**Principals of instrumental variable analysis practical 2020**

**Solutions**

This practical uses a simulated random sample from the population. It simulates a study investigating the effects of two types of anti-inflammatory drug, traditional NSAIDs (e.g. ibuprofen) vs COX-2 selective inhibitors (COX-2s e.g. celecoxib). The dataset contains data from 100,000 patients, it is a patient level file, i.e. each patient has a single row. In the dataset the exposure is indicated by the variable ‘prescribed\_cox\_2’. It equals one if the patient had a COX-2 and zero if they had a traditional NSAID. The outcome of interest is whether the patient subsequently had a gastrointestinal complication (variable ‘has\_gi\_event’), equal to one, or did not have a complication, when the outcome is equal to zero. The dataset is called iv\_practical\_2020.dta. There are 100,000 observations with variables on treatment, physician who prescribed the treatment, age, and sex. We will use the information on the physician who prescribed the treatment to create an instrument and will estimate the effects of prescribing COX-2s versus traditional NSAIDs.

1. Open the dataset and describe the variables

use iv\_practical\_2020, clear

describe

summarize

1. Create the instrument – the physician’s previous prescription. The variable visit order indicates the order in which the patients visited their GP. The command below will create a variable equal to one if the physician previously prescribed a COX-2 and equal to zero if they previously prescribed a traditional NSAID.

bysort physician\_id (visit\_order):gen prior\_prescription= prescribed\_cox\_2[\_n-1]

1. Test the first instrumental variable assumption (the relevance assumption). Are the physicians’ previous prescription associated with their subsequent prescriptions? What happens when you adjust for age and sex? How should estimates from a linear probability model be interpreted?

. regress prescribed\_cox\_2 prior\_prescription, robust

Linear regression Number of obs = 99,000

F(1, 98998) = 1214.10

Prob > F = 0.0000

R-squared = 0.0121

Root MSE = .49551

------------------------------------------------------------------------------------

| Robust

prescribed\_cox\_2 | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------------+----------------------------------------------------------------

prior\_prescription | .1101998 .0031627 34.84 0.0000 .104001 .1163986

\_cons | .4109381 .0021311 192.83 0.0000 .4067611 .4151151

------------------------------------------------------------------------------------

. regress prescribed\_cox\_2 prior\_prescription age female, robust

Linear regression Number of obs = 99,000

F(3, 98996) = 9135.67

Prob > F = 0.0000

R-squared = 0.1650

Root MSE = .45557

------------------------------------------------------------------------------------

| Robust

prescribed\_cox\_2 | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------------+----------------------------------------------------------------

prior\_prescription | .109838 .0029103 37.74 0.0000 .104134 .1155421

age | .0349172 .0002573 135.68 0.0000 .0344128 .0354216

female | -.1751715 .0028953 -60.50 0.0000 -.1808462 -.1694968

\_cons | -1.59644 .0156438 -102.05 0.0000 -1.627102 -1.565778

------------------------------------------------------------------------------------

R2 is the proportion of variability explained in the outcome variable of a regression by the covariates. The R2 values for the prior prescription instrument is 1.2%. Since the R2 statistics is small we know that the resulting IV estimates will be imprecise and have wide confidence intervals.

The F statistics for the (first stage) regression of prescribed COX-2 on prior prescription is 1214. Economists say an instrument with a first-stage F statistic less than 10 is a weak instrument, i.e. an instrument which will give an IV estimate with a relatively large finite sample bias. However, it is important not to select instruments with F statistics greater than 10 in the same dataset as performing estimation of the causal effect as this can lead to bias via winner’s curse.

The linear probability model estimates the association between the instrument and the exposure on the absolute probability scale i.e. risk differences.

1. Investigate the plausibility of the third instrumental variable assumption, independence. Do the instruments associate with the measured confounders?

. regress age prior\_prescription, robust

Linear regression Number of obs = 99,000

F(1, 98998) = 0.33

Prob > F = 0.5658

R-squared = 0.0000

Root MSE = 4.9892

------------------------------------------------------------------------------------

| Robust

age | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------------+----------------------------------------------------------------

prior\_prescription | .0182647 .0318033 0.57 0.5658 -.0440693 .0805987

\_cons | 59.99414 .0216285 2773.85 0.0000 59.95175 60.03653

------------------------------------------------------------------------------------

. regress female prior\_prescription, robust

Linear regression Number of obs = 99,000

F(1, 98998) = 0.24

Prob > F = 0.6211

R-squared = 0.0000

Root MSE = .5

------------------------------------------------------------------------------------

| Robust

female | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------------+----------------------------------------------------------------

prior\_prescription | .0015757 .0031876 0.49 0.6211 -.004672 .0078235

\_cons | .499212 .0021658 230.50 0.0000 .4949671 .5034569

------------------------------------------------------------------------------------

There was little evidence of association between any of the instruments and age or sex. Of course we cannot check for associations with unmeasured confounders. The simulated dataset includes unmeasured confounders.

1. Estimate the multivariable adjusted linear regression of prescribed COX-2 and the outcome, has GI event. Are prescriptions of COX-2s associated with a higher or lower risk of gastrointestinal events? What happens when you adjust for the observed covariates?

. regress has\_gi\_event prescribed\_cox\_2, robust

Linear regression Number of obs = 100,000

F(1, 99998) = 481.68

Prob > F = 0.0000

R-squared = 0.0050

Root MSE = .23894

----------------------------------------------------------------------------------

| Robust

has\_gi\_event | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-----------------+----------------------------------------------------------------

prescribed\_cox\_2 | .0339215 .0015456 21.95 0.0000 .0308921 .0369508

\_cons | .0454512 .0008977 50.63 0.0000 .0436917 .0472107

----------------------------------------------------------------------------------

. regress has\_gi\_event prescribed\_cox\_2 age female, robust

Linear regression Number of obs = 100,000

F(3, 99996) = 431.90

Prob > F = 0.0000

R-squared = 0.0136

Root MSE = .2379

----------------------------------------------------------------------------------

| Robust

has\_gi\_event | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-----------------+----------------------------------------------------------------

prescribed\_cox\_2 | .0529007 .0016752 31.58 0.0000 .0496173 .056184

age | -.00428 .0001677 -25.53 0.0000 -.0046087 -.0039514

female | .0228715 .0015322 14.93 0.0000 .0198685 .0258746

\_cons | .2820668 .0099824 28.26 0.0000 .2625014 .3016323

The unadjusted linear regression estimate shows that patients prescribed COX-2s are 3.3 (95% CI: 3.0, 3.7) percentage points more likely to have a gastro-intestinal adverse event compared to those prescribed traditional NSAIDs. This observational estimate increases on adjustment for age and sex to 5.3 (95% CI: 5.0, 5.6). The true effect of COX-2s on the gastro-intestinal events was simulated in this dataset as -8.3 per 100 patients treated (i.e. fewer events in those prescribed COX-2s) so we can see that this analysis is still biased by an unmeasured confounder.

1. The instrumental variable ratio estimator (Wald type) is the instrument-outcome association divided by the instrument-exposure association. Estimate this using the instrument you created above (prior\_prescription). Use the following code to save your estimates as scalars. You will need to run this code as a single block for this to work. Compare your instrumental variable estimate with the unadjusted and adjusted estimates above.

. regress prescribed\_cox\_2 prior\_prescription, robust

Linear regression Number of obs = 99,000

F(1, 98998) = 1214.10

Prob > F = 0.0000

R-squared = 0.0121

Root MSE = .49551

------------------------------------------------------------------------------------

| Robust

prescribed\_cox\_2 | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------------+----------------------------------------------------------------

prior\_prescription | .1101998 .0031627 34.84 0.0000 .104001 .1163986

\_cons | .4109381 .0021311 192.83 0.0000 .4067611 .4151151

------------------------------------------------------------------------------------

. scalar denominator=\_b[prior\_prescription]

. regress has\_gi\_event prior\_prescription, robust

Linear regression Number of obs = 99,000

F(1, 98998) = 13.63

Prob > F = 0.0002

R-squared = 0.0001

Root MSE = .23933

------------------------------------------------------------------------------------

| Robust

has\_gi\_event | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------------+----------------------------------------------------------------

prior\_prescription | -.0056154 .0015208 -3.69 0.0002 -.0085961 -.0026347

\_cons | .0636023 .0010571 60.17 0.0000 .0615304 .0656741

------------------------------------------------------------------------------------

. scalar numerator = \_b[prior\_prescription]

. display scalar(numerator)/scalar(denominator)

-.05095637

The ratio estimate is -5.1 fewer events per 100 patients treated. Since this dataset was simulated under the instrumental variable assumptions this IV estimate is very close to the true effect of -8. Of course it is not possible to know the “true” effect in real datasets.

1. Now perform two-stage least squares estimation, again using the prior prescription as the instrumental variable. Compare this with the ratio estimate above.

. ivregress 2sls has\_gi\_event (prescribed\_cox\_2 = prior\_prescription), robust

Instrumental variables (2SLS) regression Number of obs = 99,000

Wald chi2(1) = 13.28

Prob > chi2 = 0.0003

R-squared = .

Root MSE = .24248

----------------------------------------------------------------------------------

| Robust

has\_gi\_event | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-----------------+----------------------------------------------------------------

prescribed\_cox\_2 | -.0509564 .0139806 -3.64 0.0003 -.0783579 -.0235549

\_cons | .0845422 .0065355 12.94 0.0000 .0717327 .0973516

----------------------------------------------------------------------------------

Instrumented: prescribed\_cox\_2

Instruments: prior\_prescription

We can see that with a single instrument two-stage least squares gives the same estimate as the ratio estimator.

1. Perform two-stage least squares estimation manually, i.e. by fitting the first stage regression of the exposure (prescription) on the physicians’ prior prescription, saving the predicted values, then regressing the outcome on the predicted values of the exposure. Compare the standard error and confidence interval with those from two-stage least squares.

. regress prescribed\_cox\_2 prior\_prescription

Source | SS df MS Number of obs = 99,000

-------------+---------------------------------- F(1, 98998) = 1216.94

Model | 298.792426 1 298.792426 Prob > F = 0.0000

Residual | 24306.8039 98,998 .245528232 R-squared = 0.0121

-------------+---------------------------------- Adj R-squared = 0.0121

Total | 24605.5964 98,999 .248543888 Root MSE = .49551

------------------------------------------------------------------------------------

prescribed\_cox\_2 | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------------+----------------------------------------------------------------

prior\_prescription | .1101998 .003159 34.88 0.0000 .1040082 .1163913

\_cons | .4109381 .0021463 191.46 0.0000 .4067314 .4151448

------------------------------------------------------------------------------------

. predict prescribed\_cox\_2\_hat

(option xb assumed; fitted values)

(1,000 missing values generated)

. regress has\_gi\_event prescribed\_cox\_2\_hat, robust

Linear regression Number of obs = 99,000

F(1, 98998) = 13.63

Prob > F = 0.0002

R-squared = 0.0001

Root MSE = .23933

--------------------------------------------------------------------------------------

| Robust

has\_gi\_event | Coef. Std. Err. t P>|t| [95% Conf. Interval]

---------------------+----------------------------------------------------------------

prescribed\_cox\_2\_hat | -.0509564 .0138001 -3.69 0.0002 -.0780043 -.0239084

\_cons | .0845422 .0064507 13.11 0.0000 .0718989 .0971854

--------------------------------------------------------------------------------------

.

end of do-file

We see that we obtain the same IV estimate of -0.05 but with a different standard error. You should only report the standard error and confidence intervals from ivregress. Usually when performing two-stage least squares using this manual approach you will find a smaller standard error on the IV estimate. However that is not a general rule, and it is possible for ivregress to have larger standard errors than the manual approach (you should still use the standard error and confidence interval limits from ivregress). In this simulated example there the manual two-stage least squares standard errors and the ivregress estimated standard errors are very similar. This will not be the case for many empirical examples – therefore it is always better to use ivregress or ivreg2 than manually calculating 2SLS.

1. We can potentially increase the power of instrumental variables by including more of the physicians’ prior prescriptions. Create two further variables indicating the physician’s last but two, and last but three prescriptions. Investigate the first stage R2 from the regress of the exposure (prescribed\_cox\_2) on these variables and compare to the R2 from question 3 above.

. bysort physician\_id (visit\_order):gen prior\_prescription2= prescribed\_cox\_2[\_n-2]

(2,000 missing values generated)

. bysort physician\_id (visit\_order):gen prior\_prescription3= prescribed\_cox\_2[\_n-3]

(3,000 missing values generated)

.

. regress prescribed\_cox\_2 prior\_prescription, robust

Linear regression Number of obs = 99,000

F(1, 98998) = 1214.10

Prob > F = 0.0000

R-squared = 0.0121

Root MSE = .49551

------------------------------------------------------------------------------------

| Robust

prescribed\_cox\_2 | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------------+----------------------------------------------------------------

prior\_prescription | .1101998 .0031627 34.84 0.0000 .104001 .1163986

\_cons | .4109381 .0021311 192.83 0.0000 .4067611 .4151151

------------------------------------------------------------------------------------

. regress prescribed\_cox\_2 prior\_prescription2, robust

Linear regression Number of obs = 98,000

F(1, 97998) = 1310.64

Prob > F = 0.0000

R-squared = 0.0132

Root MSE = .49523

-------------------------------------------------------------------------------------

| Robust

prescribed\_cox\_2 | Coef. Std. Err. t P>|t| [95% Conf. Interval]

--------------------+----------------------------------------------------------------

prior\_prescription2 | .1150178 .003177 36.20 0.0000 .1087908 .1212448

\_cons | .4086686 .0021406 190.92 0.0000 .4044731 .4128641

-------------------------------------------------------------------------------------

. regress prescribed\_cox\_2 prior\_prescription3, robust

Linear regression Number of obs = 97,000

F(1, 96998) = 1251.14

Prob > F = 0.0000

R-squared = 0.0128

Root MSE = .49536

-------------------------------------------------------------------------------------

| Robust

prescribed\_cox\_2 | Coef. Std. Err. t P>|t| [95% Conf. Interval]

--------------------+----------------------------------------------------------------

prior\_prescription3 | .1129802 .0031941 35.37 0.0000 .1067198 .1192406

\_cons | .4097004 .0021524 190.35 0.0000 .4054817 .4139191

-------------------------------------------------------------------------------------

We can see that each of the prior prescriptions predicts whether the physician will subsequently prescribe a COX-2. The F-stat is a function of R2 and the sample size and is a test of the strength of the instruments. The more relevant instruments included, the stronger the instrument will be, and the more precise the instrumental variable estimates will be.

1. Now perform two-stage least squares estimation using each instrument separately. What do notice about these three estimates compared to the estimate in question 7? What do you notice about the sample size? Why has this happened?

. ivregress 2sls has\_gi\_event (prescribed\_cox\_2 = prior\_prescription), robust

Instrumental variables (2SLS) regression Number of obs = 99,000

Wald chi2(1) = 13.28

Prob > chi2 = 0.0003

R-squared = .

Root MSE = .24248

----------------------------------------------------------------------------------

| Robust

has\_gi\_event | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-----------------+----------------------------------------------------------------

prescribed\_cox\_2 | -.0509564 .0139806 -3.64 0.0003 -.0783579 -.0235549

\_cons | .0845422 .0065355 12.94 0.0000 .0717327 .0973516

----------------------------------------------------------------------------------

Instrumented: prescribed\_cox\_2

Instruments: prior\_prescription

. ivregress 2sls has\_gi\_event (prescribed\_cox\_2 = prior\_prescription2), robust

Instrumental variables (2SLS) regression Number of obs = 98,000

Wald chi2(1) = 17.60

Prob > chi2 = 0.0000

R-squared = .

Root MSE = .24289

----------------------------------------------------------------------------------

| Robust

has\_gi\_event | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-----------------+----------------------------------------------------------------

prescribed\_cox\_2 | -.0565514 .0134805 -4.20 0.0000 -.0829726 -.0301302

\_cons | .0870636 .0063106 13.80 0.0000 .074695 .0994322

----------------------------------------------------------------------------------

Instrumented: prescribed\_cox\_2

Instruments: prior\_prescription2

. ivregress 2sls has\_gi\_event (prescribed\_cox\_2 = prior\_prescription3), robust

Instrumental variables (2SLS) regression Number of obs = 97,000

Wald chi2(1) = 5.66

Prob > chi2 = 0.0174

R-squared = .

Root MSE = .24092

----------------------------------------------------------------------------------

| Robust

has\_gi\_event | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-----------------+----------------------------------------------------------------

prescribed\_cox\_2 | -.0325962 .0137032 -2.38 0.0174 -.0594539 -.0057384

\_cons | .0759935 .0063985 11.88 0.0000 .0634528 .0885343

----------------------------------------------------------------------------------

Instrumented: prescribed\_cox\_2

Instruments: prior\_prescription3

The three IV estimates using the three prior prescriptions as separate instruments are -5.1 (95%CI: -2.3, -7.8), -5.7 (95%CI: -3.0, -8.3), and -3.3 (95%CI: -0.6, 5.9). A weighted average of these three estimates gives the IV estimate using them as multiple instruments as in the next question. The formula for the weights is known and can be found in Angrist, J.D. and Imbens, G.W. (1995), Two-Stage Least Squares estimation of average causal effects in models with variable treatment intensity, Journal of the American Statistical Association 90, 431-442.

The sample size has fallen because we do not have any values for the first two, and first three observations for each GP for instrument based on the two and three lagged prescriptions respectively.

1. Now perform two-stage least squares estimation using the three instruments as multiple instruments.

. ivregress 2sls has\_gi\_event (prescribed\_cox\_2 = prior\_prescription prior\_prescription2 prior\_prescription3), robust

Instrumental variables (2SLS) regression Number of obs = 97,000

Wald chi2(1) = 28.61

Prob > chi2 = 0.0000

R-squared = .

Root MSE = .24201

----------------------------------------------------------------------------------

| Robust

has\_gi\_event | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-----------------+----------------------------------------------------------------

prescribed\_cox\_2 | -.0468763 .0087639 -5.35 0.0000 -.0640532 -.0296994

\_cons | .0825892 .0041704 19.80 0.0000 .0744154 .090763

----------------------------------------------------------------------------------

Instrumented: prescribed\_cox\_2

Instruments: prior\_prescription prior\_prescription2 prior\_prescription3

This estimate is more precise than using any one instrument alone.

1. Investigate the first-stage diagnostics (partial f-statistic), endogeneity test, and overidentification test using -ivregress- postestimation commands. Interpret the endogeneity test and over-identification test p-values.

. ivregress 2sls has\_gi\_event (prescribed\_cox\_2 = prior\_prescription\*), robust

Instrumental variables (2SLS) regression Number of obs = 97,000

Wald chi2(1) = 28.61

Prob > chi2 = 0.0000

R-squared = .

Root MSE = .24201

----------------------------------------------------------------------------------

| Robust

has\_gi\_event | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-----------------+----------------------------------------------------------------

prescribed\_cox\_2 | -.0468763 .0087639 -5.35 0.0000 -.0640532 -.0296994

\_cons | .0825892 .0041704 19.80 0.0000 .0744154 .090763

----------------------------------------------------------------------------------

Instrumented: prescribed\_cox\_2

Instruments: prior\_prescription prior\_prescription2 prior\_prescription3

. estat firststage, all

First-stage regression summary statistics

--------------------------------------------------------------------------

| Adjusted Partial Robust

Variable | R-sq. R-sq. R-sq. F(3,96996) Prob > F

-------------+------------------------------------------------------------

prescribed~2 | 0.0310 0.0310 0.0310 1060.82 0.0000

--------------------------------------------------------------------------

Shea's partial R-squared

--------------------------------------------------

| Shea's Shea's

Variable | Partial R-sq. Adj. Partial R-sq.

-------------+------------------------------------

prescribed~2 | 0.0310 0.0310

--------------------------------------------------

We can see the first stage F statistic in this post-estimation output. We can also see Shea’s R-squared is the same as the first-stage R2 in this model. This is because we have not included any measured confounders in the analysis. If you include confounders the Shea’s R2 gives a more representative picture of the instrument strength since the naïve R2 will include variance explained by the confounders. This is why normally the partial R2 should be reported as a measure of instrument strength. This is reported by many commands including ivreg2.

. estat endogenous

Tests of endogeneity

Ho: variables are exogenous

Robust score chi2(1) = 89.4173 (p = 0.0000)

Robust regression F(1,96997) = 89.5939 (p = 0.0000)

One way of interpreting an endogeneity test is that it tests the difference between the exposure-outcome association estimate from linear regression and the instrumental variable estimate. The observational estimate was a 3.4 (95% CI 3.0, 3.7) percentage point increase compared to our IV estimate here of -4.7 (95% CI -3.0, 6.4). We can see that as neither estimate is spanned by the 95% confidence interval of the other estimate that the test will have a p-value less than 0.05. We can see that both variants of the endogeneity test reported by the –estat endogenous- command report very small P-values.

. estat overid

Test of overidentifying restrictions:

Score chi2(2) = 1.86595 (p = 0.3934)

If we estimate an instrumental variable with a single instrument this is termed ‘exactly identified’ because there are the same number of instruments as causal parameters. If we estimate an instrumental variable model with two or more instruments (in addition to the vector of 1s) that is termed ‘over-identified’ because there are now more instruments than causal effects.

We saw in question 11 that the IV estimate using multiple instruments is equal to a weighted average of the IV estimates using the instruments separately. An interpretation of an over-identification test is that it is a statistical test of the equality of those separate IV estimates. This can tell us whether the joint use of a set of instruments is valid. We saw that the separate IV estimates using the three instruments were very similar with overlapping confidence intervals. Hence the over-identification test reported by –estat overid- gives a large P-value. This means there is little evidence that the effects estimated by each of the instruments are different.

BONUS QUESTION 1: What 4th instrumental variable assumption could be used here? What treatment parameter can it be used to estimate?

One of the most commonly used assumptions is that the instrument (physicians preferences) has a monotonic effect on the likelihood of being prescribed a COX-2. This requires that attending a physician that prefers COX-2s rather than traditional NSAIDs, would increase the likelihood that someone would be prescribed a COX-2. Under this assumption we estimate a local average treatment effect which is the average effects of treatment in individuals whose treatment status was affected by the instrument.

BONUS QUESTION 2: If we included measured confounders in the instrumental variable estimation what interpretation would the causal parameter from two-stage least squares estimate. By reading the helpfile for ivregress can you see how to include confounders?

Included confounders in a model means that the IV estimate is conditional on those confounders.

We can include a variable list (varlist) of confounders in ivregress by specifying them after the outcome variable but before the brackets. For example:

ivregress 2sls has\_gi\_event [varlist1] (prescribed\_cox\_2 = prior\_prescription)

BONUS QUESTION 3: Using the helpfile for the ivreg2\* command, can you run all the analyses in question 10 using a single ivreg2 command?

ivreg2 has\_gi\_event (prescribed\_cox\_2 = prior\_prescription prior\_prescription3), first endog(prescribed\_cox\_2)

\* If ivreg2 is not already installed, use the following command:

ssc install ivreg2

BONUS QUESTION 4: What happens if we run the manual analysis described in Question 8 but include the covariates age and sex in the first stage, but not the second stage?

. regress prescribed\_cox\_2 prior\_prescription female age

Source | SS df MS Number of obs = 99,000

-------------+---------------------------------- F(3, 98996) = 6520.44

Model | 4059.7885 3 1353.26283 Prob > F = 0.0000

Residual | 20545.8079 98,996 .207541798 R-squared = 0.1650

-------------+---------------------------------- Adj R-squared = 0.1650

Total | 24605.5964 98,999 .248543888 Root MSE = .45557

------------------------------------------------------------------------------------

prescribed\_cox\_2 | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------------+----------------------------------------------------------------

prior\_prescription | .109838 .0029044 37.82 0.0000 .1041455 .1155305

female | -.1751715 .0028958 -60.49 0.0000 -.1808472 -.1694958

age | .0349172 .0002902 120.32 0.0000 .0343484 .035486

\_cons | -1.59644 .0175802 -90.81 0.0000 -1.630897 -1.561983

------------------------------------------------------------------------------------

. predict prescribed\_cox\_2\_hat

(option xb assumed; fitted values)

(1,000 missing values generated)

. regress has\_gi\_event prescribed\_cox\_2\_hat, robust

Linear regression Number of obs = 99,000

F(1, 98998) = 303.94

Prob > F = 0.0000

R-squared = 0.0034

Root MSE = .23894

--------------------------------------------------------------------------------------

| Robust

has\_gi\_event | Coef. Std. Err. t P>|t| [95% Conf. Interval]

---------------------+----------------------------------------------------------------

prescribed\_cox\_2\_hat | -.0689349 .0039541 -17.43 0.0000 -.0766848 -.061185

\_cons | .0928448 .0021133 43.93 0.0000 .0887028 .0969868

--------------------------------------------------------------------------------------

The manual two-stage least squares approach now gives a point estimate of -6.9 (95%CI: -6.1, -7.7) compared to the unadjusted IV estimates of -5.1 (95%CI: -2.4, -7.8) and the correctly adjusted IV estimates of 5.1 (95%CI: -2.4, 7.8). The manual two-stage least squares approach now results in biased estimates. The bias in this instance is fairly minor, but in other examples could be much larger. This can be corrected by including the same covariates in both stages of the regression. However, the easiest method is to use a routine like ivregress or ivreg2 which will guarantee that all covariates are included in both stages and reduces the possibility of error.