Exam 3

Johan Bysted, Jonathan Arve og Mathias Kold

2024-05-16

Exam 3 - Instrumental variables

In a multiple linear regression (MLR) there are 6 assumptions. Assumption 1-5 are called Gauss-Markov assumptions and assumption 6 is called the normality assumption. The first 4 assumptions exist to secure that the model is unbiased, while assumption 5 checks for heteroskedasticity and assumption 6 checks for normality in the model. The assumptions are:

MLR1: Linear in Parameters

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters of interest and u is an unobserved random error or disturbance term.

MLR2: Random Sampling

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

MLR3: No Perfect Collinearity

In the sample (and therefore in the population), none of the independent variables is constant, and there are no exact linear relationships among the independent variables.

MLR4: Zero Conditional Mean

The error u has an expected value of zero given any values of the independent variables. In other words,

$$E(u|x_1, x_2, \ldots, x_k) = 0$$

MLR5: Homoskedasticity

The error u has the same variance given any value of the explanatory variables. In other words,

$$\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$$

MLR6: Normality

The population error u is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0, \sigma^2)$.

Consider the following model: $log(earning) = \beta_0 + \beta_1 educ + \beta_2 exp + \beta_3 male + \beta_4 ethblack + \beta_5 ethhisp + u$

where earnings are hourly wages in US dollars, educ is education measured in years of schooling, exp is work experience measured in years, male is a gender dummy, ethblack and ethhisp are race dummies for African Americans and Hispanics, respectively. Additionally, we have three instruments: mother's education measured in years (meduc), father's education measured in years (feduc), and number of siblings (sibling s).

1. Estimate the model using OLS and comment on the results.

```
model <- lm(log_earning~educ+exp+male+black+hisp)
summary(model)
```

```
##
## Call:
## lm(formula = log_earning ~ educ + exp + male + black + hisp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07585 -0.28006 -0.00145  0.30775  1.98441
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.396226   0.173508   2.284  0.02280 *
## educ         0.124220   0.009451  13.143  < 2e-16 ***
## exp          0.033882   0.005046   6.715 4.99e-11 ***
## male         0.293449   0.045803   6.407 3.36e-10 ***
## black       -0.195670   0.071255  -2.746  0.00624 **
## hisp        -0.097406   0.100342  -0.971  0.33213
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5103 on 514 degrees of freedom
## Multiple R-squared:  0.3539, Adjusted R-squared:  0.3476
## F-statistic: 56.32 on 5 and 514 DF,  p-value: < 2.2e-16
```

From the model, it can be seen that the education variable has an estimate of 0.124220 which means that one extra year of education will raise the hourly wage by approximately 12.42%. The experience variable has an estimate of 0.033882 which means that one more year of experience will raise the hourly wage by approximately 3.39%. The male variable has an estimate of 0.293449 which means that if you are a male your hourly wage will be approximately 29.34% higher than if you are a woman. The black variable has an estimate of -0.19567 which means that African Americans approximately will have 19.57% lower hourly wages than non-African Americans. The Hispanics variable has an estimate of -0.097406 which means that Hispanic approximately will have 9.74% lower hourly wages than non-Hispanics.

The intercept is 0.396226 which is the value of the dependent variable, in this case hourly wage in US dollars, when all other variables have a value of 0.

It can also be seen that all the variables except the hispanic variable are statistically significant at a 5% significance level due to p-values < 0.05. They are also significant at a 1% significance level.

2. Why might we be concerned that education is endogenous?

Given the significant positive relationship between education and earnings, it is important to consider if endogeneity is biasing the results. The potential endogeneity of education needs to be tested through

instrumental variables (IV) regression. There may be factors that affect both education and earnings that are not included in the model. For example family background could influence both education and earning potential. If these factors are not accounted for, the estimated coefficient on education will capture not only the effect of education but also the effect of these omitted variables, leading to bias.

3. Are sibling, meduc, and feduc useful as instruments?

To find out whether these variables are useful instruments, they have to fulfill two conditions; they need to be correlated with educ and they need to be independent of the error term, u:

$$Cov(x, z) \neq 0$$

$$Cov(z, u) = 0$$

In this case, x is the variable, z is the instruments and u is the error term in the model. If the instruments are correlated with the error term, these are also affected by an endogeneity issue. The instruments need to be uncorrelated with the omitted variable, that creates an endogeneity issue for educ. The first condition, $Cov(x, z) \neq 0$, can be tested. A model for the variable educ as the dependent variable can be set up. When multiple IVs are used, 2SLS is used:

$$educ = \pi_0 + \pi_1 exp + \pi_2 male + \pi_3 black + \pi_4 hisp + \pi_5 sib + \pi_6 meduc + \pi_7 feduc + v$$

Setting up the hypothesis test:

$$H_0 : \pi_5 = 0, \pi_6 = 0 \; or \; \pi_7 = 0$$

$$H_1 : \pi_5 \neq 0, \pi_6 \neq 0 \; or \; \pi_7 \neq 0$$

By using F-tests, we can test the hypothesis. F-stat is calculated using the following formula:

$$F = \frac{R^2_{ur} - R^2_r}{1 - R^2_{ur}} * \frac{n - k - 1}{q}$$

```
sib <- data3$siblings; meduc <- data3$meduc; feduc <- data3$feduc
ivreg <- lm(educ~exp+male+black+hisp+sib+meduc+feduc)
summary(ivreg)
```

```
##
## Call:
## lm(formula = educ ~ exp + male + black + hisp + sib + meduc +
##     feduc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8750  -1.3575  -0.2666   1.4341   8.9572
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.240160   0.653682  17.195  < 2e-16 ***
## exp         -0.116651   0.020372  -5.726 1.75e-08 ***
## male         0.003102   0.190843   0.016   0.9870
## black       -0.631488   0.299167  -2.111   0.0353 *
## hisp         0.658727   0.438636   1.502   0.1338
## sib         -0.102192   0.046651  -2.191   0.0289 *
## meduc        0.261472   0.047138   5.547 4.67e-08 ***
## feduc        0.147712   0.035030   4.217 2.93e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.118 on 512 degrees of freedom
## Multiple R-squared:  0.2778, Adjusted R-squared:  0.2679
## F-statistic: 28.13 on 7 and 512 DF,  p-value: < 2.2e-16
```

It can be seen, that all three factors are significant at a 5% significance level.

```
ivtest <- c("sib=0","meduc=0","feduc=0")
linearHypothesis(ivreg,ivtest)
```

```
## Linear hypothesis test
##
## Hypothesis:
## sib = 0
## meduc = 0
## feduc = 0
##
## Model 1: restricted model
## Model 2: educ ~ exp + male + black + hisp + sib + meduc + feduc
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    515 2914.6
## 2    512 2297.8  3    616.75 45.809 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value < 0.05 H_0 is rejected, hence the three variables are statistically significant. Thereby it can be assumed, that the three variables are correlated with education, and can be used as instruments for education.

4. Test whether education is endogenous.

First we find the reduced form residuals from the linear regression including the three Instrument Variables. The reduced form residuals is denoted as $\hat{v}$. $\hat{v}$ is then included in the original OLS model, and if the residuals are significant, it means that there is an endogeneity issue:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_2 + \delta v + error$$

Thereby we get the nullhypothesis:

$$H_0 : \delta = 0$$

```
res <- ivreg$residuals
res_mod <- lm(log_earning~educ+exp+male+black+hisp+res)
coeftest(res_mod)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
```

```
## (Intercept) -0.064700   0.339003 -0.1909   0.84871
## educ         0.153036   0.020516  7.4593 3.742e-13 ***
## exp          0.037628   0.005567  6.7591 3.784e-11 ***
## male         0.290479   0.045775  6.3458 4.869e-10 ***
## black       -0.157544   0.075122 -2.0972   0.03647 *
## hisp        -0.069476   0.101739 -0.6829   0.49499
## res         -0.036550   0.023106 -1.5818   0.11430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since residuals are not significant, we fail to reject H_0, and thereby we cannot conclude that there is an endogeneity issue. This does not mean that educ does not have an endogeneity issue, but for these IV variables there is not.

5. Estimate the model using 2SLS employing the three described instruments. Compare with the results in question 1.

The p-value of 0.11 is not sufficient to reject that there is an endogeneity issue, why it is deemed necessary to further examine if there is an issue. To accomodate this issue, we use the 2SLS method. A regression on this form is set up:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 exp + \beta_3 male + \beta_4 black + \beta_5 hisp + u$$

where y_2=educ, the variable that is suspected to have endogeneity. Now, a regression for educ is made. The three IV's for educ is added in the regression:

$$y_2 = \pi_0 + \pi_1 exp + \pi_2 male + \pi_3 black + \pi_4 hisp + \pi_5 sib + \pi_6 meduc + \pi_7 feduc + v$$

In this regression, the following assumptions is expected to be fulfilled:

- $E(v) = 0$. The mean of the error term is 0, which indicates unbiasedness.
- $Cov(x, v) = 0$. This applies for all variables in the original model.
- $Cov(z, v) = 0$. The IV's should be independent of the error term.

```
stage1 <- lm(educ~exp+male+black+hisp+sib+meduc+feduc)
educ_fitted <- fitted(stage1)
```

Before proceeding, we need to confirm that the IV's are significant variables for educ:

```
linearHypothesis(stage1,c("meduc=0", "feduc=0", "sib=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## meduc = 0
## feduc = 0
## sib = 0
##
## Model 1: restricted model
## Model 2: educ ~ exp + male + black + hisp + sib + meduc + feduc
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
```

```
## 1    515 2914.6
## 2    512 2297.8  3    616.75 45.809 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value < 0.05, H_0 is rejected, and the IV's are significant. The fitted values of educ is used as IV. A new expression for y_2 is found: $y_2 = \hat{y_2} + v$ This new expression can be added to the original regression:

$$y_1 = \beta_0 + \beta_1(\hat{y_2} + v) + \beta_2 exp + \beta_3 male + \beta_4 black + \beta_5 hisp + u$$

```
stage2 <- lm(log_earning~educ_fitted+exp+male+black+hisp)
sls <- ivreg(log_earning~educ+exp+male+black+hisp|exp+male+black+hisp+sib+feduc+meduc)
```

Setting the new 2SLS and the OLS up:

```
screenreg(list(OLS=model,TwoSLS=sls), digits=4)
```

```
##
## ========================================
##              OLS           TwoSLS
## ----------------------------------------
## (Intercept)    0.3962 *     -0.0647
##               (0.1735)      (0.3426)
## educ           0.1242 ***    0.1530 ***
##               (0.0095)      (0.0207)
## exp            0.0339 ***    0.0376 ***
##               (0.0050)      (0.0056)
## male           0.2934 ***    0.2905 ***
##               (0.0458)      (0.0463)
## black         -0.1957 **    -0.1575 *
##               (0.0713)      (0.0759)
## hisp          -0.0974       -0.0695
##               (0.1003)      (0.1028)
## ----------------------------------------
## R^2            0.3539        0.3422
## Adj. R^2       0.3476        0.3358
## Num. obs.    520           520
## ========================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

It can be seen, that when using the 2SLS, the estimate for educ increases a bit, and afro american as well as hispanic falls. Male and experience does not really change.

6. Perform the overidentification test. What do you conclude?

First, the 2SLS residuals, $\hat{u}$, from the earlier estimated 2SLS model are obtained.

```
resid2sls <- resid(sls)
```

Then the residuals are regressed on all exogenous variables including the IV's, where R-squared, R_1^2 is obtained.

```
res.aux <- lm(resid2sls ~ exp + male + black + hisp + sib + meduc + feduc)
```

The null hypothesis is that all IV's are uncorrelated with u:

$$H_0 : \text{Corr}(Z, u) = 0, \quad \text{where} \quad Z = (z_1, z_2, z_3).$$

We use $nR_1^2 \sim \chi_q^2$, where q is the number of instrumental variables from outside the model minus the total number of endogenous explanatory variables. Hypothesis will be tested:

```
r2 <- summary(res.aux)$r.squared
n <- nobs(res.aux)
teststat <- n*r2
pval <- 1-pchisq(teststat,df=2)
pval
```

```
## [1] 0.01455821
```

From the calculated p-value, H_0 is rejected at a 5% significance level due to the p-value < 0.05. This means that at least some of the IV's are not exogenous, which means that the model does not pass the test and is assumed to have overidentification.

7. Perform the entire analysis again using only meduc and feduc as instruments. Does this change your conclusions?

We perform the entire analysis again, starting from 3.3, because 3.1 and 3.2 are not about the choice of instrumental variables. The tests and methods will not be elaborated further, so for an explanation of what we are doing look at the earlier exercises.

We start by testing whether meduc and feduc are useful as instruments using the same method as in 3.3:

```
ivreg7 <- lm(educ ~ exp + male + black + hisp + meduc + feduc)
linearHypothesis(ivreg7, c("meduc=0","feduc=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## meduc = 0
## feduc = 0
##
## Model 1: restricted model
## Model 2: educ ~ exp + male + black + hisp + meduc + feduc
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    515 2914.6
## 2    513 2319.3  2    595.22 65.826 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value < 0.05 H_0 is rejected, hence the two variables are statistically significant, why it can be assumed, that meduc and feduc are correlated with education, and can be used as instruments for education.

Now we want to test if there is endogeneity issues in the new model with the two instrumental variables. The same method as in 3.4 will be used:

```
res7 <- ivreg7$residuals
res_mod7 <- lm(log_earning ~ educ + exp + male + black + hisp + res7)
coeftest(res_mod7)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -0.1657973  0.3443473 -0.4815   0.63038
## educ         0.1593561  0.0208626  7.6384 1.086e-13 ***
## exp          0.0384495  0.0055843  6.8853 1.691e-11 ***
## male         0.2898274  0.0457297  6.3378 5.107e-10 ***
## black       -0.1491819  0.0752218 -1.9832   0.04787 *
## hisp        -0.0633502  0.1017049 -0.6229   0.53364
## res7        -0.0441531  0.0233868 -1.8879   0.05960 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We fail to reject the null hypothesis at a 5% significance level, and thereby we cannot conclude that there is an endogeneity issue. If we used a 10% significance level then we could have rejected the null hypothesis and instead accepted the alternative hypothesis that there is an endogeneity issue, indicating potential endogeneity issues.

We perform the 2SLS method again as in 3.5, but this time with only meduc and feduc as instrumental variables, and we do it directly with the code in R:

```
sls2 <- ivreg(log_earning ~ educ + exp + male + black + hisp | exp + male + black + hisp + feduc + medu
```

To look whether it changes our conclusions we print it together with the OLS from 3.1 and the 2SLS from 3.5:

```
screenreg(list(OLS=model, SLS3_5 = sls, SLS3_7=sls2), digits = 4)
```

```
##
## ================================================
##              OLS           SLS3_5        SLS3_7
## ------------------------------------------------
## (Intercept)    0.3962 *     -0.0647       -0.1658
##               (0.1735)      (0.3426)      (0.3498)
## educ           0.1242 ***    0.1530 ***    0.1594 ***
##               (0.0095)      (0.0207)      (0.0212)
## exp            0.0339 ***    0.0376 ***    0.0384 ***
##               (0.0050)      (0.0056)      (0.0057)
## male           0.2934 ***    0.2905 ***    0.2898 ***
##               (0.0458)      (0.0463)      (0.0465)
## black         -0.1957 **    -0.1575 *     -0.1492
##               (0.0713)      (0.0759)      (0.0764)
## hisp          -0.0974       -0.0695       -0.0634
##               (0.1003)      (0.1028)      (0.1033)
## ------------------------------------------------
## R^2            0.3539        0.3422        0.3366
## Adj. R^2       0.3476        0.3358        0.3301
```

```
## Num. obs.    520           520           520
## ================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

In the above we can see the original model and the different results we get with using 3 IV's and 2 IV's to estimate the fitted values of education. We performed overidentification test in 3.6 when using 3 IV's, where we found that the model was indeed overidentificated, so one of the IV's was not exogenous. The same overidentification test is now performed on the new model with two IV's:

```
resid2sls2 <- resid(sls2)
res.aux2 <- lm(resid2sls2 ~ exp + male + black + hisp + meduc + feduc)
r2_2 <- summary(res.aux2)$r.squared
n2 <- nobs(res.aux2)
teststat2 <- n2*r2_2
pval2 <- 1-pchisq(teststat2,df=1) #now only 1 degree of freedom, because only 2 IV's and 1 endogenous v
pval2
```

```
## [1] 0.01593513
```

The new model with 2 IV's also rejects the null hypothesis. This means that at least one of the IV's are not exogenous, which means that the model does not pass the test and is assumed to have overidentification.