QE Problem Set 6

1 X may be endogenous because ~~the variable of~~ student ~~ability~~ academic ability is an unobserved determinant of performance on the final exam which is plausibly correlated with percentage of lectures attended. Plausibly, more academically capable students are better able to understand the content from the notes or textbook, hence have less incentive to attend lectures, and ~~do if~~ therefore attend fewer lectures. ~~Hence~~ Academic ability is also a plausible determinant of final exam performance, it is likely that more academically capable students will perform better as a result. Academic ability thus turns out to be one of the unobserved determinants "collected" in u, such that $cov(X,u) \neq 0$

Alternatively, conscientousness is an unobserved determinant of Y that is correlated with X. Plausibly, more conscientous students perform better at the final exams because they are more conscientous, and more conscientous students are also ~~also~~ likely to attend more lectures.

A control, W, is a good control if it is exogenous, such that as estimation of regression parameters is consistent. ~~and A control is likely to be exogenous if Plausibly exogenous~~ Plausible exogenous controls include high school results / admissions test results, whether or not a student has a term-time job, ~~hours spent studying in a week~~, average hours spent studying in a week during term, average number of nights out in a week during term, average hours of sleep a night during term, average caffeine intake a day during term, percentages of classes attended

Z is plausibly relevant: a student whose term time residence is closer to the lecture theatre is ~~more likely to attend lec~~ likely to attend more lectures. Z is not likely to be exogenous; a student whose term time residence is closer to the lecture theatre is also likely to reside closer to the ~~classrooms, hence more likely to attend classes, which are~~ which is a plausible ~~determinant of~~ libraries, hence have more convenient access to a conducive studying space, which is likely to affect final exam performance. Z is likely to be excluded, distance ~~of~~ from term time residence to lecture theatres is unlikely to have a direct causal effect on

final exam performance.

X could be correlated with the quality of the school (teacher experience, access to ~~educ~~ educational resources, etc.), which ~~can~~ is an unobserved determinant of Y. Perhaps girls schools are disproportionately better funded, or disproportionately private rather than public schools, and thus have better teachers or educational resources, which are determinants of year 12 exam scores.

Plausible exogenous controls include: household income, number of siblings, number of parents in the workforce, teacher years of experience, average attendance, private or public school, average number of hours of teaching a week.

Z is relevant: girls who live in the catchment area for a girls-only school are more likely to attend a girls-only school. Z is ~~plau~~ not likely to be exogenous. Plausibly, girls who live in such catchment areas are more likely to be the children of parents who are particularly invested in their child's educational attainment (and want their daughters to attend a girls school), who also invest greater resources in their child's education, which is a determinant of year 12 exam results. Z is plausibly excluded: ~~whether~~ whether a girl lives within such a catchment zone is unlikely to have a direct effect on year 12 exam results.

2a. At equilibrium,

$q_s = q_d$

$\beta_0 + \beta_1 p + u = \delta_0 + \delta_1 p + \delta_2 t + v$

~~$u = (\beta_0 + \beta_1) - (\beta_0 - \delta$~~

$u = (\delta_0 - \beta_0) + (\delta_1 - \beta_1)p + \delta_2 t + v$

~~$cov(p,u) =$~~

$= cov(p, (\delta_0 - \beta_0) + (\delta_1 - \beta_1)p + \delta_2 t + v)$

$= (\delta_1 - \beta_1) var(p) + \delta_2 cov(p,t) + cov(p,v)$

~~$cov(p,u)$ not necessarily $= 0$~~

~~$v = (\delta_0 - \beta_0) + (\delta_1 - \beta_1)p + \delta_2 t$~~

~~$v = (\beta_0 - \delta_0) + (\beta_1 - \delta_1)p - \delta_2 t + u$~~

~~$cov(p,v)$~~

~~$= (\beta_1 - \delta_1) var(p) - \delta_2 cov(p,t) + cov(p,u)$~~

$cov(p,v)$ not necessarily $= 0$

$(\beta_1 - \delta_1)p = (\delta_0 - \beta_0) + \delta_2 t - u + v$

~~$cov(p,$~~ $p = (\delta_0 - \beta_0)/(\beta_1 - \delta_1) + \delta_2 t/(\beta_1 - \delta_1)$
$- u/(\beta_1 - \delta_1) + v/(\beta_1 - \delta_1)$

$cov(p,u) = \delta_2/\beta_1 - \delta_1 \, cov(t,u) + (\delta_1 - \beta_1) var(u)$
$- 1/\beta_1 - \delta_1 \, var(u) + 1/\beta_1 - \delta_1 \, cov(v,u)$

~~Given the favourable assump~~

Even given the favourable assumptions that
$cov(t,u) = cov(v,u) = 0$,
$cov(p,u) = -1/\beta_1 - \delta_1 \, var(u) \neq 0$

$cov(p,v) = \delta_2/\beta_1 - \delta_1 \, cov(t,v) - 1/\beta_1 - \delta_1 \, var(v)$
$+ 1/\beta_1 - \delta_1 \, cov(u,v)$

Even given the favourable assumptions that
$cov(t,v) = cov(u,v) = 0$
$cov(p,v) = -1/\beta_1 - \delta_1 \, var(v) \neq 0$

ii $\delta_2 t = (\beta_0 - \delta_0) + (\beta_1 - \delta_1)p + u - v$
$t = (\beta_0 - \delta_0)/\delta_2 + \beta_1 - \delta_1/\delta_2 \, p + 1/\delta_2 u - 1/\delta_2 v$

$cov(t,u) = \beta_1 - \delta_1/\delta_2 \, cov(p,u) + 1/\delta_2 \, var(u) - 1/\delta_2 \, cov(v,u)$
Even given the favourable assumptions that
$cov(p,u) = cov(v,u) = 0$
$cov(t,u) = 1/\delta_2 \, var(u) \neq 0$

$cov(t,v) = \beta_1 - \delta_1/\delta_2 \, cov(p,v) + 1/\delta_2 \, cov(u,v) - 1/\delta_2 \, var(v)$
Even given the favourable assumptions that
$cov(p,v) = cov(v,u) = 0$
$cov(t,v) = -1/\delta_2 \, var(v) \neq 0$

b ~~cov~~ $cov(p,t) = \delta_2/\beta_1 - \delta_1 \, var(t) - 1/\beta_1 - \delta_1 \, cov(u,t)$
$+ 1/\beta_1 - \delta_1 \, cov(v,t)$

Relevance
$cov(p,t) \neq 0$
$\delta_2 \, var(t) - cov(u,t) + cov(v,t) \neq 0$

Exogeneity
~~$cov(t,u) = cov(t,v) = 0$~~ $cov(t,u) = 0$

c

3a Consider

Structural equation

$Y = \beta_0 + \beta_1 X + u$ ①

where X is possibly endogenous

First stage regressions

$X = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + v$ ②

where $E(v)=0$, $cov(z_1, v) = cov(z_2, v) = 0$ by construction

$X = \delta_0 + \delta_1 z_1 + w$ ③

where $E(w)=0$, $cov(z_1, w) = 0$ by construction

$X = \pi_0 + \pi_2 z_2 + t$ ④

where $E(t)=0$, $cov(z_2, t) = 0$ by construction

$\hat{\beta}_1 = \hat{cov}(Y, \hat{X}) / \hat{cov}(X, \hat{X})$

for $\hat{X} = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2$

~~$\hat{\beta}_1 = \hat{cov}(Y, \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2)$~~

~~$/ \hat{cov}(\gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + v, \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2)$~~

~~from ③, by definition of $\hat{X}$~~

~~$= \hat{cov}(Y, \gamma_1 z_1 + \gamma_2 z_2)$~~

~~$/ \hat{cov}(\gamma_1 z_1 + \gamma_2 z_2, \gamma_1 z_1 + \gamma_2 z_2)$~~

$\hat{\beta}_1 = \hat{cov}(Y, \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2)$

$/ \hat{cov}(X, \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2)$

by substitution

$= \hat{cov}(Y, \gamma_1 \hat{cov}(Y, z_1) + \gamma_2 \hat{cov}(Y, z_2)]$

$/ [\gamma_1 \hat{cov}(X, z_1) + \gamma_2 \hat{cov}(X, z_2)]$

by bilinearity of covariance, since γ_0 is a constant

Suppose that $\hat{\beta}_{1|z_1} = \hat{\beta}_{1|z_2} = a$

then

$\hat{\beta}_1 = $ ~~$[\gamma_1 a \hat{cov}(X, z_1) + \gamma_2 a \hat{cov}(X, z_2$~~

$[\gamma_1 a \hat{cov}(X, z_1) + \gamma_2 a \hat{cov}(X, z_2)]$

$/ [\gamma_1 \hat{cov}(X, z_1) + \gamma_2 \hat{cov}(X, z_2)]$

$= a = \beta_{1|z_1} = \beta_{1|z_2}$

b Consider the linear regression model

$\hat{u} = \phi_0 + \phi_1 z_1 + \phi_2 z_2 + s$

where $E(s)=0$, $cov(z_1, s) = cov(z_2, s) = 0$ by construction

and $\hat{u} = Y - \hat{\beta}_0 - \hat{\beta}_1 X$

$\hat{cov}(\hat{u}, z_1) = \hat{cov}(Y - \hat{\beta}_0 - \hat{\beta}_1 X, z)$

$= \hat{cov}(Y - \hat{\beta}_0 - \hat{\beta}_1(\hat{\delta}_0 + \hat{\delta}_1 z_1 + w), z_1)$

by substitution ~~of ③ example counts~~

$= \hat{cov}(Y - \hat{\beta}_0 \hat{\delta}_1 z_1, z_1)$

by bilinearity of covariance operator, since $\hat{\beta}_0$ and $\hat{\beta}_1 \hat{\delta}_0$ are constants and $\hat{cov}(w, z_1) = 0$

$= \hat{cov}(Y, z_1) - \hat{\beta}_0 \hat{\delta}_1 \hat{var}(z_1)$

$= \hat{cov}(Y, z_1) - $ ~~$\hat{cov}(Y, z_1) / \hat{cov}(X, z_1)$~~ $\hat{cov}(X, z_1)$

$- \left(\hat{cov}(Y, z_1) / \hat{cov}(X, z_1)\right)\left(\hat{cov}(X, z_1) / \hat{var}(z_1)\right)$

$(\hat{var}(z_1))$

$= 0$

By symmetry, $\widehat{cov}(\hat{u}, z_2) = 0$
since $\hat{z}_1^{\wedge}$ and $\hat{z}_2^{\wedge}$ are both linear functions of z_1 and z_2, $\widehat{cov}(\hat{u}, \hat{z}_1^{\wedge}) = \widehat{cov}(\hat{u}, \hat{z}_2^{\wedge}) = 0$

$\hat{\phi}_1 = \widehat{cov}(\hat{u}, \hat{z}_1^{\wedge}) / \widehat{var}(\hat{z}_1^{\wedge}) = 0$
$\hat{\phi}_2 = \widehat{cov}(\hat{u}, \hat{z}_2^{\wedge}) / \widehat{var}(\hat{z}_2^{\wedge}) = 0$
$\hat{\phi}_0 = \bar{\hat{u}} - \hat{\phi}_1 \bar{z}_1 - \hat{\phi}_2 \bar{z}_2 = \bar{\hat{u}}$

The unrestricted ~~ass~~ model of $\hat{u}$ on z_1 and z_2 estimated by OLS is identical to the restricted ($\phi_1 = \phi_2 = 0$) model. Hence $SSR_{un} = SSR_{rs}$

$F = ((SSR_{rs} - SSR_{un})/q) / (SSR_{un}/(n-k-1)) = 0$
$F^* = {}^{n}/_{n-1} F = {}^{2}/_{1} F = 0$
p-value = 1

~~Fail to~~ When $\hat{\beta}_{1|z_1} = \hat{\beta}_{1|z_2}$, fail to reject, at all levels of significance, the null hypothesis that $\hat{u}$ is ~~correlated with at least one of~~ uncorrelated with z_1 and z_2. When $\hat{\beta}_{1|z_1} = \hat{\beta}_{1|z_2}$, the information from the first stage regression is entirely captured in the estimation of the coefficients in the structural equation, there is no further information from which we can evaluate whether the residuals of the structural equation are correlated with the instruments. There can be no evidence of exogeneity.

could you read through this and let me know if the intuition and wording are correct?

4a on average, a one-year increase in years of completed schooling is associated with an ~~0.049~~ increase in hourly wages by a proportion $e^{0.049} = 1.0502$, holding age, black, south constant

Confidence interval

~~$C = [0.049 - 0.004 c_\alpha, 0.049 + 0.004 c_\alpha]$~~

~~$C = [\hat{\beta}_{1,educ} - c_\alpha \hat{se}(\hat{\beta}_{1,educ}), \hat{\beta}_{1,educ} + c_\alpha$~~

$C = [\hat{\beta}_{1,educ} - c_\alpha \hat{se}(\hat{\beta}_{1,educ}), \hat{\beta}_{1,educ} + c_\alpha \hat{se}(\hat{\beta}_{1,educ})]$

~~$= [0.049 - 0.004 c_\alpha, 0.049 + 0.004 c_\alpha]$~~

$= [0.049 - 0.004 c_\alpha, 0.049 + 0.004 c_\alpha]$

where $P(-c_\alpha < z < c_\alpha) = 0.99$,

$c_\alpha = -\Phi^{-1}(0.005) = 2.5758$

$C = [0.038697, 0.059303]$

There is a 99% probability that the interval $C = [0.038697, 0.059303]$ contains the population regression parameter (coefficient of educ in the population linear regression of lwage on educ, age, black, and south).

There is a 99% probability that the interval $[e^{0.038697} = 1.0395, e^{0.059303} = 1.0611]$ contains the population parameter for the average proportionate increase in hourly wages associated with a one year increase in years of schooling, holding age, black, south constant.

6 $\hat{\beta}_{1,educ} \xrightarrow{p} \beta_{1,educ} = \beta_{2,educ} + \beta_{2,iqscore} * \pi$

where π is the coefficient on educ in the population linear regression of iqscore on educ, age, black, south.

~~From estimated coefficient of iqscore in regression (2) and given standard error of this estimate, p-value of this coefficient is less than 1%. We have strong evidence reason to think~~

$\hat{\beta}_{1,educ} > \hat{\beta}_{2,educ}$ since there is a positive ~~rate~~ relationship between iqscore and educ, ~~and this is~~ captured by π, and a positive relationship between iqscore and lwage, captured by $\beta_{2,iqscore}$, and the OLS estimates converge in probability to the population regression parameters.

From estimated coefficient of iqscore in (2), and given the standard error on this coefficient, p-vale of this coefficient is less than 1%. We have strong reason to think that iqscore is a determinant of lwage. This determinant is unobserved in (1), and is correlated with educ ~~(hence $\hat{\beta}_{1,educ} > \hat{\beta}_{2,educ}$)~~. (hence $\pi > 0$ implied by $\hat{\beta}_{1,educ} > \hat{\beta}_{2,educ}$). Unthogonality fails in (1), hence (1) does not consistently estimate the returns to education.

other determinants of ~~educ~~ lwage that are correlated with educ, such as household income and wealth are unobserved in (2). children from households with higher income or wealth have greater access to opportunity and hence are likely to have higher incomes as a result. ~~SA~~ Such households are also more able to afford to send their children for further education. ~~Hence~~ Orthogonality fails in (2), and (2) does not consistently estimate returns to education.

c ~~So~~ $\hat{\beta}_{1,educ} < \hat{\beta}_{3,educ}$
$\hat{se}(\hat{\beta}_{1,educ}) < \hat{se}(\hat{\beta}_{3,educ})$

$\hat{\beta}_{1,educ} < \hat{\beta}_{3,educ}$ suggests that educ is endogenous in (1), and that ~~unobserved~~ omitted variables in (1) result in a negative bias. There are ~~SA~~ unobserved determinants of ~~lwage~~ lwage that are negatively correlated with lwage and positively correlated with educ (perhaps, propensity to pursue an academic career) or unobserved determinants of lwage that are positively correlated with lwage and negatively correlated with educ (perhaps risk tolerance).

$\hat{se}(\hat{\beta}_{1,educ}) < \hat{se}(\hat{\beta}_{3,educ})$ because ~~$\hat{var}(x)$~~ $\hat{var}(\hat{educ}) < \hat{var}(educ)$ where $\hat{educ}$ is the linear prediction of educ ~~as~~ from libcrd14 ~~hence~~ $\omega^2\hat{\beta}_{1,educ} < \omega^2\hat{\beta}_{3,educ}$ and $\hat{se}(\hat{\beta}_{1,educ}) < \hat{se}(\hat{\beta}_{3,educ})$
Intuitively, the component of educ uncorrelated with other unobserved determinants of lwage ~~is less~~ explains less of the variance of lwage ~~since~~ than educ since the former ~~excludes relevant~~ ~~blocks inherited about lwage~~ "contains less information" than the latter.

d ~~The models estimated in (5) and (6) are used to generate linear predictions of educ on the basis of from libcrd14 and libcrd14, dated, married, on which (3) and (4) are estimated respectively.~~

The model estimated in (5) is used to generate linear predictions of educ from libcrd14, controlling for age, black, south. Model (3) is estimated by regressing lwage on this prediction, and the controls age, black, south. Analogously for (6) and (4). ~~SAME~~.

e Relevance: the instrumental variable is correlated with the endogenous variable for which it is an instrument.
Exogeneity: the instrumental variable is uncorrelated with unobserved determinants of the dependent variable.
Exclusion: the instrumental variable does not appear

Assuming libcrd14 is relevant and exogenous
~~SA~~ Assuming no measurement error, simultaneity seems implausible

Is this correct? Why (formally) does the linear predictor of x have lower variance than x? Do we know what happens to the residuals in (3)?

What is a more precise / technical way to spell this out?

in the structural equation.

All of libcrdkt, daded, and momed are excluded from the structural equation since there is no plausible direct causal relationship between the three potential instruments and lwage.

daded and momed ~~is likely to be~~ are likely to be relevant since ~~if~~ more educated parents are ~~more~~ likely to have higher incomes and hence ~~is~~ to be more able to finance their child's education. daded and momed are likely to be exogenous since lwage is likely to be ~~is~~ partially determined by an individual's ~~access to~~ professional network, and more educated parents are likely to have more extensive and valuable networks that their children can access.

~~Whether or not libcrdkt is relevant and exogenous depends on how a library card works. If library cards are randomly assigned to different households~~

Assuming that library cards belong to library members and are necessary for borrowing books from libraries. libcrdkt is likely to be relevant: households with a library card are more likely to have parents who are more invested in their child's education, ~~hence more likely~~ hence more likely to have children who ~~have an attend~~ remain in the education system for longer. libcrdkt is unlikely to be exogenous since parents who are more invested in their child's education are likely to also be more invested in their child's career, and to support this in various ways that might affect lwage.

Test for relevance by an F test for the regression of educ on the instruments and controls of
H_0: coefficients on all instruments ~~is~~ are zero
H_1: coefficients on at least one instrument is non-zero.

Given $F = 128$ and $F = 100$ for ~~(5) and (6)~~ the given hypotheses for (5) and (6) respectively, we have corresponding p-values 0 and 0. We have very strong evidence for rejecting the null for both (5) and (6), hence very strong empirical reason to think libcrdkt and libcrdkt, daded, momed are relevant instruments.

Test for exogeneity by an F test for the regression of the ~~residuals of (5) and (6)~~ residual of (1) on the instruments and controls, of

H_0: coefficients on all instruments are zero
H_1: coefficients on at least one instrument is non-zero

f $\hat{se}(\hat{\beta}_{4,educ}) < \hat{se}(\hat{\beta}_{3,educ})$ because $\hat{var}(\hat{educ}_3) < \hat{var}(\hat{educ}_4)$ where $\hat{educ}_i$ is the linear prediction of educ on the basis of libcrd14 in (3) and libcrd14, daded, momed in (4) for $i=3$ and $i=4$ respectively. Hence $e^2_{\hat{\beta}_{3,educ}} > e^2_{\hat{\beta}_{4,educ}}$ and $\hat{se}(\hat{\beta}_{3,educ}) > \hat{se}(\hat{\beta}_{4,educ})$

~~this has no implications for the validity of daded and momed as IVs~~
This suggests that daded and momed are relevant instruments, but has no implications for their exogeneity.

5 If some school principles were successfully pressured by parents to place their children in the small classes, then treatment (placement in a small class) is nonrandom, and plausibly correlated with other unobserved determinants of academic performance. ~~If~~ children who receive the treatment are more likely to be the children of parents who successfully pressure school principles to place their child in the small classes. These parents, plausibly, are more invested in their child's education, and hence expend more resources to support their child academically, which plausibly improves their child's academic performance. If treatment is correlated with unobserved determinants of academic performance, orthogonality fails and OLS regression does not consistently estimate the causal effect of being in a small class on academic performance.

Internal validity can be restored by excluding from the OLS regression any students who enrolled in a small (normal) class ~~s~~ that were originally randomly assigned to a normal (small) class. ~~of the remaining students,~~ For the remaining students, ~~enrolm~~ ~~t~~ treatment (enrolment in a small class) is randomly assigned, and there is no element of choice. Hence treatment is uncorrelated with unobserved determinants of academic performance, orthogonality holds, and OLS regression consistently estimates the causal effect of enrolment in a small class on academic performance (within this subset of students).

Alternatively, since original assignment is correlated with treatment, random (hence uncorrelated with unobserved determinants of academic performance), and has no independent effect on academic performance (apart from through actual enrolment), it is a valid instrument for treatment. ~~The~~ IV or 2SLS regression consistently estimates the causal effect of enrolment in a small class on academic performance (in the population).

6a On average, an individual offered the opportunity to participate in the training programme has an hourly wage $0.78 higher than an individual not offered this opportunity, two years after the commencement of the ~~p~~ study.

Confidence interval
$C = [0.78 - 0.23 c_\alpha, 0.78 + 0.23 c_\alpha]$
where $P(-c_\alpha < z < c_\alpha) = 0.95$
$c_\alpha = -\Phi^{-1}(0.025) = 1.9600$
$C = [0.3292, 1.2308]$

There is a 95% probability that the interval $C = [0.3292, 1.2308]$ contains the population regression parameter (coefficient on offer in a regression of wage on offer).

6 ~~Consider the linear regression models~~
~~wage~~

Consider
structural equation
$wage = \beta_0 + \beta_1 trained + u$
where trained is possibly endogenous
First stage regression
$trained = \gamma_0 + \gamma_1 offer + v$
where $E(v) = 0$, $cov(offer, v) = 0$ by construction

By substitution we have
Reduced form regression
$wage = \beta_0 + \beta_1(\gamma_0 + \gamma_1 offer + v) + u$
$= (\beta_0 + \beta_1\gamma_0) + \beta_1\gamma_1 offer + (\beta_1 v + u)$

Assuming that offer is a valid instrument,
Relevance holds
$cov(trained, offer) \neq 0$, hence $\gamma_1 \neq 0$
Exogeneity holds
$cov(offer, u) = 0$
Exclusion holds
offer does not appear in the structural equation

~~From~~ $cov(offer, \beta_1 v + u) = \beta_1 cov(offer, v) + cov(offer, u)$
$= 0$
Orthogonality holds
OLS regression of wage on offer consistently estimates $\beta_1\gamma_1$.
OLS regression of trained on offer consistently estimates γ_1

$\hat{\beta}_1 = \hat{\beta}_1\hat{\gamma}_1 / \hat{\gamma}_1 = 0.78/0.63 = 1.2381$

On average, participation in the training programme is associated with a $1.2381 increase in hourly wages ~~#~~ 2 years after the commencement of the study.

Problem-Set-6-Question-7.R

r1454158

2022-05-31

```
# Clear the environment
rm(list = ls())

# Install packages
install.packages("estimatr")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("car")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages("margins")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
# Load packages
library(estimatr)
library(car)
```

```
## Loading required package: carData
```

```
# Load data
fertility = read.csv("fertility.csv")

# Preview data
dim(fertility)
```

```
## [1] 254654      9
```

```
head(fertility)
```

```
##   morekids boy1st boy2nd samesex agem1 black hispan othrace weeksm1
## 1        0      1      0       0    27     0      0       0       0
## 2        0      0      1       0    30     0      0       0      30
## 3        0      1      0       0    27     0      0       0       0
## 4        0      1      0       0    35     1      0       0       0
## 5        0      0      0       1    30     0      0       0      22
## 6        0      1      0       0    26     0      0       0      40
```

```
summary(fertility)
```

```
##     morekids          boy1st           boy2nd          samesex
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
```

```
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   :0.3806   Mean   :0.5144   Mean   :0.5126   Mean   :0.5056
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      agem1            black            hispan           othrace
##  Min.   :21.00   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:28.00   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :31.00   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :30.39   Mean   :0.05166   Mean   :0.07421   Mean   :0.05634
##  3rd Qu.:33.00   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :35.00   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##     weeksm1
##  Min.   : 0.00
##  1st Qu.: 0.00
##  Median : 5.00
##  Mean   :19.02
##  3rd Qu.:44.00
##  Max.   :52.00
```

```
# (a)
# Regression of weeksm1 on morekids
lm_robust(weeksm1 ~ morekids, data = fertility)
```

```
##              Estimate Std. Error  t value Pr(>|t|)  CI Lower  CI Upper     DF
## (Intercept) 21.068428 0.05606807 375.7652        0 20.958536 21.178320 254652
## morekids    -5.386996 0.08714918 -61.8135        0 -5.557806 -5.216185 254652
```

```
# On average, having more than two kids (as opposed to having exactly two kids)
# is associated with a -5.386996 change in the number of weeks worked by a woman
# in 1979

# (b)
# The coefficient of morekids in the regression of weeksm1 on morekids might not
# consistently estimate the causal effect of fertility on labour supply because
# morekids is likely to be correlated with unmodelled determinants of weeksm1.
# For example, women with physical disabilities are likely to have fewer
# children, and women with physical disabilities plausibly work fewer weeks in a
# year as a result. Orthogonality fails, and OLS regression of weeksm1 on
# morekids fails to consistently estimate the causal effect of fertility on
# labour supply.

# (c)
# Regression of morekids on boy1st with agem1, black, hispan, othrace as
# controls
lm_robust(morekids ~ boy1st + agem1 + black + hispan + othrace,
          data = fertility)
```

```
##                Estimate   Std. Error    t value      Pr(>|t|)    CI Lower
## (Intercept) -0.099598896 0.0084873828 -11.734936  8.599840e-32 -0.11623394
## boy1st      -0.009032707 0.0019066601  -4.737450  2.165409e-06 -0.01276971
## agem1        0.015361345 0.0002765887  55.538589  0.000000e+00  0.01481924
## black        0.100026321 0.0044483651  22.486086 7.304466e-112  0.09130764
## hispan       0.151216693 0.0041243904  36.664010 1.607555e-293  0.14313300
## othrace      0.027298251 0.0046497681   5.870884  4.340113e-09  0.01818483
```

```
##                 CI Upper     DF
## (Intercept) -0.082963852 254648
## boy1st      -0.005295704 254648
## agem1        0.015903452 254648
## black        0.108744998 254648
## hispan       0.159300388 254648
## othrace      0.036411673 254648
```

```
# H0: coefficient on boy1st in above regression = 0
# H1: coefficient on boy1st in above regression != 0
# p-value = 2.165409e-06
# Reject the null hypothesis that fertility is uncorrelated with whether or not
# the first child is a boy at any level of significance greater than
# 2.165409e-06 (including all conventional levels of significance). There is
# strong empirical evidence of a relationship between fertility and whether or
# not a woman's first child is a boy. One explanation of this relationship is
# that, in the 1970s, it was more expensive to raise a boy than a girl, hence
# women whose first child was a boy are more inclined to have more children.

# Regression of morekids on samesex, with agem1, black, hispan, othrace as
# controls
lm_robust(morekids ~ samesex + agem1 + black + hispan + othrace,
          data = fertility)
```

```
##                Estimate  Std. Error    t value      Pr(>|t|)    CI Lower
## (Intercept) -0.13953190 0.008459565 -16.493980  4.361026e-61 -0.15611242
## samesex      0.06800810 0.001900188  35.790187 7.836676e-280  0.06428378
## agem1        0.01538984 0.000275982  55.763925  0.000000e+00  0.01484892
## black        0.10052366 0.004442611  22.627156 3.031776e-113  0.09181626
## hispan       0.15122585 0.004116031  36.740698 9.757187e-295  0.14315854
## othrace      0.02750691 0.004639962   5.928261  3.065518e-09  0.01841271
##                CI Upper     DF
## (Intercept) -0.12295138 254648
## samesex      0.07173242 254648
## agem1        0.01593076 254648
## black        0.10923105 254648
## hispan       0.15929316 254648
## othrace      0.03660111 254648
```

```
# H0: coefficient on samesex in above regression = 0
# H1: coefficient on samesex in above regression != 0
# p-value = 7.836676e-280
# Reject the null hypothesis that fertility is uncorrelated with whether or not
# the first two children are of the same sex at any level of significance
# greater than 7.836676e-280 (including all conventional levels of
# significance). There is strong empirical evidence of a relationship between
# fertility and whether or not a woman's first two children are of the same sex.
# One explanation of this relationship is that, in the 1970s, families valued
# having children of different sexes, hence continued to have children if the
# first two were of the same sex.

# (d)
# samesex is relevant since there is a strong correlation between samesex and
# morekids (from the above regression). samesex is plausibly exogenous since
# (assuming there is no sex-selective abortion) whether or not a woman's first
```

```r
# two children are of the same sex is (more or less) random. samesex is
# plausibly excluded since there seems to be no direct causal relationship
# between whether or not a woman's first two children are of the same sex and
# the number of weeks she works in 1979.

# IV regression of weeksm1 on morekids with agem1, black, hispan, othrace as
# controls, and with samesex as an instrument for morekids
iv1 = iv_robust(weeksm1 ~ morekids + agem1 + black + hispan + othrace |
            samesex + agem1 + black + hispan + othrace,
          data = fertility, diagnostics = TRUE)

# Calculate residuals
fertility$u_hat = fertility$weeksm1 - iv1$fitted.values

# Regression of residuals on samesex, with agem1, black, hispan, othrace as
# controls
reg1 = lm_robust(u_hat ~ samesex + agem1 + black + hispan + othrace,
                 data = fertility)

# Hypothesis test
linearHypothesis(reg1, "samesex=0", test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## samesex = 0
##
## Model 1: restricted model
## Model 2: u_hat ~ samesex + agem1 + black + hispan + othrace
##
##   Res.Df Df  F Pr(>F)
## 1 254649
## 2 254648  1  0      1
```

```r
# H0: coefficient of samesex in the regression of the residual on samesex,
# agem1, black, hispan, othrace = 0
# H1: coefficient of samesex in the regression of the residual on samesex,
# agem1, black, hispan, othrace !=o
# p-value = 1
# Fail to reject the null hypothesis that the residual and samesex are
# uncorrelated at all levels of significance. There is strong evidence that
# samesex is exogenous.

# (e)
# IV regression of weeksm1 on morekids, with samesex as an instrument for
# weeksm1
iv_robust(weeksm1 ~ morekids | samesex, data = fertility)
```

```
##              Estimate Std. Error   t value     Pr(>|t|)  CI Lower  CI Upper
## (Intercept) 21.421092  0.4872506 43.963193 0.000000e+00 20.466094 22.376091
## morekids    -6.313685  1.2746857 -4.953131 7.307535e-07 -8.812035 -3.815335
##                 DF
## (Intercept) 254652
## morekids    254652
```

```
# Estimate of coefficient of morekids is more negative, the associated standard
# error is significantly larger (in the IV regression as oppose to the
# regression in (a). The difference in the value of the estimate suggests that
# omitted variable bias in the regression in (a) is positive. morekids is
# positively (negatively) correlated with negative (positive) unmodelled
# determinants of weeksm1. The standard error is larger in the IV regression
# because the component of morekids predicted by samesex is a poorer predictor
# of weeksm1 than morekids.

# (f)
# IV regression of weeksm1 on morekids, with agem1, black, hispan, othrace as
# controls, and with samesex as an instrument for morekids
iv_robust(weeksm1 ~ morekids + agem1 + black +hispan + othrace |
            samesex + agem1 + black + hispan + othrace,
          data = fertility)
```

```
##                Estimate Std. Error    t value      Pr(>|t|)   CI Lower
## (Intercept) -4.7918935 0.38979270 -12.293441  1.004902e-34 -5.5558768
## morekids    -5.8210509 1.24640073  -4.670288  3.009303e-06 -8.2639631
## agem1        0.8315975 0.02264086  36.729935 1.446269e-294  0.7872220
## black       11.6232731 0.23180281  50.142934  0.000000e+00 11.1689458
## hispan       0.4041802 0.26080282   1.549754  1.212018e-01 -0.1069863
## othrace      2.1309620 0.21099514  10.099579  5.606068e-24  1.7174172
##                CI Upper     DF
## (Intercept) -4.0279102 254648
## morekids    -3.3781388 254648
## agem1        0.8759730 254648
## black       12.0776004 254648
## hispan       0.9153468 254648
## othrace      2.5445068 254648
```

```
# Estimate of coefficient of morekids is more negative, the associated standard
# error is similar (in the IV regression with agem1, black, hispan, othrace as
# controls as opposed to the IV regression without controls).

# (g)
# Construct dummies
fertility$bothboys = as.numeric(fertility$boy1st == 1 & fertility$boy2nd == 1)
fertility$bothgirls = as.numeric(fertility$boy1st == 0 & fertility$boy2nd == 0)

# IV regression of weeksm1 on morekids, with agem1, black, hispan, othrace as
# controls, with bothboys and bothgirls as instruments for morekids
iv_robust(weeksm1 ~ morekids + agem1 + black +hispan + othrace |
            bothboys + bothgirls + agem1 + black + hispan + othrace,
          data = fertility)
```

```
##                Estimate Std. Error    t value      Pr(>|t|)   CI Lower
## (Intercept) -4.7513163 0.38891549 -12.216835  2.583958e-34 -5.5135802
## morekids    -5.4313132 1.21866591  -4.456770  8.323935e-06 -7.8198659
## agem1        0.8256123 0.02228542  37.047200 1.261299e-299  0.7819335
## black       11.5842728 0.23040216  50.278490  0.000000e+00 11.1326907
## hispan       0.3452363 0.25783473   1.338983  1.805775e-01 -0.1601129
## othrace      2.1203341 0.21092133  10.052725  9.027938e-24  1.7069340
##                CI Upper     DF
## (Intercept) -3.9890523 254648
```

```
## morekids    -3.0427606 254648
## agem1        0.8692911 254648
## black       12.0358549 254648
## hispan       0.8505854 254648
## othrace      2.5337343 254648
```

```
# Estimate of coefficient of morekids is less negative, the associated standard
# error is similar (in the IV regression with bothboys and bothgirls as
# instruments as opposed to the IV regression with samesex as an instrument).

# (h)
# IV regression of weeksm1 on morekids, with agem1, black, hispan, othrace as
# controls, with bothboys and bothgirls as instruments for morekids
iv2 = iv_robust(weeksm1 ~ morekids + agem1 + black +hispan + othrace |
                  bothboys + bothgirls + agem1 + black + hispan + othrace,
                data = fertility, diagnostics = TRUE)

summary(iv2)
```

```
##
## Call:
## iv_robust(formula = weeksm1 ~ morekids + agem1 + black + hispan +
##     othrace | bothboys + bothgirls + agem1 + black + hispan +
##     othrace, data = fertility, diagnostics = TRUE)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper     DF
## (Intercept)  -4.7513    0.38892 -12.217  2.584e-34  -5.5136  -3.9891 254648
## morekids     -5.4313    1.21867  -4.457  8.324e-06  -7.8199  -3.0428 254648
## agem1         0.8256    0.02229  37.047 1.261e-299   0.7819   0.8693 254648
## black        11.5843    0.23040  50.278  0.000e+00  11.1327  12.0359 254648
## hispan        0.3452    0.25783   1.339  1.806e-01  -0.1601   0.8506 254648
## othrace       2.1203    0.21092  10.053  9.028e-24   1.7069   2.5337 254648
##
## Multiple R-squared:  0.04345 ,   Adjusted R-squared:  0.04343
## F-statistic:  1390 on 5 and 254648 DF,  p-value: < 2.2e-16
##
## Diagnostics:
##                   numdf  dendf   value p.value
## Weak instruments      2 254647 667.764  <2e-16 ***
## Wu-Hausman            1 254647   0.432   0.511
## Overidentifying       1     NA   2.224   0.136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# H0: bothboys and bothgirls are exogenous
# H1: at least one of bothboys and bothgirls is endogenous
# p-value = 0.136
# Fail to reject the null hypothesis that the instruments bothboys and bothgirls
# are exogenous at any level of significance lower than 0.136 (including all
# conventional levels of significance). There is empirical reason to think
# bothboys and bothgirls are exogenous instruments.
```