OLS and Two-Stage Least Squares Regression Examples

# Using the Instrumental Variables (IV) Health Data Set compare the OLS regression solution(s) to Two-Stage Least Squares (TSLS) regression solutions

The Medical Expenditure Panel Survey (MEPS), part of the Agency for Healthcare Research and Quality is an enormous endeavor. The main home page is at <https://meps.ahrq.gov/mepsweb/index.jsp> with many additional webpages flowing from that. There is also a github site at <https://github.com/HHS-AHRQ/MEPS>. I browsed these website for a bit but there is so much information that I didn’t get too far.

I did a Boolean search on “instrumental variables” and got a few hits including a couple working papers. I am attaching one of the working papers involving a discussion of instrumental variables to your assignment. This working paper is “Estimated Effects of Increased Coverage on Prescription Drug Expenditures among Seniors” by Sing et al. I am also attaching two papers referenced by this working paper, “Prescription Drug Insurance and Its Effect on Utilization and Health of the Elderly” by Kahn, et al and “Estimating the Effects of Prescription Drug Coverage for Medicare Beneficiaries” by Shea, et al. The researchers in these two papers seem to have different outcomes from using instrumental variables with Shea, et al being more positive.

A quote from Kahn, et al’s paper says,

*We also experimented with instrumental variables approach. However, the instruments (Medicaid eligibility thresholds, managed care payment rates, employer characteristics)* ***while statistically significant in first stage, were weak and second stage estimates were too imprecise to be informative****. (Kahn etal, 2007)*

A quote from Shea et al’s paper says,

*Third, although we find some evidence of selection based on nonobservable factors (the RESID results), the estimated magnitude of the effect is small. As a consequence,* ***we find a close correspondence in estimates of moral hazard in drug demand between models with strong covariate control for observable health characteristics and those with residuals derived from instrumental variable estimators****. This finding is important from a policy perspective as it suggests that the most important factors underlying the selection process can be captured through information in beneficiaries’ Medicare claims files, and to that extent can be managed through risk adjustment. (Shea et al, 2007)*

Needless to say, for our assignment we are only using a very small part of the data that is available about healthcare and healthcare costs. I am providing scripts for this assignment using both gretl and R/RStudio. The R/RStudio script uses the AER package for R. Take time to compare the results of these scripts. The OLS regression results for both gretl and R/RStudio are very close but not exact. The TSLS results from R (either from the AER package command igreg or from the step-by-step) as well as the “tsls” and step-by-step output from gretl are very close for the second stages. I’ve also included a next step in the R/RStudio script, i.e. to use instrumental variables in a system of equations rather than in a single equation.

## Conduct an OLS regression

1. Using the variable logmedexpense (the log of medical expenses) as the dependent variable and healthinsu (health insurance), illnesses, age and logincome (the log of income) as independent variables, conduct an OLS regression. Use the results to answer the following questions. (Since we have taken the log of both medical expense and income we can interpret the results as percentages.)

How much more (or less depending on the sign) as a percentage are medical expenses for those with health insurance?

## Conduct a TSLS regression

1. Using the same variables plus ssiratio (social security income ratio) as an instrument, conduct a two-stage least squares (TSLS) regression. Use the results of this regression to answer the following questions. (You might want to put your OLS and TSLS results in a table or Excel to compare them more easily.)
   1. For the TSLS regression how much more (or less depending on the sign) as a percentage are medical expenses for those with health insurance?

## Compare your results

1. Now compare the results for illnesses and age.
   1. From the OLS results, how much more (or less) as a percentage are medical expenses for those with illnesses?
   2. From the TSLS results, are medical expenses \_\_\_\_\_\_ than those shown in the results of the OLS regression? Enter more or less.
   3. Consider how age affects medical expenses. The \_\_\_\_\_\_ indicates more expense with increasing age. Enter OLS or TSLS. (Hint, how does the coefficient (or slope) affect the cost? For example, is the OLS slope more or less (negative) than the TSLS slope for age? If the OLS slope is less, then doesn’t the OLS result indicate that the cost of medical expenses decreases less with age?)

# Using the Fertility.gdt file, evaluate how fertility affects labor supply using both an OLS regression and a Two-Stage Least Squares (TSLS) regression

For this part of PS#4 you will consider data in the Firtility.gdt file. The reference for these data is:

Joshua D. Angrist; William N. Evans, 2009, "Replication data for: Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size", <https://doi.org/10.7910/DVN/4W9GW2>, Harvard Dataverse, V1, UNF:3:gmuGDmy3Gcf/k1/lAJqw/A== [fileUNF]

Pay careful attention to how the gretl and R/RStudio scripts are built so that you can build what you need to in the future. I leave much of the detailed explanation of this analysis to the R-bloggers webpage and only point out what is directly relevant to this specific assignment. Use your results from gretl and R/RStudio to answer the following questions.

## Conduct an OLS regression

1. First, conduct an OLS regression for the variable weeksm1 on the indicator variable morekids. How many weeks less do women with more than two children work than women with two or less children?
2. Now, open your model table from the Icon view in gretl. What is the difference between the number of weeks less that women with more than two children work for the OLS model built in gretl? Enter your answer rounded to two decimal places.
3. In fact comparing the gretl model table and R/RStudio table built in its Viewer, are there any differences in the number of weeks worked by women with more than two kids?
4. Next, conduct an IV regression for the same variables using the instrumental variable samesex. Ultimately, the question will be, “does the instrumental variable have any effect”? But before the question, consider what information you have available to you. A seminal paper by Bound, Jaeger and Baker from 1995 provides some insight. I’m attaching it to this assignment for your files. Here is a quote from the R-bloggers webpage <https://www.r-bloggers.com/2013/09/detecting-weak-instruments-in-r/>.

*Fortunately, it is possible to quantitatively measure the strength of the relationship between the IVs and the endogenous variables. The so-called weak IV problem was underlined in paper by Bound, Jaeger, and Baker (1995). When the relationship between the IVs and the endogenous variable is not sufficiently strong, IV estimators do not correctly identify causal effects.*

*The Bound, Jaeger, and Baker paper represented a very important contribution to the econometrics literature. As a result of this paper, empirical studies that use IV almost always report some measure of the instrument strength. A secondary result of this paper was the establishment of a literature that evaluates different methods of testing for weak IVs. Staiger and Stock (1997) furthered this research agenda, formalizing the relevant asymptotic theory and recommending the now ubiquitous “rule-of-thumb” measure: a first-stage partial-F test of less than 10 indicates the presence of weak instruments. (Bound, Jaeger, Baker, 1995)*

This provides you with the reference for the rule-of-thumb F-statistic < 10’s origins. In addition, there is another R-bloggers webpage <https://cran.r-project.org/web/packages/ivreg/vignettes/Diagnostics-for-2SLS-Regression.html> that provides additional information. Unless “diagnostics=FALSE” is part of the ivreg command, the ivreg package outputs three tests. From this R-bloggers webpage, here is a description of these tests.

*By default, summary() prints the results of three “diagnostic” tests for 2SLS regression. These tests (which can be suppressed by setting the argument diagnostics=FALSE) are not the focus of the vignette and so we’ll comment on them only briefly:*

* *A good instrumental variable is highly correlated with one or more of the explanatory variables while remaining uncorrelated with the errors. If an endogenous regressor is only weakly related to the instrumental variables, then its coefficient will be estimated imprecisely. We hope for a large test statistic and small p-value in the diagnostic test for weak instruments, as is the case for both regression equations in the Kmenta model.*
* *Applied to 2SLS regression, the Wu–Hausman test is a test of endogenity. If all of the regressors are exogenous, then both the OLS and 2SLS estimators are consistent, and the OLS estimator is more efficient, but if one or more regressors are endogenous, then the OLS estimator is inconsistent. A large test statistic and small p-value, as in the example, suggests that the OLS estimator is inconsistent and the 2SLS estimator is therefore to be preferred.*
* *The Sargan test is a test of overidentification. That is, in an overidentified regression equation, where there are more instrumental variables than coefficients to estimate, as in Kmenta’s demand equation, it’s possible that the instrumental variables provide conflicting information about the values of the coefficients. A large test statistic and small p-value for the Sargan test suggest, therefore, that the model is misspecified. In the example, we obtain a moderately small p-value of 0.084 by chance even though we know (by the manner in which the data were constructed) that the demand equation is correct. The Sargan test is inapplicable to a just-identified regression equation, with an equal number of instrumental variables and coefficients, as in Kmenta’s supply equation. (Retrieved from R-bloggers at* [*https://cran.r-project.org/web/packages/ivreg/vignettes/Diagnostics-for-2SLS-Regression.html*](https://cran.r-project.org/web/packages/ivreg/vignettes/Diagnostics-for-2SLS-Regression.html) *, 2022)*

Answer this question from the R/RStudio output. From the first line of the diagnostic tests, the value of the F-statistic, \_\_\_\_\_ (enter the value rounded to two decimal places) as well as the P-value provided by the “weak estimates” (that is nearly) \_\_\_\_\_\_ (enter this value rounded to two decimal places) indicates that the instrument (samesex) is correlated or highly correlated with the endogenous variable, morekids.

1. Consider the gretl model output for the same “Weak Instrument test”. What is the F-statistic computed by gretl? Enter your answer rounded to two decimal places.
2. With the use of the instrumental variable samesex, now how many weeks less do women with more than two children work than women with two or less children?
3. Next, test for endogeneity of the regressor morekids. For example, you could use the Hausman test for endogeneity computed by the R/RStudio script to reverse engineer your answer to this question. What is the P-value from the Hausman test?
4. A P-value equal to 0.77 output from the Hausman test means that we should \_\_\_\_\_\_ the null hypothesis and consider the instrument, samesex, to be valid (relevant) and the TSLS regression to be preferred.
5. Using samesex as an instrumental variable for a TSLS regression, how many weeks less does a woman with more than two children work than women with two or less children? Enter the number of week rounded to two decimal places. Use either the gretl or R/RStudio output.
6. Do the results change in any significant way if more instruments are added to the analysis, i.e. agem1, black, hispan, or othrace?
7. In the last model using morekids, agem1, black, hispan, and othrace as independent variables, select the variable or variables below that are NOT statistically significant.
   1. morekids
   2. agem1
   3. black
   4. hispan
   5. othrace
   6. All independent variables are statistically significant

Although the graphical output that is included in this assignment will not be included in the final exam, it is fascinating that the graphical output by R/RStudio indicates that the answers between the OLS regression and the IV regression are so similar that the 95% confidence intervals overlap except for the intercept in the 2nd IV model that includes all additional variables. Fascinating!