CAUSAL EFFECT OF TYPE OF HEALTH INSURANCE COVERAGE ON BMI: AN INSTRUMENTAL VARIABLE ANALYSIS

KAY ROYO

Abstract. This study explores the effectiveness of health insurance coverage on health management by examining its relationship with Body Mass Index (BMI). Through the application of instrumental variable (IV) analysis and the consideration of omitted variables in observational data, the study aims to estimate the causal effect and generate valuable insights for intervention development. Understanding the impact of health insurance on BMI is crucial for informing effective public health policies and improving health outcomes.

1. Introduction

Obesity is a significant public health concern in the United States. According to the Centers for Disease Control and Prevention (CDC), the prevalence of obesity among adults rose significantly from 30.5% in 1999-2000 to 41.9% in 2017-2020. This increase highlights the growing urgency to address this issue. The financial implications of obesity are also substantial. In 2019, the estimated annual medical cost associated with obesity in the United States was approximately $173 billion. Furthermore, the medical costs for adults with obesity were $1,861 higher compared to those with a healthy weight [1].

This analysis focuses on the association between the type of health insurance coverage and BMI (Body Mass Index), as a measure of health outcome. While BMI is not a direct measure of overall health, it is widely used as an indicator of obesity and can be associated with various health outcomes. The association between obesity and various health outcomes is well-documented. Higher BMI values and obesity are often linked to an increased risk of chronic conditions such as cardiovascular disease, stroke, type 2 diabetes, and certain cancers [5]. Numerous studies have consistently demonstrated that the growing prevalence of obesity is linked to economic factors such as poverty, unemployment, and income levels [6]. These findings suggest that individuals from lower socioeconomic backgrounds are at a higher risk of obesity or encounter greater difficulties in maintaining a healthy lifestyle. Furthermore, despite the availability of different health insurance options, individuals with disadvantaged socioeconomic status may face challenges in accessing optimal health insurance coverage with comprehensive health management programs.

Understanding the relationship between health insurance coverage and BMI is crucial for several reasons. Health insurance coverage plays a critical role in facilitating access to healthcare services, including preventive care, screenings, and treatment options. Individuals with adequate health insurance coverage may have better opportunities for weight management and obesity prevention. Moreover, disparities in health insurance coverage exist, particularly among different socioeconomic groups. Studying the association between health insurance coverage and BMI can shed light on how disparities in access to healthcare may contribute to differences in obesity rates among different populations.Therefore, this study has the potential to provide valuable insights that can aid the development of effective public health policies and interventions aimed at reducing obesity rates and improving health outcomes.

The primary objective of this analysis is to employ instrumental variable (IV) analysis on observational data to estimate the causal effect of the type of health insurance coverage on BMI. While randomized experiments are considered the most reliable method for determining causal effects, observational studies are often more practical and feasible. In observational studies, the treatment assignment is not controlled by the researcher but occurs naturally. Although they may be subject to certain limitations, such as potential confounding factors and selection biases, observational studies can still provide valuable insights into causal relationships. Using the IV method, the causal effect of health insurance coverage on BMI can be estimated using an instrumental variable, such as employment. Based on the previous studies done on this subject, it can be hypothesized that a positive correlation between health insurance coverage and BMI exists [2].

In this study, instrumental variables are utilized to address the issue of endogeneity where the relationship between the independent variable (health insurance coverage) and the dependent variable (BMI) is confounded by unobserved factors. IVs can provide a means to establish causality in the relationship between the independent variable and the dependent variable with reduced bias from endogeneity. Through the utilization of IV analysis, the issue of endogeneity can be mitigated by introducing an exogenous source of variation, such as employment status, in the type of health insurance coverage. This approach ensures that the relationship between health insurance and BMI is examined independently of other factors that may influence BMI.

2. Data and Methodology

2.1. Data

To study the relationship between health insurance coverage and BMI, a subset of secondary data from the 2021 California Health Interview Survey (CHIS 2021) is utilized. The CHIS 2021 dataset, a population-based survey collected by UCLA-Center for Health Policy Research (UCLA-CHPR) between March 2021 and November 2021, contains individual records of survey responses provided by adult participants. This dataset is representative of the non-institutionalized population residing in households across California. Approximately 90% of adult interviews were conducted online, while the remaining 10% were carried out via telephone. The primary purpose of this survey is to gather information on health status and healthcare-related issues among Californians, with the aim of informing public health policy-making in local communities within the state. The survey questions included in this observational study have been consistently modified over time to address evolving public health concerns. This dataset includes a total of 782 features, covering diverse topics such as housing, childhood experiences, physical and mental health, healthcare discrimination, cannabidiol and alcohol use, intimate partner violence, and other demographic information. It's important to note that geographical information is not publicly accessible within the dataset due to privacy concerns.

The preprocessing of this data involves several steps. To simplify the data, a substantial number of features are first removed. Additionally, the selected health insurance coverage variable is re-coded and transformed into a binary variable. Specifically, all types of commercial or private insurance are consolidated into a single group, while government and state-funded health insurance, such as Medicaid and Medicare, are merged into another group. This re-coding allows for a more simplified and meaningful comparison of the different types of health insurance coverage in the study.

The exclusion criteria for this study involves the removal of subjects without health insurance at the time of the survey, as the main interest is to compare the effects of public and non-public insurance. Pregnant subjects are also excluded due to the significant changes in body composition associated with pregnancy. Including pregnant women could introduce bias as pregnancy is a temporary condition and may not accurately represent the broader population.

2.2. Exploratory Data Analysis

After applying the exclusion criteria, the final study population consists of $23,405$ total subjects. Among them, $10,116$ individuals (43.2%) have public insurance, while $13,289$ individuals (56.8%) have private insurance coverage. Using the categorical obesity variable, the overall prevalence of obesity among the $23,405$ individuals is 60.4%. Specifically, the percentage of individuals with obesity in the private insurance group is 59.1%, while the percentage of individuals with obesity in the public insurance group is 62.2%. The average BMI for all subjects included in this analysis is 27.44 (standard deviation: 6.14, range: $12.94 - 65.60$) as shown in Table 1. Table 1 also displays that subjects classified as obese have an average BMI of 30.9. The age range of the participants ranges from 18-85 years. It is worth noting that this dataset does not contain any missing values, ensuring the completeness of the data for analysis.

According to Figure 4 in the appendix section, it is evident that the continuous response variable of interest, BMI, exhibits a significant right-skewness. In IV analysis, the normality assumption is not required. However, it is important to note that the validity of statistical inference, such as hypothesis tests and confidence intervals, in IV analysis relies on large sample sizes and asymptotic theory, which assumes the normality of the errors. Previous studies have shown that better coverage rates can be obtained using normally distributed data even when the sample size is small [9]. Therefore, to obtain more accurate results, inverse transformation is implemented to normalize BMI. As shown in Figure 3, there is a minimal difference in the mean BMI between public and private insurance. The cross tabulations results in Table 2 reveals a statistically significant relationship between obesity and the type of health insurance coverage (p-value = 0). Additionally, there is a statistically significant association between employment and the type of health insurance coverage (p-value = 0). However, no significant association is observed between employment and obesity (p-value = 0.92).

2.3. Method

To establish the causal effect or unbiased estimate of health insurance coverage on BMI, employing IV analysis is a suitable approach. Given that the CHIS 2021 dataset is observational, gathered without interventions or variable manipulations, this method proves valuable due to the potential presence of endogeneity or unobserved confounding variables in this type of study. Unlike methods such as propensity scores, regression, and matching, which only control for measured confounders, instrumental variable method can be used to control for unmeasured confounders. These considerations are essential for accurately determining the causal relationship between the variables of interest.

Commonly used in economics and health studies, IV analysis aims to identify the variables that have a causal impact on the determination of the treatment participants receive. In the context of this analysis, the objective is to identify the variable that influences the type of insurance individuals obtain. The chosen instrument variable is utilized to capture the variation in the treatment that is not affected by unmeasured confounders. Subsequently, this variation that is free from unmeasured confounding factors is employed to estimate the causal effect of the treatment on the response [3]. Observational studies, such as CHIS 2021, lack the ability to control variables and minimize bias through random treatment assignment, as seen in randomized experiments. However, instrumental variable methods can help extract the randomization portion that is present in the study, enabling the establishment of robust causal relationships despite the limitations of observational designs.

A simple estimation of the impact of health insurance coverage on health would involve comparing the outcomes of individuals with public insurance to those with private insurance. However, the selection of health insurance is not a random process. It is influenced by multiple factors such as employment, income, and other possible demographic characteristics. Many important factors affecting the decision-making process regarding health insurance are not observable. Consequently, utilizing a simple regression analysis would lead to biased estimates of the true effect of health insurance coverage on BMI. In essence, regression estimates would capture solely the strength of the association, without providing insights into the direction and magnitude of causation.

For this analysis, specific health and demographic information from the CHIS 2021 dataset is utilized. The binary variable representing type of health insurance coverage is treated as an endogenous treatment, while the numerical BMI serves as the dependent variable. Several measured confounders or exogenous variables, including *age*, *race*, *education*, *gender*, *self-reported education*, *total household income*, *general health condition*, *percent of life in the US*, and *family size*, are included in the model as control variables. These control variables are assumed to be predetermined and independent of other variables in the model. In this study, various potential candidates for the instrumental variable are examined to identify the most suitable IV to use in the analysis.

IV regression is a method used to estimate causal relationships between variables in the presence of endogeneity in large samples. One common approach to implement IV regression is through the two-stage least squares (2SLS) method. Other methods used to implement IV regression include generalized method of moments (GMM) approach. The 2SLS method consists of two distinct stages. In the first stage (2), the instrument-treatment correlation is estimated through regression analysis. This stage yields an adjusted treatment variable, which captures the portion of the treatment explained by the instrument and is devoid of any influence from unmeasured confounders. The instrument is assumed to be exogenous, meaning it is randomly assigned, thereby rendering the adjusted treatment exogenous as well. The second stage (3) revolves around estimating the treatment effect using the adjusted treatment variable obtained from the first stage. These steps are carried out in subsequent sections to execute the necessary procedures for the 2SLS estimation. IV regression (1) and 2SLS with one endogenous regressor, can be generally described as follows.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ji} + \beta_{j+1} W_{1i} + \cdots + \beta_{j+k} W_{ki} + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where
Y_i is the dependent variable
$X_{1i}, \ldots, X_{ji}$ are j endogenous variables
$W_{1i}, \ldots, W_{ki}$ are the k included exogenous variables that are uncorrelated with ϵ_i
$\beta_0, \ldots, \beta_{j+k}$ are $1 + j + k$ unknown regression coefficients
$Z_{1i}, \ldots, Z_{mi}$ are the m instrumental variables (excluded exogenous variables)
ϵ_i is the error term

Stage 1:

$$\hat{X}_i = \alpha_0 + \alpha_1 Z_{1i} + \cdots + \alpha_k Z_{mi} + \alpha_{j+1} W_{1i} + \cdots + \alpha_{j+k} W_{ki} + \eta_i \tag{2}$$

Stage 2:

$$Y_i = \beta_0 + \beta_1 \hat{X}_{1i} + \beta_2 W_{1i} + \cdots + \beta_{1+k} W_{ki} + \epsilon_i \tag{3}$$

where both ϵ and η are error terms

It is also important to recognize the limitations of IV analysis. This approach is not easily applicable to non-linear models [11]. The linearity assumption is relevant for the relationship between the endogenous variable and the instrumental variable. Furthermore, even a slight correlation between the instrumental variable and the error term can introduce significant bias in the instrumental variable estimator. Moreover, the performance of 2SLS, the method used to implement IV regression, can be suboptimal in small samples and when errors exhibit heteroskedasticity [10]. Therefore, the use of heteroskedasticity-robust standard errors is recommended to address potential heteroskedasticity in the model.

2.4. Instrumental Variables

Prior to the implementation of the 2SLS method, a comprehensive assessment of various instrumental variable candidates listed in Table 4 is conducted. The initial step in instrumental variable analysis involves identifying instrumental variables that affect the assignment of treatment but is uncorrelated with unmeasured confounders and has no direct impact on the outcome, except through its influence on the treatment. The selection of the most suitable instrumental variable is critical prior to conducting 2SLS as the validity of the instrumental variable approach primarily depend on the quality of the chosen instrument. Figure 1 shows the Directed Acyclic Graph (DAG) of the causal relationships among the variables in this study and the main assumptions for instrumental variable. An ideal instrumental variable should meet the following criteria.

1. Relevance: A valid IV is associated with the treatment i.e. $\mathrm{corr}(X, Z) \neq 0$.
2. Exogeneity or effective random assignment: IV must be uncorrelated with the error term or unmeasured confounders, $\mathrm{corr}(\epsilon, Z) = 0$.
3. Exclusion: IV indirectly affects the outcome solely through the treatment variable.

Additional criteria may be required especially in the case of nonlinear models since these criteria are primarily useful for linear models [13]. Although it is infeasible to test all the assumptions and features thoroughly, there are methods available to test certain aspects

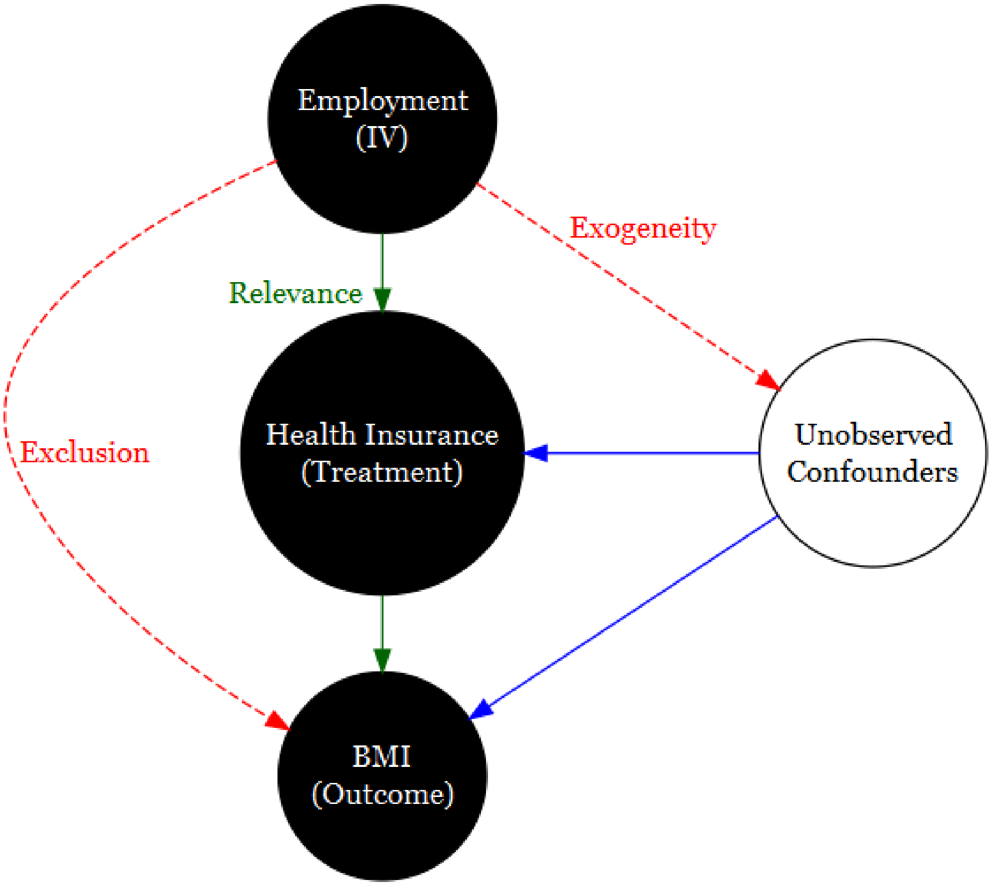


FIGURE 1. DAG of Causal Relationships

of these assumptions. Even if there is no perfect instrumental variable, conducting an IV analysis can still provide valuable information about the treatment effect [3]. To identify the most suitable IV for this analysis, several statistical tests are performed. Table 4 presents the results of the *Chi-square test*, which examines the association between the categorical IV candidates and the categorical treatment variable, health insurance coverage. The obtained p-values indicate that all variables in the table are significantly associated with the treatment at a significance level of $\alpha = 0.01$.

Additionally, the *Kruskal-wallis test*, a non-parametric statistical test, is used to assess whether there are statistically significant differences in the distribution of the numerical dependent variable, BMI, among the groups defined by each potential IV. The results reveal that only the employment variable does not demonstrate a statistically significant difference in BMI ($p - value = 0.6951$). This suggests a lack of direct association between employment and BMI. This finding is further supported by fitting separate ANOVA models, where each variable is individually treated as a predictor of BMI. The ANOVA model with employment as the predictor demonstrates no significant difference in the average BMI across different employment levels ($p - value = 0.275$). Thus, employment appears to be the only variable that satisfies the Relevance and Exclusion criteria.

As testing the Exogeneity criterion directly is challenging, the sensitivity analysis results can be utilized to confirm the adequacy of employment as an instrumental variable. Given that the exclusion condition is satisfied by employment as an instrumental variable, it can be assumed that the exogeneity condition is also met. This assumption arises from the fact that the exclusion condition implies that the instrumental variable has no direct effect on the error term. Therefore, employment can be considered exogenous in the analysis, meaning it is uncorrelated with the error term and affects the outcome variable solely through its impact on the treatment variable. It is worth noting that some of the potential IVs examined that do not meet all the criteria can instead be incorporated as control variables due to their significant impact on the outcome variable. This approach ensures that their influence is appropriately accounted for in the analysis.

3. Results

3.1. Inferential Analysis

3.1.1. *OLS*

For this analysis, a multivariate linear regression model is used as a baseline model to compare the results of 2SLS to. The *Likelihood Ratio test* is conducted to assess the significance of each potential control variable. The test suggests that the variable citizenship can be excluded from the set of control variables since it does not have a statistically significant effect in the model. However, it should be noted that citizenship does exhibit significant differences in BMI among its categories, as indicated by the fitted ANOVA model summary. Therefore, the final baseline model is defined as follows.

$$\begin{aligned}\mathrm{BMI}^{-1} = \beta_0 &+ \beta_1(\text{health\_insurance}) + \beta_2(\text{age}) + \beta_3(\text{race}) + \beta_4(\text{gender}) \\ &+ \beta_5(\text{family\_size}) + \quad \beta_6(\text{education}) + \beta_7(\text{hhtotal\_income}) \\ &+ \beta_8(\text{percent\_life\_US}) + \beta_9(\text{gen\_health\_cond}) + \epsilon\end{aligned} \tag{4}$$

The summary of the fitted model that excludes the instrumental variable employment can be found in Table 5 of the Appendix section in this report. It is important to note that the variable health_insurance in this model is potentially an endogenous variable. Although it is not required, the response variable BMI is transformed using Box-Cox transformation to improve the normality of its distribution. Based on the OLS regression output (Table 5) used as a benchmark, the variable *insurancePublic* shows a positive and significant coefficient of 0.00099390. This suggests that the response variable, BMI, is expected to be higher for individuals with public insurance compared to the reference category, which is private insurance. However, it is important to interpret this result with caution as the endogeneity of health_insurance can lead to potential biases in the analysis. This means that this naive model incorrectly estimates the effect of health_insurance on BMI because of omitted variable bias. To fix the endogeneity problem, an instrument variable can be used to remove the endogeneity from health_insurance and instead use an exogeneity-only version of health_insurance. The variable *employment* is a potential instrument that can address the issue of endogeneity in health_insurance. This instrument meets the criteria outlined in the previous section, making it suitable for addressing the correlation of health_insurance with the error term.

3.1.2. *2SLS*

For this analysis, a multivariate IV regression model is used to estimate the relationship between BMI and the treatment variable health_insurance, while accounting for the control variables defined in the previous section. In this model, employment is treated as an instrumental variable. In order to validate the findings obtained from fitting the IV regression model using *ivreg* from the *AER* package in R, the two stages of 2SLS are also fitted separately. The model shown in (5) is fitted in the first stage to obtain the new treatment variable, which is adjusted using the IV. The model defined in (6) is the final model fitted in stage 2, which uses the predicted values of the health_insurance* obtained from the first stage. These predicted values are used in stage 2 to estimate the causal effect of the treatment on the outcome BMI.

$$\begin{aligned}\text{health\_insurance}^* = {} & \beta_0 + \beta_1(\text{age}) + \beta_2(\text{race}) + \beta_3(\text{gender}) + \beta_4(\text{family\_size}) \\ & + \beta_5(\text{education}) + \beta_6(\text{hhtotal\_income}) + \beta_7(\text{percent\_life\_US}) \\ & + \beta_8(\text{gen\_health\_cond}) + \beta_9(\text{employment}) + \eta\end{aligned} \tag{5}$$

$$\begin{aligned}\text{BMI}^{-1} = {} & \beta_0 + \beta_1(\text{health\_insurance}^*) + \beta_2(\text{age}) + \beta_3(\text{race}) + \beta_4(\text{gender}) \\ & + \beta_5(\text{family\_size}) + \quad \beta_6(\text{education}) + \beta_7(\text{hhtotal\_income}) \\ & + \beta_8(\text{percent\_life\_US}) + \beta_9(\text{gen\_health\_cond}) + \epsilon\end{aligned} \tag{6}$$

The IV regression model (7) is also fitted using the *ivreg* function from the *AER* package in R. In this model, the variables age, race, gender, family size, education, household total income, percentage of life in the US, and general health conditions are considered exogenous regressors. They are assumed to be unaffected by the error term and are used as control variables in the analysis. On the other hand, employment is treated as an instrumental variable, which is used to address the endogeneity of the treatment variable, health insurance. Health insurance variable is considered endogenous, as it is not randomly assigned and is potentially correlated with the error term and influenced by unobserved factors. By using employment as an IV, the goal is to identify the causal effect of health insurance on BMI. Since both methods follow the same approach, the estimated results are nearly identical as expected with a minimal difference between the estimated coefficients and their standard errors.

$$\begin{aligned}\text{BMI}^{-1} = {} & \text{health\_insurance} + \text{age} + \text{race} + \text{gender} \\ & + \text{family\_size} + \text{education} + \text{hhtotal\_income} \\ & + \text{percent\_life\_US} + \text{gen\_health\_cond} \;\mid\; \text{age} + \text{race} \\ & + \text{gender} + \text{family\_size} + \text{education} + \text{hhtotal\_income} \\ & + \text{percent\_life\_US} + \text{gen\_health\_cond} + \text{employment} + \epsilon\end{aligned} \tag{7}$$

The summary output of the stage 1 model presented in Table 6 provides evidence that the instrumental variable employment is relevant for the endogenous variable health_insurance even after controlling for other variables. In this case, it is safe to conclude that the proposed IV employment is correlated with the endogenous variable health_insurance since the coefficient for employment is highly statistical significant with pvalue $< 2e - 16$. This finding is consistent with the results shown in Table 4 discussed in the previous section. Moreover, the summary shown in Table 8 suggests that the variable insurancePublic has a coefficient of 0.004079, which is positive and statistically significant. This indicates that there is a significant association between having public insurance and BMI. The positive coefficient suggests that individuals with public insurance are expected to have higher BMI compared to the reference category, private insurance.

3.2. Sensitivity Analysis

3.2.1. *OLS vs. 2SLS*

By conducting a comparison between the outcomes derived from the 2SLS method using *ivreg* and the benchmark OLS approach, the impact of utilizing an instrumental variable method becomes evident. The model diagnostic plots of the OLS and 2SLS models, depicted in Figures 6 and 7 respectively, exhibit similarities. These plots indicate that neither of these methods exhibit significant deviations from the model assumptions, such as homoskedasticity of errors. As depicted in Figure 2, it is clear that the estimated coefficients tend to be notably higher when employing the instrumental variable, accompanied by slightly wider confidence intervals in comparison to OLS. Table 9 presents the estimated coefficients and corresponding standard errors for both methodologies side by side, where the standard errors are also marginally greater in the 2SLS method compared to OLS.

In Figure 2 below, it is evident that the coefficient for the treatment variable insurancePublic is significantly higher in the 2SLS method compared to OLS. Several factors can contribute to this difference. One possible reason is that the instrumental variable (IV) estimate captures the local average treatment effect (LATE), which represents the impact of the treatment on a specific subgroup of individuals [7]. In contrast, OLS estimates the average treatment effect (ATE) across the entire population. In other words, the OLS estimate describes the average difference in BMI for individuals with different types of health insurance coverage. The higher coefficients observed in 2SLS can be attributed to its ability to address endogeneity issues. Unlike OLS, which is susceptible to bias and underestimates the true causal effect in the presence of endogeneity, 2SLS utilizes instrumental variables to mitigate endogeneity and provide more accurate treatment effect estimates. Consequently, the higher coefficients obtained in 2SLS reflect a closer approximation of the genuine causal relationship between the treatment and the outcome.

Table 10 presents the evaluation results for the OLS and IV regression (2SLS) models using various statistical measures. The OLS model exhibits higher values for both the R-squared and adjusted R-squared compared to 2SLS. Furthermore, the AIC and BIC values are lower for the OLS model than the 2SLS model. The difference in R-squared values between OLS and 2SLS may be due to several factors. One possibility is that the incorporation of instrumental variables in the 2SLS model may introduce additional noise or measurement error, resulting in a lower R-squared value. However, it is important to note that comparing R-squared

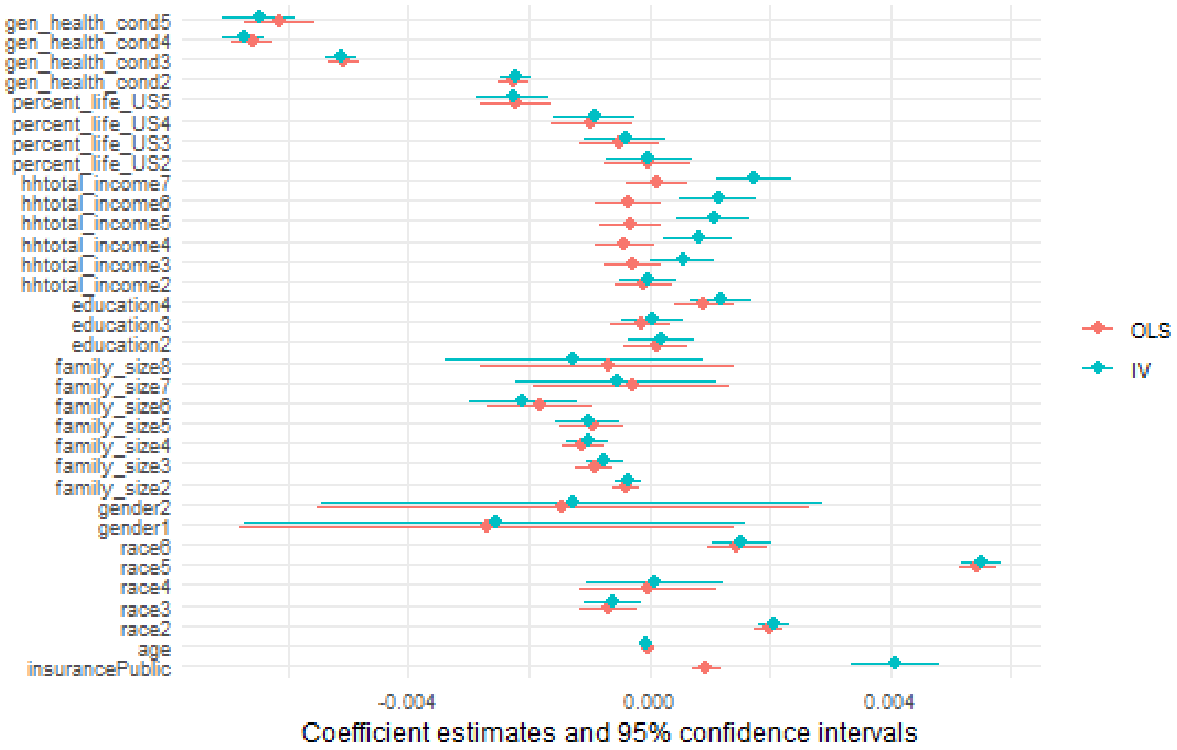


Figure 2. Coefficient Estimates: OLS vs 2SLS

values between OLS and 2SLS should be done cautiously because the methods serve different purposes. OLS aims to estimate the overall relationship between independent and dependent variables, while 2SLS focuses on estimating causal relationships by addressing endogeneity. Despite the differences in R-squared values, it is noteworthy that both the OLS and 2SLS models exhibit comparable levels of predictive accuracy or goodness-of-fit, as evidenced by their equivalent RMSE values. This suggests that both models perform similarly in terms of their ability to make accurate predictions.

3.2.2. *2SLS*

The following assumptions form the foundation of the general IV regression framework and are important to consider when using instrumental variable analysis to estimate causal relationships between variables [8].

1. Exogeneity of control variables: The control variables $W_{1i}, \cdots, W_{ri}$ are exogenous, meaning that their conditional expectation given other variables is zero, i.e., $\mathbb{E}(\epsilon_i | W_{1i}, \cdots, W_{ri}) = 0$. This assumption ensures that the control variables are not correlated with the error term.
2. Independence and identically distributed (i.i.d.) observations: The variables $(X_{1i}, \cdots, X_{ki}; W_{1i}, \cdots, W_{ri}; Z_{1i}, \cdots, Z_{mi})$ are independent and identically distributed draws from their joint distribution. This assumption allows for reliable statistical inference.
3. Finite fourth moments: All variables involved in the regression, including the instrument variables, have nonzero finite fourth moments. This assumption implies that extreme outliers are unlikely and helps ensure the stability of the regression estimates.

4. Valid instrument variables: The instrument variables $Z_{1i}, \cdots, Z_{mi}$ are valid. This means that they are correlated with the endogenous treatment variable but not directly associated with the outcome variable, conditional on other relevant variables. Valid instruments are crucial for obtaining consistent and unbiased estimates of the causal effect.

The *Durbin-Wu-Hausman test*, which compares the estimates from the IV regression to those from the alternative estimator OLS, is conducted to assess the presence of endogeneity in the regression model. The test yields a significant result with a p-value of less than 2e-16, indicating that the null hypothesis, which assumes the endogenous treatment to be exogenous, can be rejected. This implies that there is evidence of endogeneity in the model, leading to different estimates between IV regression and OLS. When all the regressors are exogenous, both the OLS and 2SLS estimators are consistent, and the OLS estimator is more efficient. However, if one or more regressors are endogenous, the OLS estimator becomes inconsistent.

The significant p-value obtained using Wu-Hausman test suggests that the OLS estimates are biased and inconsistent, highlighting the need to rely on the IV estimates for more reliable estimation of the causal relationship between BMI and health insurance. By employing instrumental variables to address endogeneity, the IV regression helps mitigate potential biases and ensures the provision of consistent estimates for the causal effects. Therefore, based on the significant Wu-Hausman test result, it can be concluded that the presence of endogeneity impacts the estimates, making the IV regression a preferable approach for accurately capturing the causal relationship between BMI and health insurance.

Moreover, Table 11 presents the IV regression model fitted using the *ivmodel* package in R. The first stage regression result section shows a large F-statistic, which indicates that the IV is not weak and the 2SLS estimator is reliable. The second section of the table, which includes several k-class estimators, shows the estimated causal effect using the 2SLS estimator is 0.0040793 with a significant p-value. Since there is only one IV in this analysis, 2SLS and LIML show the same results. Overall, the result suggests that there is a significant treatment effect, as indicated by the small p-value from the *Anderson-Rubin test*.

In this analysis, all the control variables are assumed to be exogenous as they are considered predetermined. This implies that these variables are determined independently of the model and prior to the acquisition of the treatment, which in this case is health insurance. For instance, the control variable race is considered exogenous in this analysis as it is not influenced by BMI, health insurance, or employment.

Furthermore, the *ivreg* package in R automatically conducts a diagnostic test for Weak Instruments, which uses the Kleibergen-Paap rk Wald F statistic, when fitting an IV regression model. This test involves an F-test on the instrument in the first stage to assess the strength of the IV. Identifying the presence of weak instruments is crucial as they tend to introduce more bias than they reduce variance. In this analysis, the test yields a large statistic and a significant p-value of less than 2e-16. This indicates that the instrumental variable employment is highly correlated with one or more of the explanatory variables while remaining uncorrelated with the errors. With such a significant p-value, the null hypothesis of weak instruments can be rejected. This finding is consistent with the results of other tests discussed in previous sections, further supporting the conclusion that the IV employment is a valid instrument that meets the criteria for validity. Consequently, assumption 4 holds.

In order to draw causal inferences using IV regression, it is important to have a sufficiently large sample size since estimates may be biased or show large sampling variability if the sample is insufficient [9]. Given that the dataset used in this analysis is sufficiently large, it is reasonable to conclude that this condition is satisfied. The Residuals vs. Fitted values plot included in Figure 7 does not show a clear pattern or curvature. Hence, the linearity assumption is also not violated. Additionally, the Scale-Location plot suggests a minor departure from perfect homoscedasticity, which is addressed by employing heteroskedasticity-robust standard errors. This cautious approach accounts for the slight deviation from the homoscedasticity assumption.

The Residual vs. Leverage plot in Figure 7, obtained from the second stage of the manual implementation of 2SLS, indicates the absence of significant outliers. Therefore, the IV regression model satisfies the assumption of no influential observations, supporting the third assumption. Furthermore, the assumption of independence and identically distributed (i.i.d.) observations, which is commonly applied to continuous or numeric variables, may not be directly applicable to imbalanced categorical variables from an observational study. Since all variables in the IV regression model, except for the response variable, are categorical, the notion of i.i.d. assumption may be less relevant. Instead, it is more appropriate to focus on other assumptions and diagnostic tests that are specifically tailored for categorical variables, such as assessing the independence of categories within each variable or examining the goodness of fit of the categorical model. Thus, explicitly checking the i.i.d. assumption for categorical variables may not be necessary.

4. Discussion

Based on the analysis results, it is evident that the employment variable serves as a valuable instrumental variable for modeling the causal relationship between health insurance coverage and BMI. The instrumental variable approach effectively addresses the issue of endogeneity, enhancing the reliability of the estimated effects. Comparing the results to the benchmark OLS regression, the IV regression shows a significant difference in the estimated coefficient of the treatment variable, indicating the successful mitigation of endogeneity. The findings indicate that the choice of health insurance can have a significant impact on an individual's weight and overall health, which supports the hypothesis.

The IV regression estimates suggest that private insurance may be more effective in managing weight and obesity compared to public health insurance. This finding aligns with several plausible reasons. Private insurance plans typically offer more comprehensive coverage for weight management and obesity-related services, including preventive care, specialized programs, nutrition counseling, and coverage for medications or surgeries. In addition, private insurance plans provide greater flexibility and choice in selecting healthcare providers, including those specialized in obesity management. They also tend to have fewer restrictions and limitations on accessing weight management services. Conversely, public health insurance programs often have stricter eligibility criteria and limited coverage options for obesity-related interventions. Overall, these findings support the notion that private insurance offers more beneficial resources and opportunities for individuals in managing their weight and obesity. While additional research may be required to precisely characterize the effect size, the results suggest that public health insurance plans should prioritize the improvement of their health management programs.

References

[1] U.S. Department of Health and Human Services. (n.d.). Assessing your weight and health risk. National Heart Lung and Blood Institute.

[2] Mylona, E. K., Benitez, G., Shehadeh, F., Fleury, E., Mylonakis, S. C., Kalligeros, M., & Mylonakis, E. (2020). The association of obesity with health insurance coverage and demographic characteristics: a statewide cross-sectional study. *Medicine*, 99(27), e21016. https://doi.org/10.1097/MD.0000000000021016

[3] Baiocchi M, Cheng, J., & Small, D. S. (2014). Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13), 2297–2340. https://doi.org/10.1002/sim.6128.

[4] Penna ND, Stevens JP, Stretch R. Instrumental Variable Analysis of Electronic Health Records. 2016 Sep 10. In: Secondary Analysis of Electronic Health Records [Internet]. Cham (CH): Springer; 2016. Chapter 19. Available from: https://www.ncbi.nlm.nih.gov/books/NBK543623/ doi: 10.1007/978-3-319-43742-2_19

[5] Health risks of overweight & obesity. (2022, November 18). NIDDK - National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/weight-management/adult-overweight-obesity/health-risks

[6] Akil, L., & Ahmad, H. A. (2011). Effects of socioeconomic factors on obesity rates in four southern states and Colorado. Ethnicity & disease, 21(1), 58–62.

[7] Card, D. (1999), The causal effect of education on earnings. In: Ashenfelter, O. C., and D. Card (eds). Handbook of Labor Economics Vol. 3A. Amsterdam: Elsevier, 1801–1863.

[8] Introduction to econometrics with R. (n.d.). Retrieved June 1, 2023, from https://www.econometrics-with-r.org/ivr.html

[9] Maydeu-Olivares, A., Shi, D.,& Rosseel, Y. (2019). Instrumental variables two-stage least squares (2SLS) vs. maximum likelihood structural equation modeling of causal effects in linear regression models. Structural Equation Modeling: A Multidisciplinary Journal, 26(6), 876–892. https://doi.org/10.1080/10705511.2019.1607740

[10] Huntington-Klein, N. (n.d.). Chapter 19 - Instrumental variables. Retrieved June 8, 2023, from https://theeffectbook.net/ch-InstrumentalVariables.html

[11] Foster, E. M. (1997). Instrumental variables for logistic regression: An illustration. Social Science Research, 26(4), 487–504. https://doi.org/10.1006/ssre.1997.0606

[12] Kang, H., Jiang, Y., Zhao, Q., and Small, D.S. (2021). ivmodel: An R Package for Inference and Sensitivity Analysis of Instrumental Variables Models with One Endogenous Variable. Observational Studies 7(2), 1-24. doi:10.1353/obs.2021.0029.

[13] Contributors to Wikimedia projects. (2023, April 5). Instrumental variables estimation. Wikipedia. https://en.wikipedia.org/wiki/Instrumental_variables_estimation

Appendix

Table 1. Descriptive Statistics

	age	bmi	bmi^{-1}
Mean	51.88	27.44	0.04
Std.Dev	17.09	6.14	0.01
Min	18.00	12.94	0.02
Q1	40.00	23.10	0.03
Median	55.00	26.36	0.04
Q3	65.00	30.36	0.04
Max	85.00	65.60	0.08
MAD	22.24	5.22	0.01
IQR	25.00	7.26	0.01
CV	0.33	0.22	0.20
Skewness	-0.18	1.24	0.10
SE.Skewness	0.02	0.02	0.02
Kurtosis	-0.78	2.38	0.02
N.Valid	23405.00	23405.00	23405.00
Pct.Valid	100.00	100.00	100.00

obese	Yes	No
min	25.00	12.94
q1	26.88	20.83
median	29.37	22.47
mean	30.9	22.16
q3	33.29	23.74
max	65.60	29.82

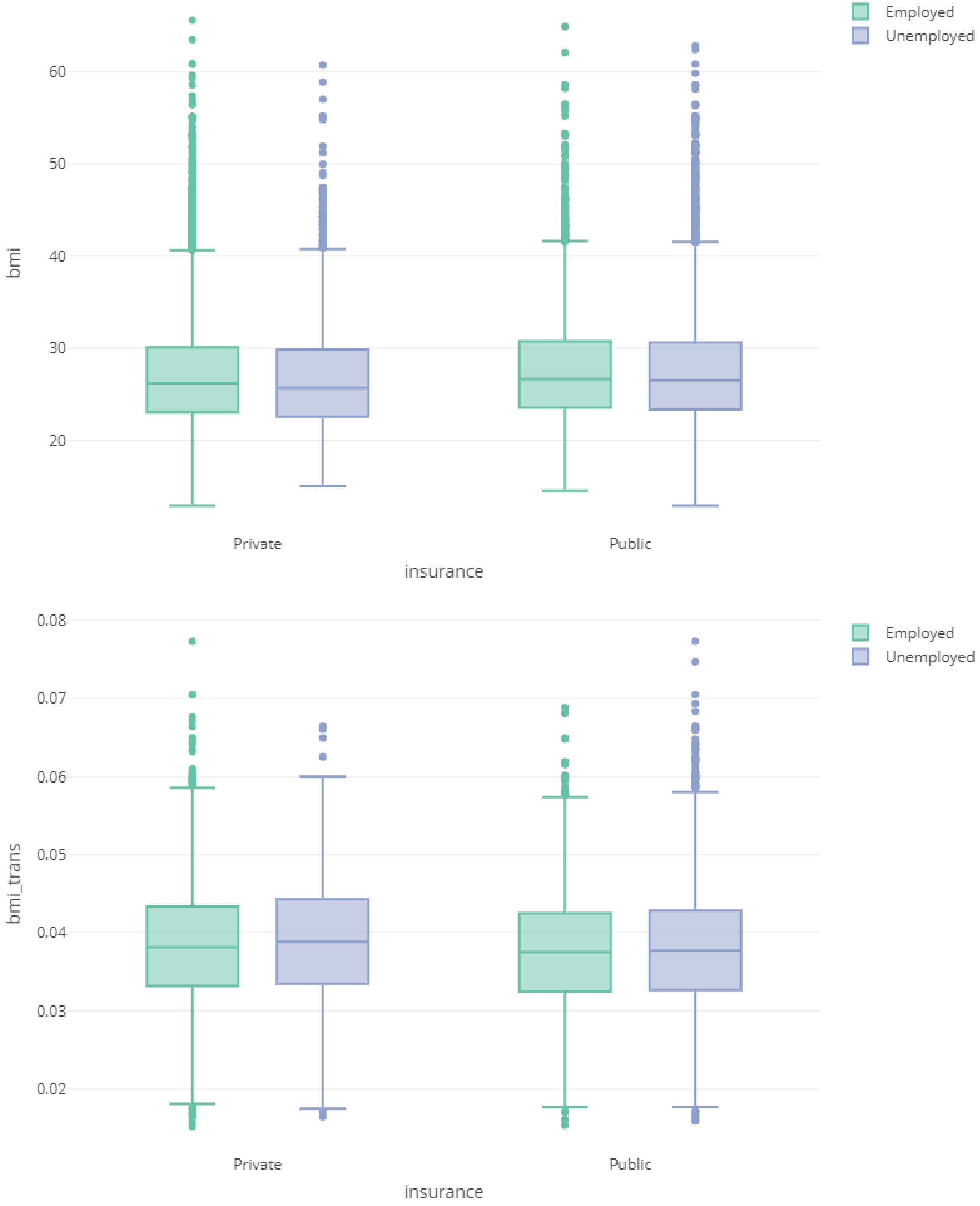


FIGURE 3. Box Plot: Response, Treatment, Instrumental Variables

Table 2. Cross tabulations

Employment	Employed	Unemployed
Insurance		
Private	10942	2347
Public	3008	7108
Chi.squared	df	p.value
6598.769	1	0
Odds Ratio	Lo - 95%	Hi - 95%
11.02	10.36	11.72
Risk Ratio	Lo - 95%	Hi - 95%
2.77	2.68	2.86

Obese	1	2
Insurance		
Private	7849	5440
Public	6291	3825
Chi.squared	df	p.value
23.3193	1	0
Odds Ratio	Lo - 95%	Hi - 95%
0.88	0.83	0.93
Risk Ratio	Lo - 95%	Hi - 95%
0.95	0.93	0.97

Employment	Employed	Unemployed
Obese		
1	8432	5708
2	5518	3747
Chi.squared	df	p.value
0.0101	1	0.92
Odds Ratio	Lo - 95%	Hi - 95%
1.00	0.95	1.06
Risk Ratio	Lo - 95%	Hi - 95%
1.00	0.98	1.02

Table 3. Data Summary

Variable	Stats / Values	Freqs (% of Valid)	NA
bmi [numeric]	Mean (sd) : 27.4 (6.1) min $<$ med $<$ max: 12.9 $<$ 26.4 $<$ 65.6 IQR (CV) : 7.3 (0.2)	1995 distinct values	0
obese [factor]	Yes, No	14140(60.4%), 9265(39.6%)	0
insurance [factor]	Private, Public	13289(56.8%), 10116(43.2%)	0
employment [factor]	Employed, Unemployed	14140(60.4%), 9265(39.6%)	0
education [factor]	$<$ HS education, HS, Some college, College degree & above	941(4.0%), 2704(11.6%), 6475(27.7%), 13285(56.8%)	0
age [numeric]	Mean (sd) : 51.9 (17.1) min $<$ med $<$ max: 18 $<$ 55 $<$ 85 IQR (CV) : 25 (0.3)	14 distinct values	0
citizenship [factor]	US-Born, Naturalized, Non-citizen	17590(75.2%), 4297(18.4%), 1518(6.5%)	0
race [factor]	Hispanic, White (Non-hispanic), African American (NH), American Indian (NH), Asian (NH), Other (2 or more races)	5491(23.5%), 12037(51.4%), 1006(4.3%), 148(0.6%), 3806(16.3%), 917(3.9%)	0
percent_life_US [factor]	0-20, 21-40, 41-60, 61-80, 81+	589(2.5%), 982(4.2%), 1593(6.8%), 1522(6.5%), 18719(80.0%)	0
hhtotal_income [factor]	< 10000, $10,000 - 39,999$, $40,000 - 69,999$, $79,000 - 99,999$, $100,000 - 129,000$, $130,000 - 159,000$, $160,000+$	1006(4.3%), 4878(20.8%) 4120(17.6%), 3650(15.6%), 2993(12.8%), 2019(8.%), 4739(20.2%)	0
gen_health_cond [factor]	Excellent, Very good, Good, Fair, Poor	4195(17.9%), 8499(36.3%) 7276(31.1%), 2796(11.9%), 639(2.7%)	0
family_size [factor]	1, 2, 3, 4, 5, 6, 7, 8	9215(39.4%), 8127(34.7%), 2821(12.1%), 2142(9.2%) 739(3.2%), 249(1.1%), 70(0.3%), 42(0.2%)	0
gender [factor]	No response,Male,Female	11(0.0%), 10103(43.2%), 13291(56.8%)	0
bmi_trans [numeric]	Mean (sd) : 0 (0) min $<$ med $<$ max: 0 $<$ 0 $<$ 0.1 IQR (CV) : 0 (0.2)	1995 distinct values	0

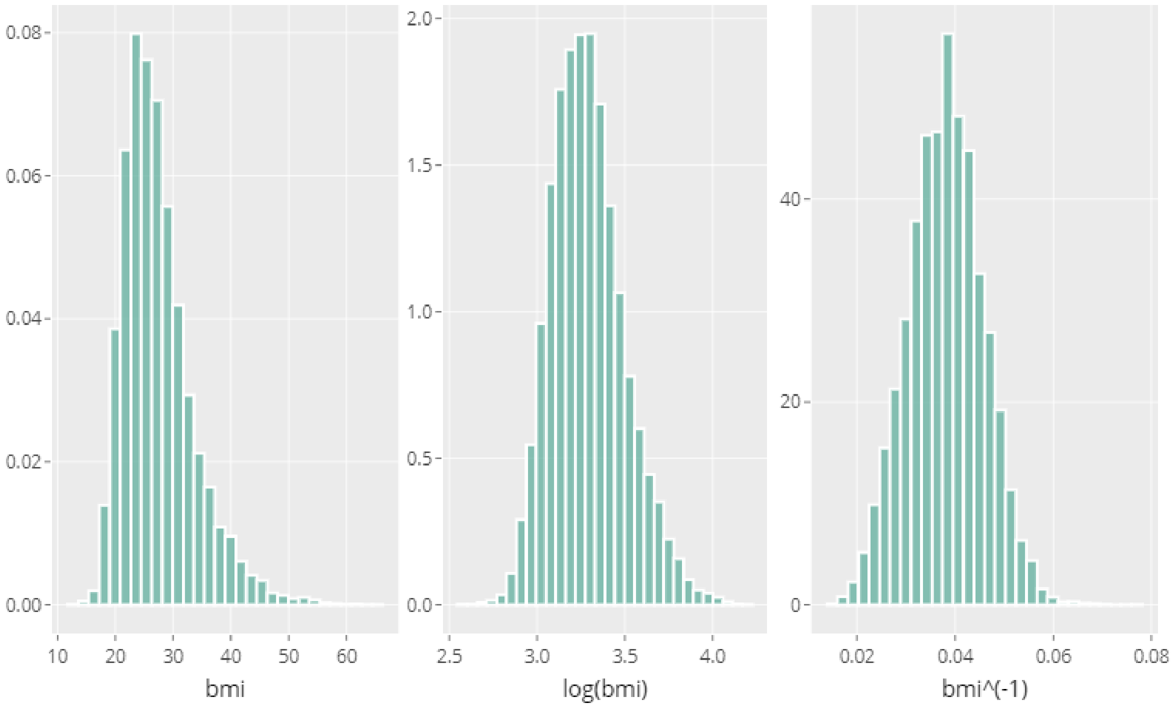


FIGURE 4. Histogram of BMI

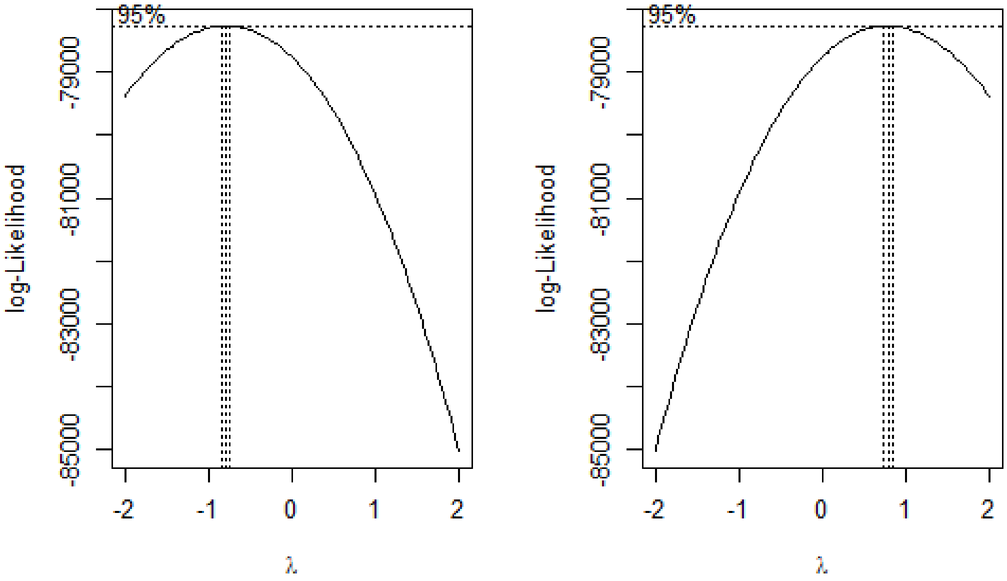


FIGURE 5. Box-cox Plot: BMI (left) and BMI^{-1} (right)

Table 4. Proposed IVs (* = Chi-square test, ** = Kruskal-wallis test, ***= ANOVA)

Variable	Treatment (Insurance Coverage)*	Outcome (BMI)**	Outcome (BMI)***
employment	p-value < 2.2e-16	p-value = 0.6951	p-value = 0.275
family_size	p-value < 2.2e-16	p-value = 0.003808	p-value = 0.00471
education	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16
hhtotal_income	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16
citizenship	p-value = 0.001534	p-value < 2.2e-16	p-value < 2.2e-16
percent_life_US	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16
gen_health_cond	p-value < 2.2e-16	p-value < 2.2e-16	p-value < 2.2e-16

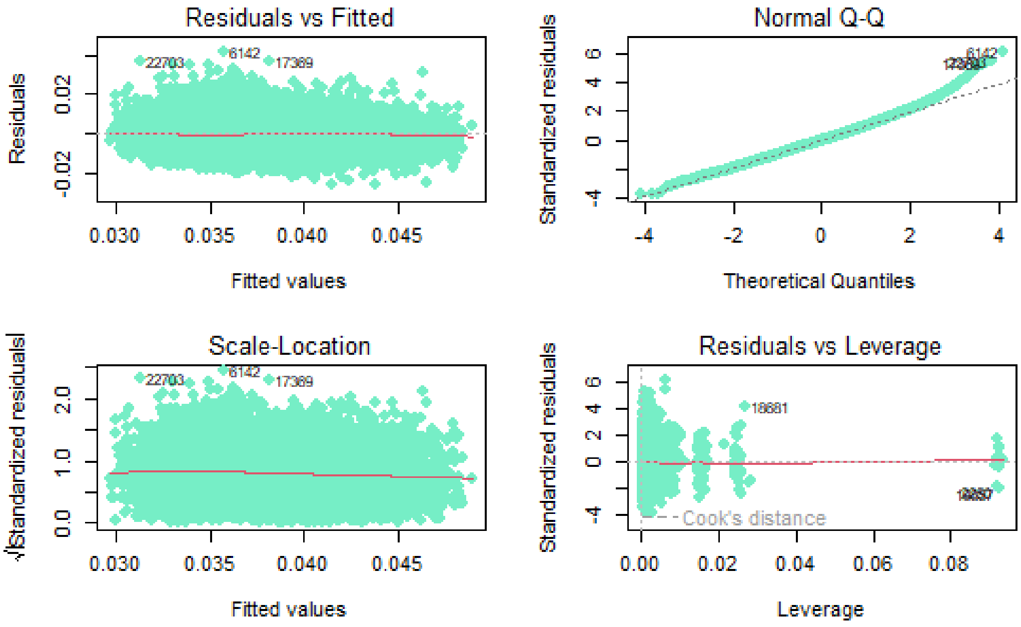


Figure 6. Model Disagnostics Plot: OLS

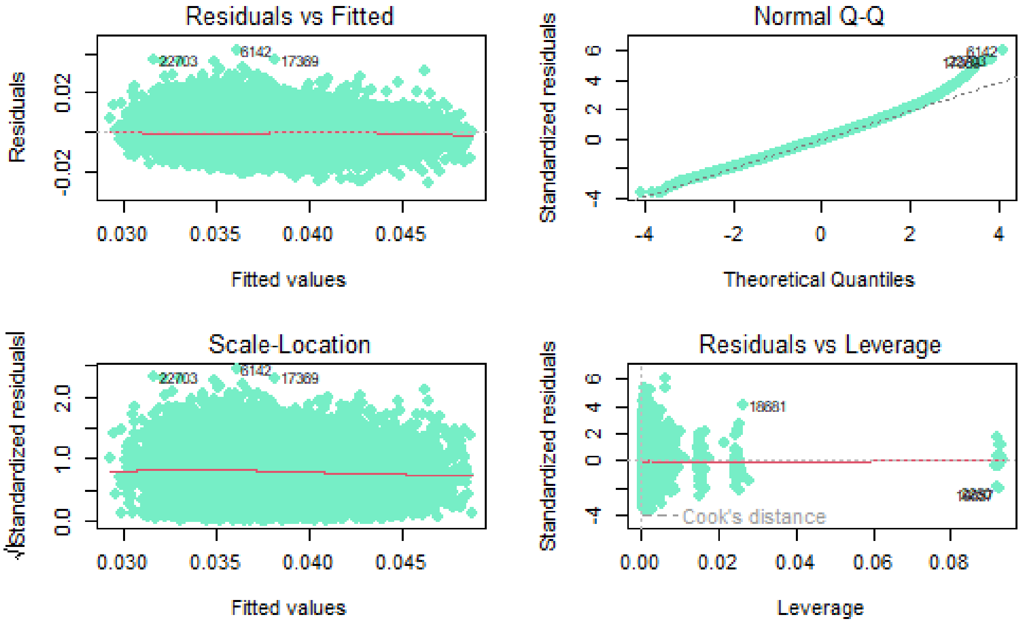
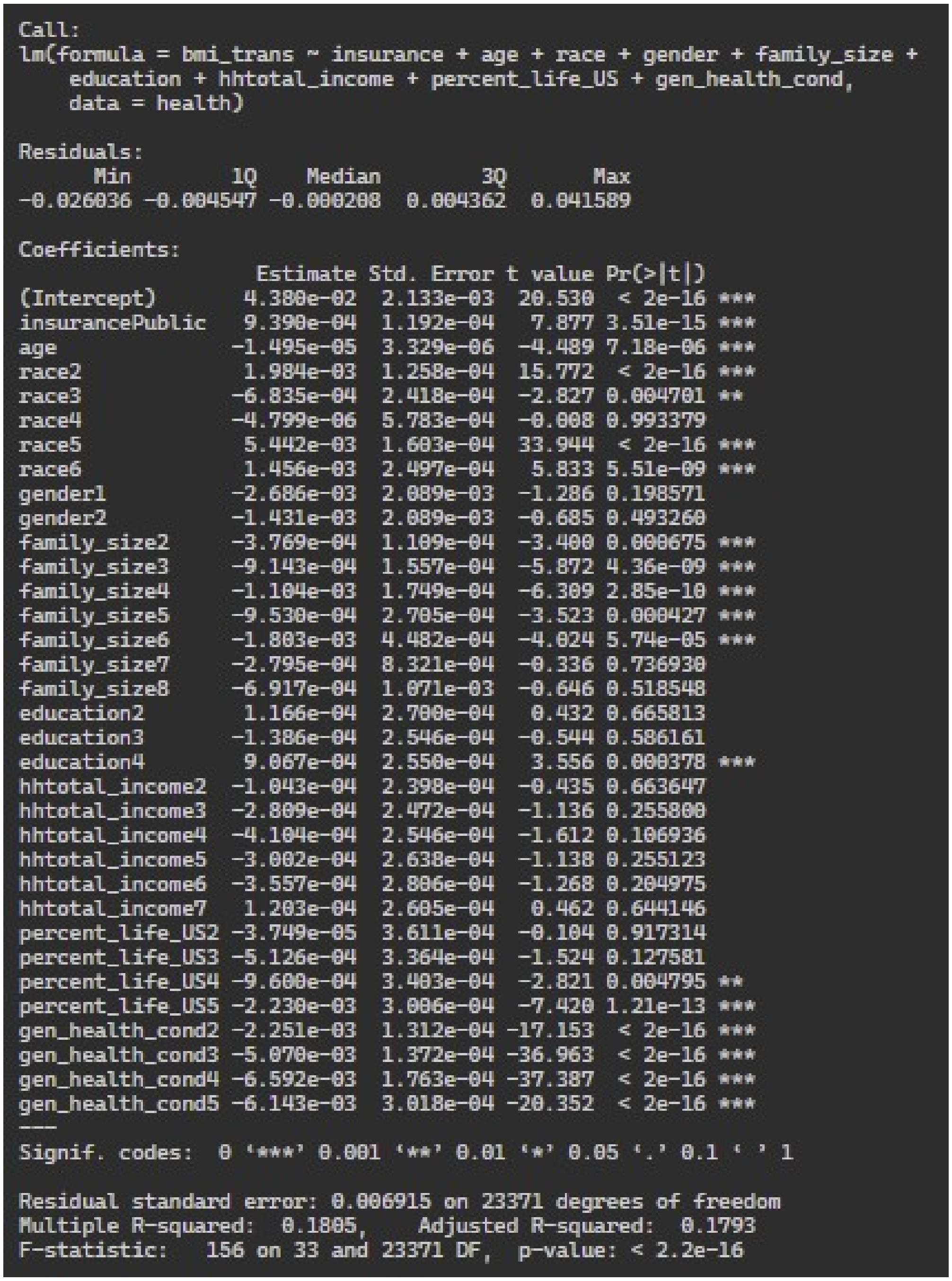


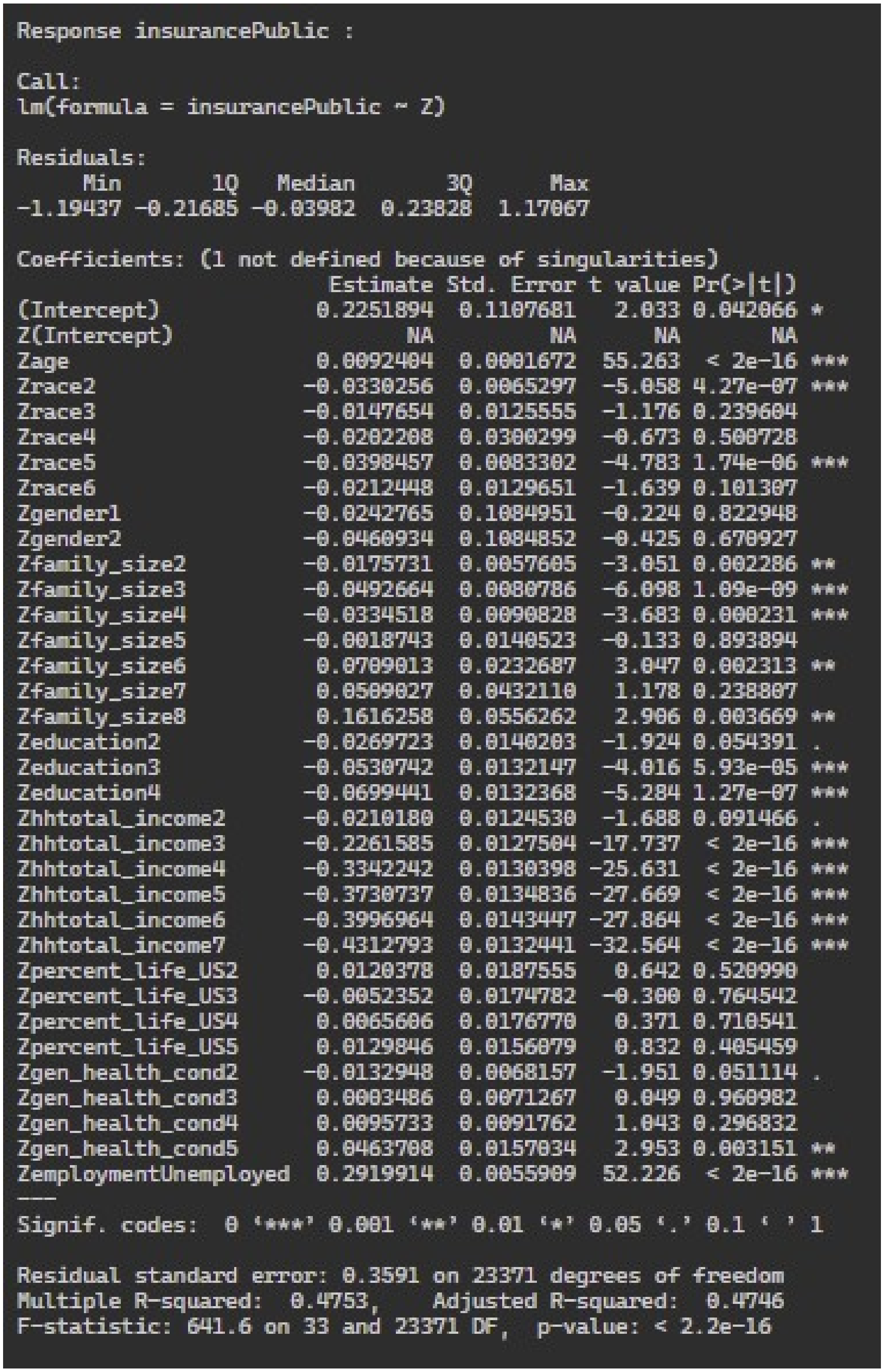
Figure 7. Model Disagnostics Plot: 2SLS Stage 2

Table 5. OLS Model Summary



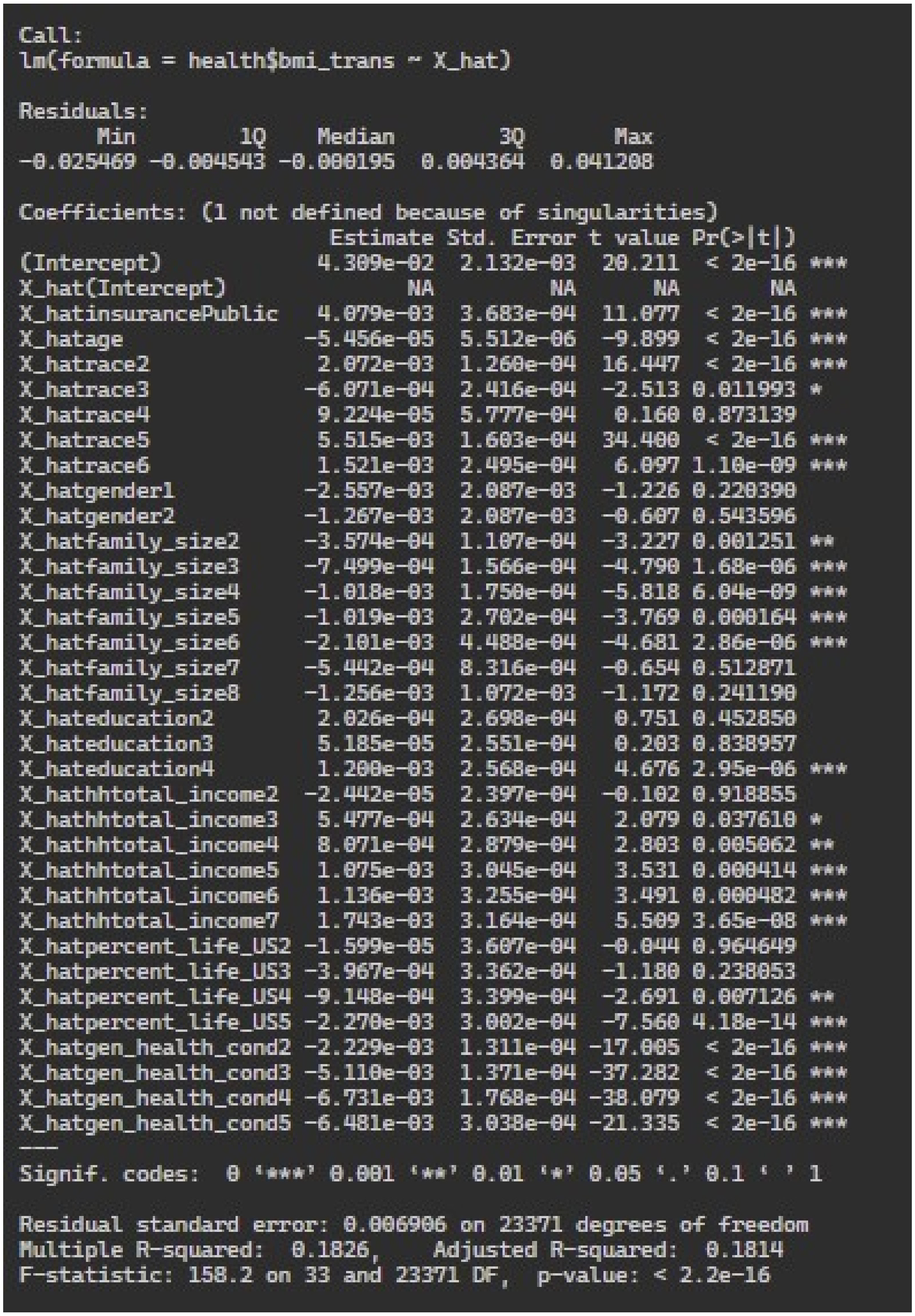
```
Call:
lm(formula = bmi_trans ~ insurance + age + race + gender + family_size +
    education + hhtotal_income + percent_life_US + gen_health_cond,
    data = health)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.026036 -0.004547 -0.000208  0.004362  0.041589 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4.380e-02  2.133e-03  20.530  < 2e-16 ***
insurancePublic   9.390e-04  1.192e-04   7.877 3.51e-15 ***
age              -1.495e-05  3.329e-06  -4.489 7.18e-06 ***
race2             1.984e-03  1.258e-04  15.772  < 2e-16 ***
race3            -6.835e-04  2.418e-04  -2.827 0.004701 ** 
race4            -4.799e-06  5.783e-04  -0.008 0.993379    
race5             5.442e-03  1.603e-04  33.944  < 2e-16 ***
race6             1.456e-03  2.497e-04   5.833 5.51e-09 ***
gender1          -2.686e-03  2.089e-03  -1.286 0.198571    
gender2          -1.431e-03  2.089e-03  -0.685 0.493260    
family_size2     -3.769e-04  1.109e-04  -3.400 0.000675 ***
family_size3     -9.143e-04  1.557e-04  -5.872 4.36e-09 ***
family_size4     -1.104e-03  1.749e-04  -6.309 2.85e-10 ***
family_size5     -9.530e-04  2.705e-04  -3.523 0.000427 ***
family_size6     -1.803e-03  4.482e-04  -4.024 5.74e-05 ***
family_size7     -2.795e-04  8.321e-04  -0.336 0.736930    
family_size8     -6.917e-04  1.071e-03  -0.646 0.518548    
education2        1.166e-04  2.700e-04   0.432 0.665813    
education3       -1.386e-04  2.546e-04  -0.544 0.586161    
education4        9.067e-04  2.550e-04   3.556 0.000378 ***
hhtotal_income2  -1.043e-04  2.398e-04  -0.435 0.663647    
hhtotal_income3  -2.809e-04  2.472e-04  -1.136 0.255800    
hhtotal_income4  -4.104e-04  2.546e-04  -1.612 0.106936    
hhtotal_income5  -3.002e-04  2.638e-04  -1.138 0.255123    
hhtotal_income6  -3.557e-04  2.806e-04  -1.268 0.204975    
hhtotal_income7   1.203e-04  2.605e-04   0.462 0.644146    
percent_life_US2 -3.749e-05  3.611e-04  -0.104 0.917314    
percent_life_US3 -5.126e-04  3.364e-04  -1.524 0.127581    
percent_life_US4 -9.600e-04  3.403e-04  -2.821 0.004795 ** 
percent_life_US5 -2.230e-03  3.006e-04  -7.420 1.21e-13 ***
gen_health_cond2 -2.251e-03  1.312e-04 -17.153  < 2e-16 ***
gen_health_cond3 -5.070e-03  1.372e-04 -36.963  < 2e-16 ***
gen_health_cond4 -6.592e-03  1.763e-04 -37.387  < 2e-16 ***
gen_health_cond5 -6.143e-03  3.018e-04 -20.352  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.006915 on 23371 degrees of freedom
Multiple R-squared:  0.1805,	Adjusted R-squared:  0.1793 
F-statistic:   156 on 33 and 23371 DF,  p-value: < 2.2e-16
```

Table 6. Manual 2SLS Model Summary: Stage 1



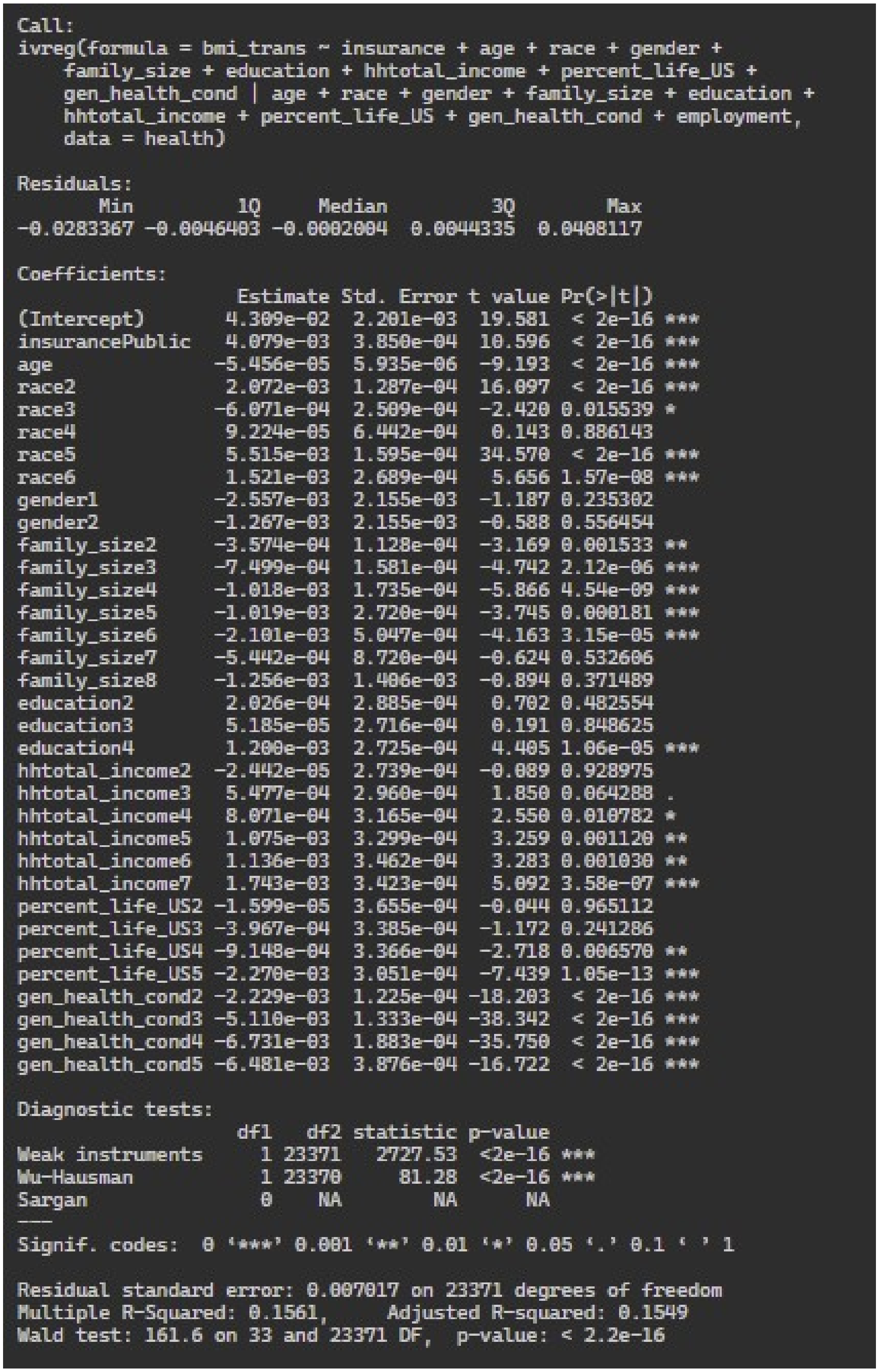
```
Response insurancePublic :

Call:
lm(formula = insurancePublic ~ Z)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.19437 -0.21685 -0.03982  0.23828  1.17067 

Coefficients: (1 not defined because of singularities)
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)            0.2251894  0.1107681   2.033 0.042066 *  
Z(Intercept)                  NA         NA      NA       NA    
Zage                   0.0092404  0.0001672  55.263  < 2e-16 ***
Zrace2                -0.0330256  0.0065297  -5.058 4.27e-07 ***
Zrace3                -0.0147654  0.0125555  -1.176 0.239604    
Zrace4                -0.0202208  0.0300299  -0.673 0.500728    
Zrace5                -0.0398457  0.0083302  -4.783 1.74e-06 ***
Zrace6                -0.0212448  0.0129651  -1.639 0.101307    
Zgender1              -0.0242765  0.1084951  -0.224 0.822948    
Zgender2              -0.0460934  0.1084852  -0.425 0.670927    
Zfamily_size2         -0.0175731  0.0057605  -3.051 0.002286 ** 
Zfamily_size3         -0.0492664  0.0080786  -6.098 1.09e-09 ***
Zfamily_size4         -0.0334518  0.0090828  -3.683 0.000231 ***
Zfamily_size5         -0.0018743  0.0140523  -0.133 0.893894    
Zfamily_size6          0.0709013  0.0232687   3.047 0.002313 ** 
Zfamily_size7          0.0509027  0.0432110   1.178 0.238807    
Zfamily_size8          0.1616258  0.0556262   2.906 0.003669 ** 
Zeducation2           -0.0269723  0.0140203  -1.924 0.054391 .  
Zeducation3           -0.0530742  0.0132147  -4.016 5.93e-05 ***
Zeducation4           -0.0699441  0.0132368  -5.284 1.27e-07 ***
Zhhtotal_income2      -0.0210180  0.0124530  -1.688 0.091466 .  
Zhhtotal_income3      -0.2261585  0.0127504 -17.737  < 2e-16 ***
Zhhtotal_income4      -0.3342242  0.0130398 -25.631  < 2e-16 ***
Zhhtotal_income5      -0.3730737  0.0134836 -27.669  < 2e-16 ***
Zhhtotal_income6      -0.3996964  0.0143447 -27.864  < 2e-16 ***
Zhhtotal_income7      -0.4312793  0.0132441 -32.564  < 2e-16 ***
Zpercent_life_US2      0.0120378  0.0187555   0.642 0.520990    
Zpercent_life_US3     -0.0052352  0.0174782  -0.300 0.764542    
Zpercent_life_US4      0.0065606  0.0176770   0.371 0.710541    
Zpercent_life_US5      0.0129846  0.0156079   0.832 0.405459    
Zgen_health_cond2     -0.0132948  0.0068157  -1.951 0.051114 .  
Zgen_health_cond3      0.0003486  0.0071267   0.049 0.960982    
Zgen_health_cond4      0.0095733  0.0091762   1.043 0.296832    
Zgen_health_cond5      0.0463708  0.0157034   2.953 0.003151 ** 
ZemploymentUnemployed  0.2919914  0.0055909  52.226  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3591 on 23371 degrees of freedom
Multiple R-squared:  0.4753,	Adjusted R-squared:  0.4746 
F-statistic: 641.6 on 33 and 23371 DF,  p-value: < 2.2e-16
```

Table 7. Manual 2SLS Model Summary: Stage 2



```
Call:
lm(formula = health$bmi_trans ~ X_hat)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.025469 -0.004543 -0.000195  0.004364  0.041208 

Coefficients: (1 not defined because of singularities)
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)             4.309e-02  2.132e-03  20.211  < 2e-16 ***
X_hat(Intercept)               NA         NA      NA       NA    
X_hatinsurancePublic    4.079e-03  3.683e-04  11.077  < 2e-16 ***
X_hatage               -5.456e-05  5.512e-06  -9.899  < 2e-16 ***
X_hatrace2              2.072e-03  1.260e-04  16.447  < 2e-16 ***
X_hatrace3             -6.071e-04  2.416e-04  -2.513 0.011993 *  
X_hatrace4              9.224e-05  5.777e-04   0.160 0.873139    
X_hatrace5              5.515e-03  1.603e-04  34.400  < 2e-16 ***
X_hatrace6              1.521e-03  2.495e-04   6.097 1.10e-09 ***
X_hatgender1           -2.557e-03  2.087e-03  -1.226 0.220390    
X_hatgender2           -1.267e-03  2.087e-03  -0.607 0.543596    
X_hatfamily_size2      -3.574e-04  1.107e-04  -3.227 0.001251 ** 
X_hatfamily_size3      -7.499e-04  1.566e-04  -4.790 1.68e-06 ***
X_hatfamily_size4      -1.018e-03  1.750e-04  -5.818 6.04e-09 ***
X_hatfamily_size5      -1.019e-03  2.702e-04  -3.769 0.000164 ***
X_hatfamily_size6      -2.101e-03  4.488e-04  -4.681 2.86e-06 ***
X_hatfamily_size7      -5.442e-04  8.316e-04  -0.654 0.512871    
X_hatfamily_size8      -1.256e-03  1.072e-03  -1.172 0.241190    
X_hateducation2         2.026e-04  2.698e-04   0.751 0.452850    
X_hateducation3         5.185e-05  2.551e-04   0.203 0.838957    
X_hateducation4         1.200e-03  2.568e-04   4.676 2.95e-06 ***
X_hathhtotal_income2   -2.442e-05  2.397e-04  -0.102 0.918855    
X_hathhtotal_income3    5.477e-04  2.634e-04   2.079 0.037610 *  
X_hathhtotal_income4    8.071e-04  2.879e-04   2.803 0.005062 ** 
X_hathhtotal_income5    1.075e-03  3.045e-04   3.531 0.000414 ***
X_hathhtotal_income6    1.136e-03  3.255e-04   3.491 0.000482 ***
X_hathhtotal_income7    1.743e-03  3.164e-04   5.509 3.65e-08 ***
X_hatpercent_life_US2 -1.599e-05  3.607e-04  -0.044 0.964649    
X_hatpercent_life_US3 -3.967e-04  3.362e-04  -1.180 0.238053    
X_hatpercent_life_US4 -9.148e-04  3.399e-04  -2.691 0.007126 ** 
X_hatpercent_life_US5 -2.270e-03  3.002e-04  -7.560 4.18e-14 ***
X_hatgen_health_cond2 -2.229e-03  1.311e-04 -17.005  < 2e-16 ***
X_hatgen_health_cond3 -5.110e-03  1.371e-04 -37.282  < 2e-16 ***
X_hatgen_health_cond4 -6.731e-03  1.768e-04 -38.079  < 2e-16 ***
X_hatgen_health_cond5 -6.481e-03  3.038e-04 -21.335  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.006906 on 23371 degrees of freedom
Multiple R-squared:  0.1826,    Adjusted R-squared:  0.1814 
F-statistic: 158.2 on 33 and 23371 DF,  p-value: < 2.2e-16
```

Table 8. Multivariate 2SLS (*ivreg*) Model Summary: Heteroskedasticity-robust S.E.



```
Call:
ivreg(formula = bmi_trans ~ insurance + age + race + gender +
    family_size + education + hhtotal_income + percent_life_US +
    gen_health_cond | age + race + gender + family_size + education +
    hhtotal_income + percent_life_US + gen_health_cond + employment,
    data = health)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0283367 -0.0046403 -0.0002004  0.0044335  0.0408117 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4.309e-02  2.201e-03  19.581  < 2e-16 ***
insurancePublic   4.079e-03  3.850e-04  10.596  < 2e-16 ***
age              -5.456e-05  5.935e-06  -9.193  < 2e-16 ***
race2             2.072e-03  1.287e-04  16.097  < 2e-16 ***
race3            -6.071e-04  2.509e-04  -2.420 0.015539 *  
race4             9.224e-05  6.442e-04   0.143 0.886143    
race5             5.515e-03  1.595e-04  34.570  < 2e-16 ***
race6             1.521e-03  2.689e-04   5.656 1.57e-08 ***
gender1          -2.557e-03  2.155e-03  -1.187 0.235302    
gender2          -1.267e-03  2.155e-03  -0.588 0.556454    
family_size2     -3.574e-04  1.128e-04  -3.169 0.001533 ** 
family_size3     -7.499e-04  1.581e-04  -4.742 2.12e-06 ***
family_size4     -1.018e-03  1.735e-04  -5.866 4.54e-09 ***
family_size5     -1.019e-03  2.720e-04  -3.745 0.000181 ***
family_size6     -2.101e-03  5.047e-04  -4.163 3.15e-05 ***
family_size7     -5.442e-04  8.720e-04  -0.624 0.532606    
family_size8     -1.256e-03  1.406e-03  -0.894 0.371489    
education2        2.026e-04  2.885e-04   0.702 0.482554    
education3        5.185e-05  2.716e-04   0.191 0.848625    
education4        1.200e-03  2.725e-04   4.405 1.06e-05 ***
hhtotal_income2  -2.442e-05  2.739e-04  -0.089 0.928975    
hhtotal_income3   5.477e-04  2.960e-04   1.850 0.064288 .  
hhtotal_income4   8.071e-04  3.165e-04   2.550 0.010782 *  
hhtotal_income5   1.075e-03  3.299e-04   3.259 0.001120 ** 
hhtotal_income6   1.136e-03  3.462e-04   3.283 0.001030 ** 
hhtotal_income7   1.743e-03  3.423e-04   5.092 3.58e-07 ***
percent_life_US2 -1.599e-05  3.655e-04  -0.044 0.965112    
percent_life_US3 -3.967e-04  3.385e-04  -1.172 0.241286    
percent_life_US4 -9.148e-04  3.366e-04  -2.718 0.006570 ** 
percent_life_US5 -2.270e-03  3.051e-04  -7.439 1.05e-13 ***
gen_health_cond2 -2.229e-03  1.225e-04 -18.203  < 2e-16 ***
gen_health_cond3 -5.110e-03  1.333e-04 -38.342  < 2e-16 ***
gen_health_cond4 -6.731e-03  1.883e-04 -35.750  < 2e-16 ***
gen_health_cond5 -6.481e-03  3.876e-04 -16.722  < 2e-16 ***

Diagnostic tests:
                   df1   df2 statistic p-value    
Weak instruments     1 23371   2727.53  <2e-16 ***
Wu-Hausman           1 23370     81.28  <2e-16 ***
Sargan               0    NA        NA      NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.007017 on 23371 degrees of freedom
Multiple R-Squared: 0.1561,     Adjusted R-squared: 0.1549 
Wald test: 161.6 on 33 and 23371 DF,  p-value: < 2.2e-16
```

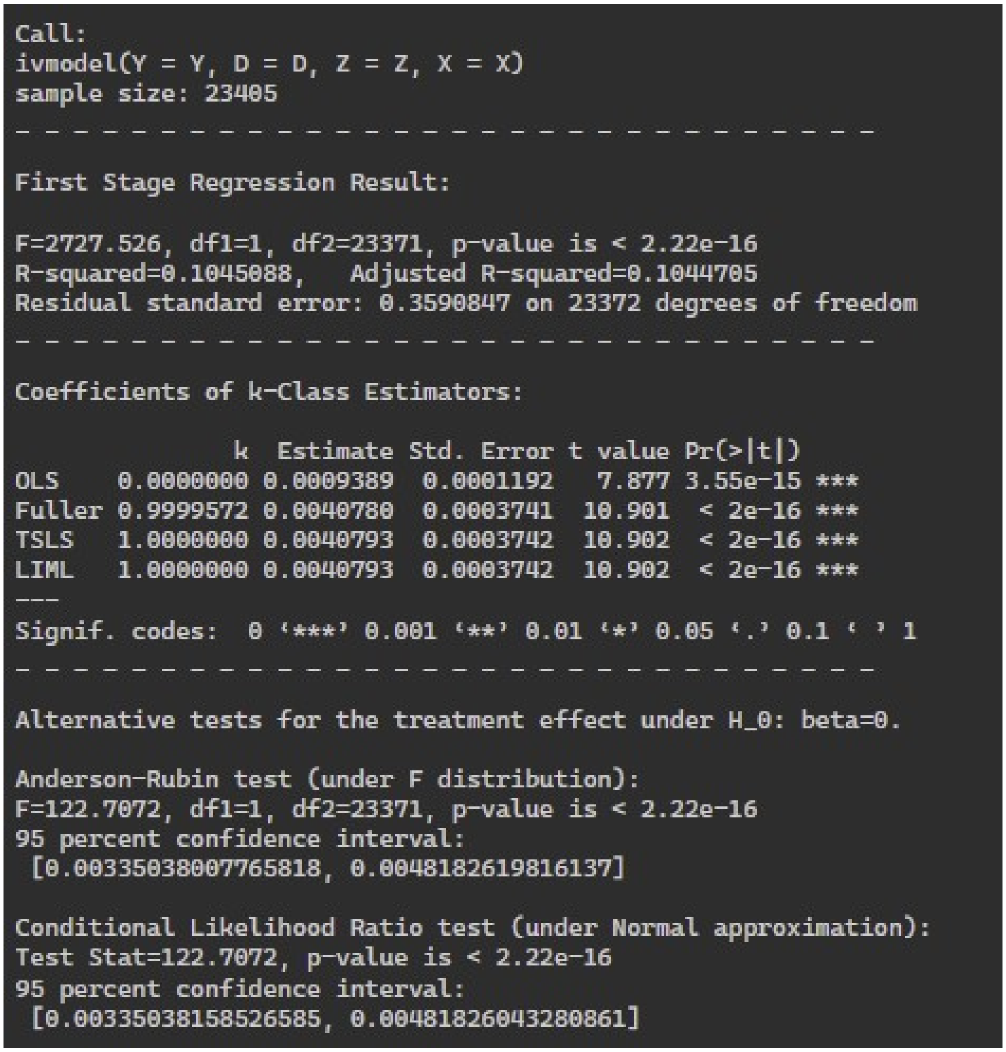
Table 9. Estimated Coefficients (Standard Errors): OLS vs. 2SLS

	OLS	2SLS
(Intercept)	0.04380 (0.00213)	0.04309 (0.00217)
insurancePublic	0.000939 (0.000119)	0.004079 (0.000374)
age	-1.49e-05 (3.33e-06)	-5.46e-05 (5.60e-06)
race2	0.001984 (0.000126)	0.002072 (0.000128)
race3	-0.000684 (0.000242)	-0.000607 (0.000245)
race4	-4.80e-06 (5.78e-04)	9.22e-05 (5.87e-04)
race5	0.005442 (0.000160)	0.005515 (0.000163)
race6	0.001456 (0.000250)	0.001521 (0.000253)
gender1	-0.00269 (0.00209)	-0.00256 (0.00212)
gender2	-0.00143 (0.00209)	-0.00127 (0.00212)
family_size2	-0.000377 (0.000111)	-0.000357 (0.000113)
family_size3	-0.000914 (0.000156)	-0.000750 (0.000159)
family_size4	-0.001104 (0.000175)	-0.001018 (0.000178)
family_size5	-0.000953 (0.000270)	-0.001019 (0.000275)
family_size6	-0.001803 (0.000448)	-0.002101 (0.000456)
family_size7	-0.000280 (0.000832)	-0.000544 (0.000845)
family_size8	-0.000692 (0.001071)	-0.001256 (0.001089)
education2	0.000117 (0.000270)	0.000203 (0.000274)
education3	-1.39e-04 (2.55e-04)	5.18e-05 (2.59e-04)
education4	0.000907 (0.000255)	0.001200 (0.000261)
hhtotal_income2	-1.04e-04 (2.40e-04)	-2.44e-05 (2.44e-04)
hhtotal_income3	-0.000281 (0.000247)	0.000548 (0.000268)
hhtotal_income4	-0.000410 (0.000255)	0.000807 (0.000293)
hhtotal_income5	-0.000300 (0.000264)	0.001075 (0.000309)
hhtotal_income6	-0.000356 (0.000281)	0.001136 (0.000331)
hhtotal_income7	0.000120 (0.000260)	0.001743 (0.000321)
percent_life_US2	-3.75e-05 (3.61e-04)	-1.60e-05 (3.67e-04)
percent_life_US3	-0.000513 (0.000336)	-0.000397 (0.000342)
percent_life_US4	-0.000960 (0.000340)	-0.000915 (0.000345)
percent_life_US5	-0.002230 (0.000301)	-0.002270 (0.000305)
gen_health_cond2	-0.002251 (0.000131)	-0.002229 (0.000133)
gen_health_cond3	-0.005070 (0.000137)	-0.005110 (0.000139)
gen_health_cond4	-0.006592 (0.000176)	-0.006731 (0.000180)
gen_health_cond5	-0.006143 (0.000302)	-0.006481 (0.000309)

Table 10. Model Evaluation: OLS vs. 2SLS (ivreg)

	OLS	2SLS
R^2	0.180	0.156
adjusted R^2	0.179	0.155
AIC	−166378.4	−165693.5
BIC	−166096.3	−165411.3
Log-likelihood	83224.190	
F	155.954	
RMSE	0.01	0.01

Table 11. IV Model Summary (*ivmodel*)



```
Call:
ivmodel(Y = Y, D = D, Z = Z, X = X)
sample size: 23405
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

First Stage Regression Result:

F=2727.526, df1=1, df2=23371, p-value is < 2.22e-16
R-squared=0.1045088,   Adjusted R-squared=0.1044705
Residual standard error: 0.3590847 on 23372 degrees of freedom
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Coefficients of k-Class Estimators:

               k  Estimate Std. Error t value Pr(>|t|)
OLS    0.0000000 0.0009389  0.0001192   7.877 3.55e-15 ***
Fuller 0.9999572 0.0040780  0.0003741  10.901  < 2e-16 ***
TSLS   1.0000000 0.0040793  0.0003742  10.902  < 2e-16 ***
LIML   1.0000000 0.0040793  0.0003742  10.902  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Alternative tests for the treatment effect under H_0: beta=0.

Anderson-Rubin test (under F distribution):
F=122.7072, df1=1, df2=23371, p-value is < 2.22e-16
95 percent confidence interval:
 [0.00335038007765818, 0.0048182619816137]

Conditional Likelihood Ratio test (under Normal approximation):
Test Stat=122.7072, p-value is < 2.22e-16
95 percent confidence interval:
 [0.00335038158526585, 0.00481826043280861]
```

Code

https://github.com/kayannr/health_insurance_IV/blob/main/IV_RCode_STA250KR.Rmd