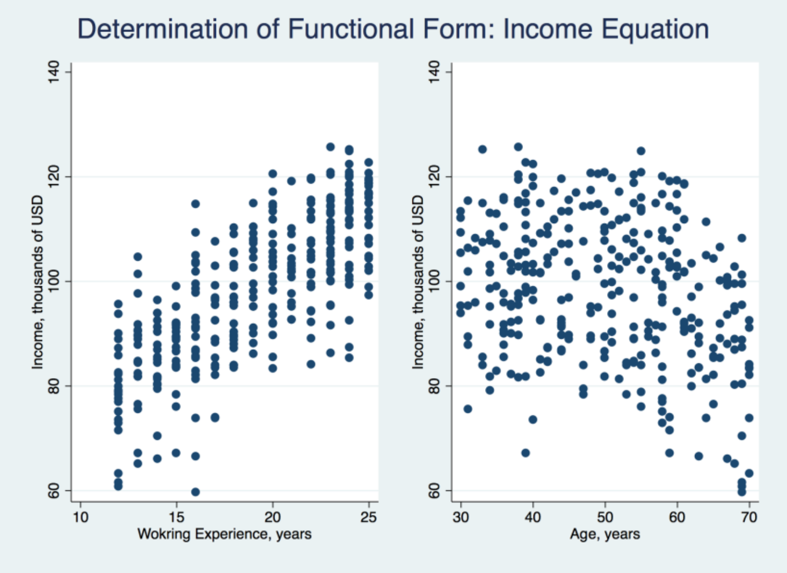
It is very clear from the two scatterplots that the relationship between health and log(age) is still non-linear (left) while the association between health and age^2 takes on a linear form (right). Hence, we decided to fit the following health equation into our system where we examined the log model in sensitivity analysis.

$$\text{Health} = \alpha_0 + \alpha_1\text{Income} + \alpha_2\text{Inherited Health} + \alpha_3\text{Age} + \alpha_4\text{Age}^2 + \varepsilon_1$$

Income Equation

Figure 3 reports the two scatterplots where an individual's observed income was plotted against working experience and age, respectively:

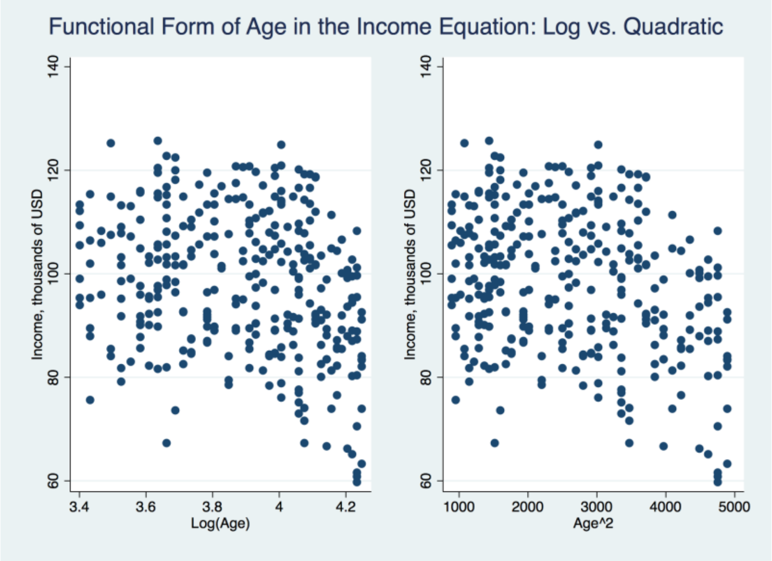
Figure 3. Determination of functional form in the income equation



The relationship between an individual's income and working experience appears to be linear except for eight people who have reported to work for about 10-20 years but had exceptionally low income (left). We will address these potential outliers in sensitivity analyses.

Age is not linearly associated with income as demonstrated by the scatterplot on the right. In this way, we plotted income against the log-transformed age and the age^2 in Figure 4 on the next page:

Figure 4. Functional form of age in the income equation: log vs. quadratic



Taking the natural log of age does not improve the linearity between income and age (left), which is consistent with our findings in the health equation. There appears to be a very weak negative linear association between income and age^2. In this way, we have determined the functional form for the income equation as below:

$$\text{Income} = \beta_0 + \beta_1\text{Health} + \beta_2\text{Working Experience} + \beta_3\text{Age} + \beta_4\text{Age}^2 + \varepsilon_2$$

Estimating the System of Equations

We present the regression results of the 2SLS model in Table 2. Coefficient estimation for each predictor was reported with p-values in the round brackets and 95% CI in the square brackets. We also reported the R^2 and Durbin-Wu-Hausman test results for both equations.

Table 2. Summary of 2SLS regression results

Equations	**Income Equation**	**Health Equation**
Variables	Income	Health
Health	0.42 [0.38-0.46] (0)	-
Income	-	0.29 [0.24-0.35] (0)
Inherited health	-	0.72 [0.68-0.76] (0)
Age	1.26 [0.81-1.71] (0)	1.06 [0.60-1.51] (0)
Age^2	-0.01 [-0.02 - -0.01] (0)	-0.02 [-0.03- -0.02] (0)
Working experience	2.07 [1.93-2.20] (0)	-
Intercept	-7.46 [-18.70-3.78] (0.19)	-3.83 [-14.84-7.17] (0.50)
R^2	0.87	0.94
Durbin-Wu-Hausman	0	0

Durbin-Wu-Hausman test revealed that our uses of an individual's inherited health (p-value=0) and working experience (p-value=0) as an instrumental variable in the respective health equation and income equation were both valid. Both equations were highly significant, as evident by the big R^2 score achieved. We therefore concluded that 87% of variations of an individual's income were explained by the income equation while 94% of variations of observed health status were captured by the health equation.

Sensitivity Analysis

We conducted two major steps in our sensitivity analysis, comprising (1) perform model diagnostics of the primary model, and (2) compare an alternative 2SLS model where log-transformed age was entered into both equations.

Model Diagnostics of the Primary Model

We further verified the validity of using inherited health and working experience as an instrumental variable in the respective health and income equation by checking the t-statistics. Since both t-statistics far exceeded the rule-of-thumb benchmark values (38.58 for inherited health and 29.91 for working experience), we concluded that both inherited health and working experience were very strong instrumental variables.

We plotted the residuals from the health and income equation in Figure 5 and statistically verified the presence of normality using the Shapiro-Wilk Test. Since both series of residuals had a bell-shape distribution and we failed to reject the normality assumption (p-value=0.17 for the health equation and p-value=0.69 for the income equation), we could not reject the hypothesis that the two regression errors were normally distributed.

We then plotted fitted income/health vs. respective residuals to assess whether the variance of the residual was constant, i.e., heteroskedasticity (Figure 6). Because the points were randomly scattered, we inferred that there was no heteroskedasticity in the two regression equations.

We further checked for heteroskedasticity and the presence of outliers by plotting the residuals against each predictor for the health equation in Figure 7 and for the income equation in Figure 8. We found that the residuals from both equations were randomly scattered at all levels of each predictor. Therefore, both equations neither produced heteroscedasticity nor significant outliers.

Since we have detected no significant violations of model assumptions, we concluded our primary model to be valid and our results to be largely robust.

Figure 5. Histograms of residuals from the two equations

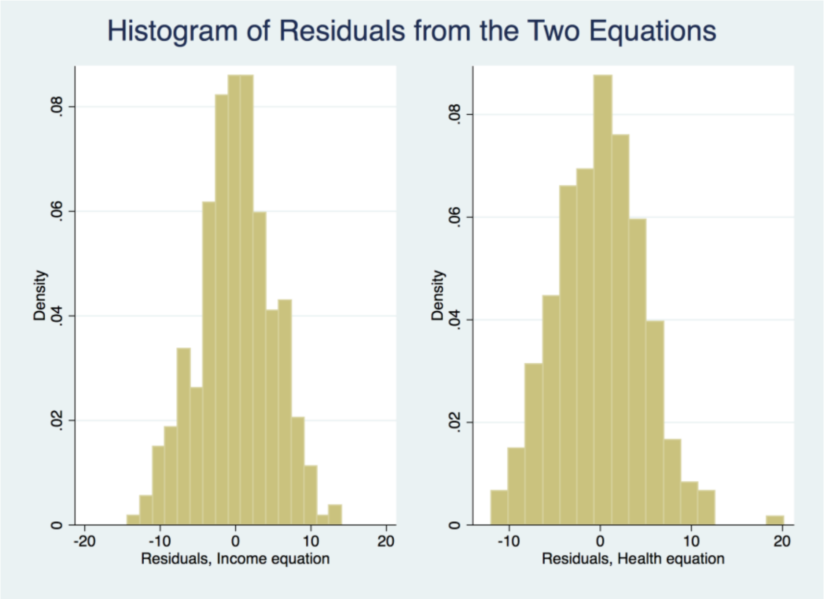


Figure 6. Scatterplots of fitted value vs. residuals to check heteroscedasticity

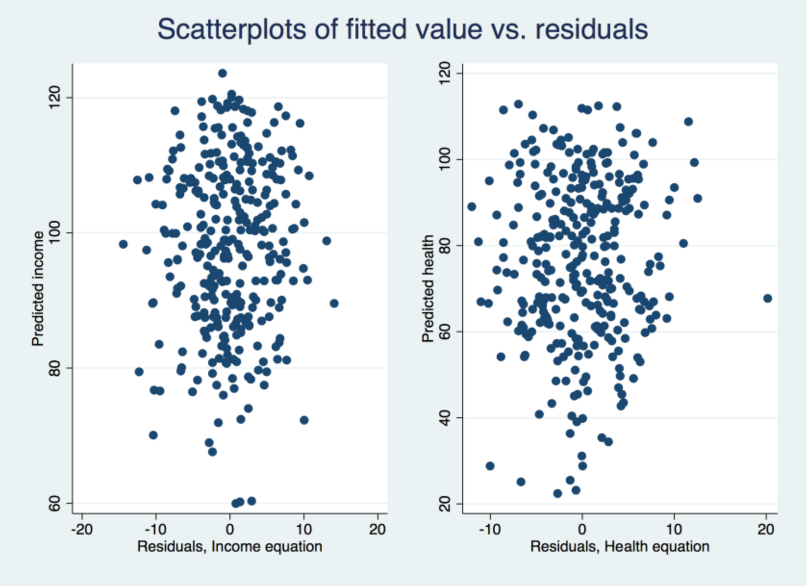


Figure 7. Residuals vs. predictor of the health equation

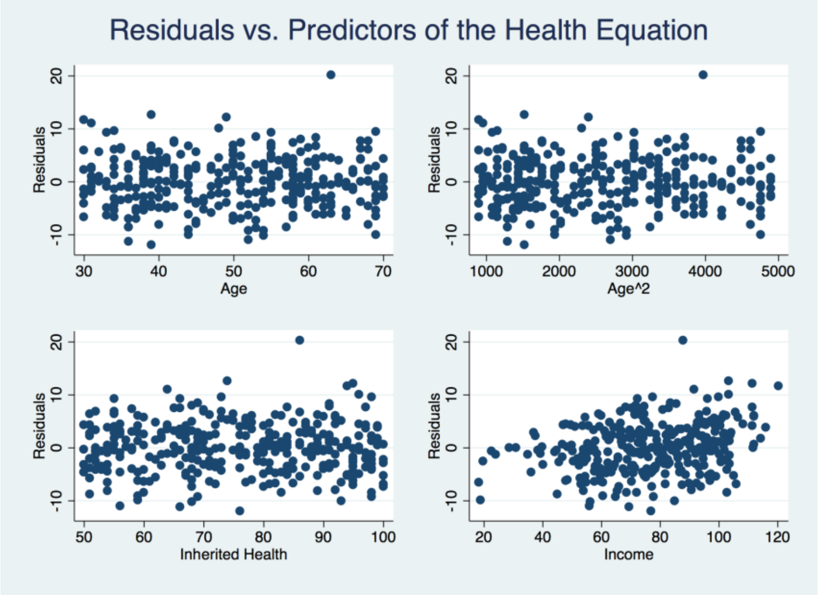
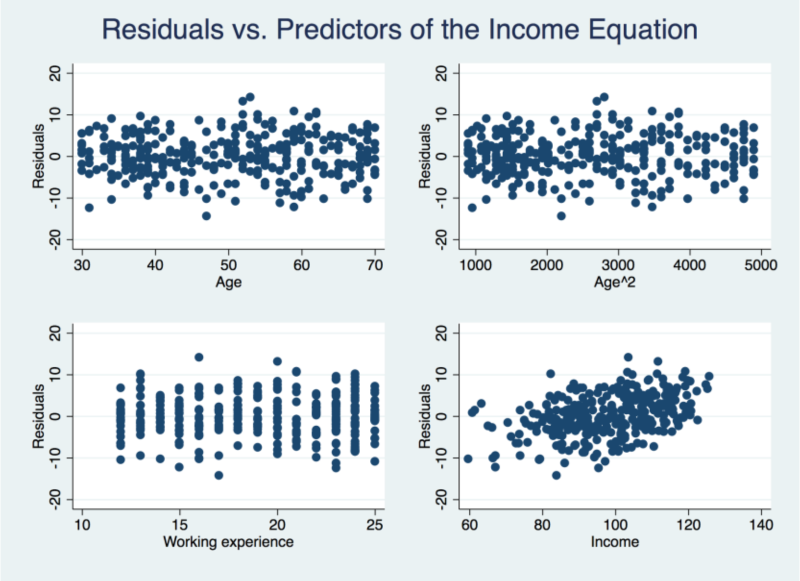


Figure 8. Residuals vs. predictor of the income equation



Model Comparison: Log(age) vs. age^2

We used the log-transformed age variable instead of age and age^2 in 2SLS and re-estimate the regression results. Comparison of regression results is provided in Table 4 below.

Table 3. Comparison of 2SLS regression results: log model vs. primary model

Models	**Primary Model**		**Log Model**	
Equations	Income Equation	Health Equation	Income Equation	Health Equation
Variables	Income	Health	Income	Health
Health	0.42 [0.38-0.46] (0)	-	0.42 [0.37-0.46] (0)	-
Income	-	0.29 [0.24-0.35] (0)	-	0.31 [0.24-0.38] (0)
Inherited health	-	0.72 [0.68-0.76] (0)	-	0.71 [0.66-0.76] (0)
Age	1.26 [0.81-1.71] (0)	1.06 [0.60-1.51] (0)	-	-
Age^2	-0.01 [-0.02 - -0.01] (0)	-0.02 [-0.03- -0.02] (0)	-	-
Log(Age)	-	-	8.20 [4.79-11.62] (0)	-50.09 [-53.04 --47.16] (0)
Working experience	2.07 [1.93-2.20] (0)	-	2.08 [1.94-2.22] (0)	-
Intercept	-7.46 [-18.70-3.78] (0.19)	-3.83 [-14.84-7.17] (0.50)	-5.22 [-20.75-10.30] (0.51)	187.12 [173.01 – 201.23] (0)
R^2	0.87	0.94	0.87	0.91
Hausman	0	0	0	0

The log model gave largely consistent regression results compared with our primary model. Specifically, in the log model, the direct effect of health on income remained unchanged while income had a slightly increased influence on health in the health equation. Like our primary model, the log model achieved very high R^2 scores and passed the Durbin-Wu-Hausman test. In order to statistically compare the fit of the two models, we conducted the

RESET to see if any of them was at risks of omitted variables and had a mis-specified form.

Summary of RESET results is provided in Table 4:

Table 4. Compare the Ramsey Regression Equation Specification Error Test (RESET) between the log model and the original model

RESET Results	Primary Model	Log Model
F-statistics	1.72	14.13
P-value	0.16	0

The log model had an inadequate model fit as tested by RESET. Therefore, we choose our primary model over the log model and conclude again that our original regression results are robust.

DISCUSSION

In the present report, we have derived a 2SLS model where an individual's income and health are simultaneously determined:

$$\widehat{\text{Health}} = -3.83 + 0.29\text{Income} + 0.72\text{Inherited Health} + 1.06\text{Age} - 0.02\ \text{Age}^2 + \varepsilon_1$$

$$\widehat{\text{Income}} = -7.46 + 0.42\text{Health} + 2.07\text{Working Experience} + 1.26\text{Age} - 0.01\text{Age}^2 + \varepsilon_2$$

Effect of Working Experience on Income

An individual's working experience has a direct and significant effect on income as one additional year of working is associated with a $2.07K or $2070 increase in income.

Effect of Inherited Health on Health

An individual's inherited health has a weak but significant influence on current health as a one-point increase in the inherited health index is associated with a 0.72-points increase in the health status of the individual.

Effect of Age on Income and Health

Our simultaneous equation system suggests that an individual's age has a significant influence on both health status and income. Specifically, a one-year increase of age is associated with a 1.06-points increase in health index and a $1.26K or $1260 increase in the individual's income. Furthermore, the age effect on health and income decreases as people get older, as implied by the negative coefficient of the age^2 variable in both equations.

Effect of Income on Health

Income is positively associated with health where each $1K increase in income is linked with a 0.29-points increase in the individual's health index.

Effect of Health on Income

We found a positive relationship between higher income and better health as each 1-point increase in the health index is associated with a significant $0.42 or $420 increase in income.

CONCLUSION

We built a two-stage least-square model that simultaneously determined an individual's income and health. The system was well-defined as we have demonstrated in sensitivity analysis.