ECO-305 (Econometrics 2) Project

Tackling Endogeneity by the Demonstration of Two Stage Least Squares (2SLS) Method on MROZ Dataset

Created by

1. Aayushman (20004)
2. Rajkishan (20223)
3. Vaibhav Mahadwad (20299)
4. Nikhil Deepak Patil (20328)
5. Prateek Mishra (20347)

INDIAN INSTITUTE OF SCIENCE EDUCATION AND
RESEARCH BHOPAL

DEPARTMENT OF ECONOMIC SCIENCES
IISER BHOPAL

Demonstration of Two Stage Least Squares (2SLS) Method on MROZ Dataset ECO-305 Project

1 WHAT IS TWO STAGE LEAST SQUARES (2SLS) METHOD?

The 2SLS method is a statistical technique that is used in the analysis of structural equations.They are used when the regression model has the problem of endogeneity.
Endogenous variables have values that are determined by other variables in the system. Having endogenous regressors in a model will cause ordinary least squares estimators to fail, as one of the assumptions of OLS is that there is no correlation between an predictor variable and the error term.
The solution to this is to use **Instrumental Variables**. It is a variable that is uncorrelated with the error term, but correlated with a particular independent variable.

2 MROZ DATASET

This dataset is is based on the sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions.

It's variables are as follows:

- work - work at home in 1975?
- hours - wife's hours of work in 1975
- kidslt6 - number of children less than 6 years old in household
- kidsge6 - number of children between ages 6 and 18 in household
- age - wife's age
- educ - wife's educational attainment, in years
- wage - wife's average hourly earnings, in 1975 dollars
- repwage - wife's wage reported at the time of the 1976 interview
- hushrs - husband's hours worked in 1975
- husage - husband's age
- huseduc - husband's educational attainment, in years
- huswage - husband's wage, in 1975 dollars
- faminc - family income, in 1975 dollars
- mtr -
- motheduc - wife's mother's educational attainment, in years
- fatheduc - wife's father's educational attainment, in years
- unem - unemployment rate in county of residence, in percentage points
- city - lives in large city (SMSA) ?
- exper - actual years of wife's previous labor market experience
- nwifeinc -
- lwage - log of wages earned
- expersq - square of experience, in years*years

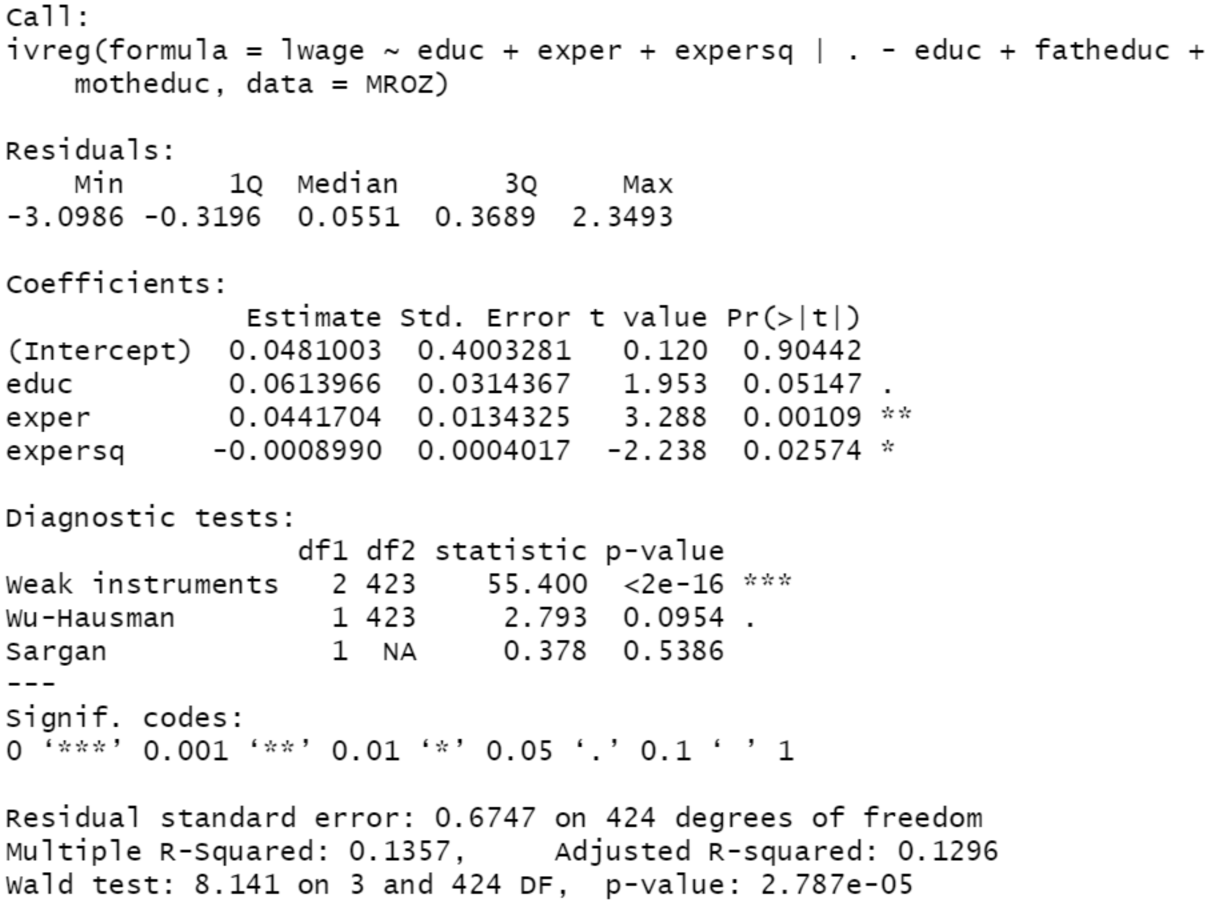
3 ECONOMETRIC MODEL

3.1 GENERAL MODEL FOR 2SLS

The 2SLS model for our MROZ dataset is as follows:

$$lwage = \alpha + \beta_1 educ + \beta_2 exper + \beta_3 expersq$$

The results of the regression followed by it's summary is as follows:



```
Call:
ivreg(formula = lwage ~ educ + exper + expersq | . - educ + fatheduc +
    motheduc, data = MROZ)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0986 -0.3196  0.0551  0.3689  2.3493 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  0.0481003  0.4003281   0.120  0.90442   
educ         0.0613966  0.0314367   1.953  0.05147 . 
exper        0.0441704  0.0134325   3.288  0.00109 **
expersq     -0.0008990  0.0004017  -2.238  0.02574 * 

Diagnostic tests:
                 df1 df2 statistic p-value    
Weak instruments   2 423    55.400  <2e-16 ***
Wu-Hausman         1 423     2.793  0.0954 .  
Sargan             1  NA     0.378  0.5386    
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6747 on 424 degrees of freedom
Multiple R-Squared: 0.1357,	Adjusted R-squared: 0.1296 
Wald test: 8.141 on 3 and 424 DF,  p-value: 2.787e-05 
```

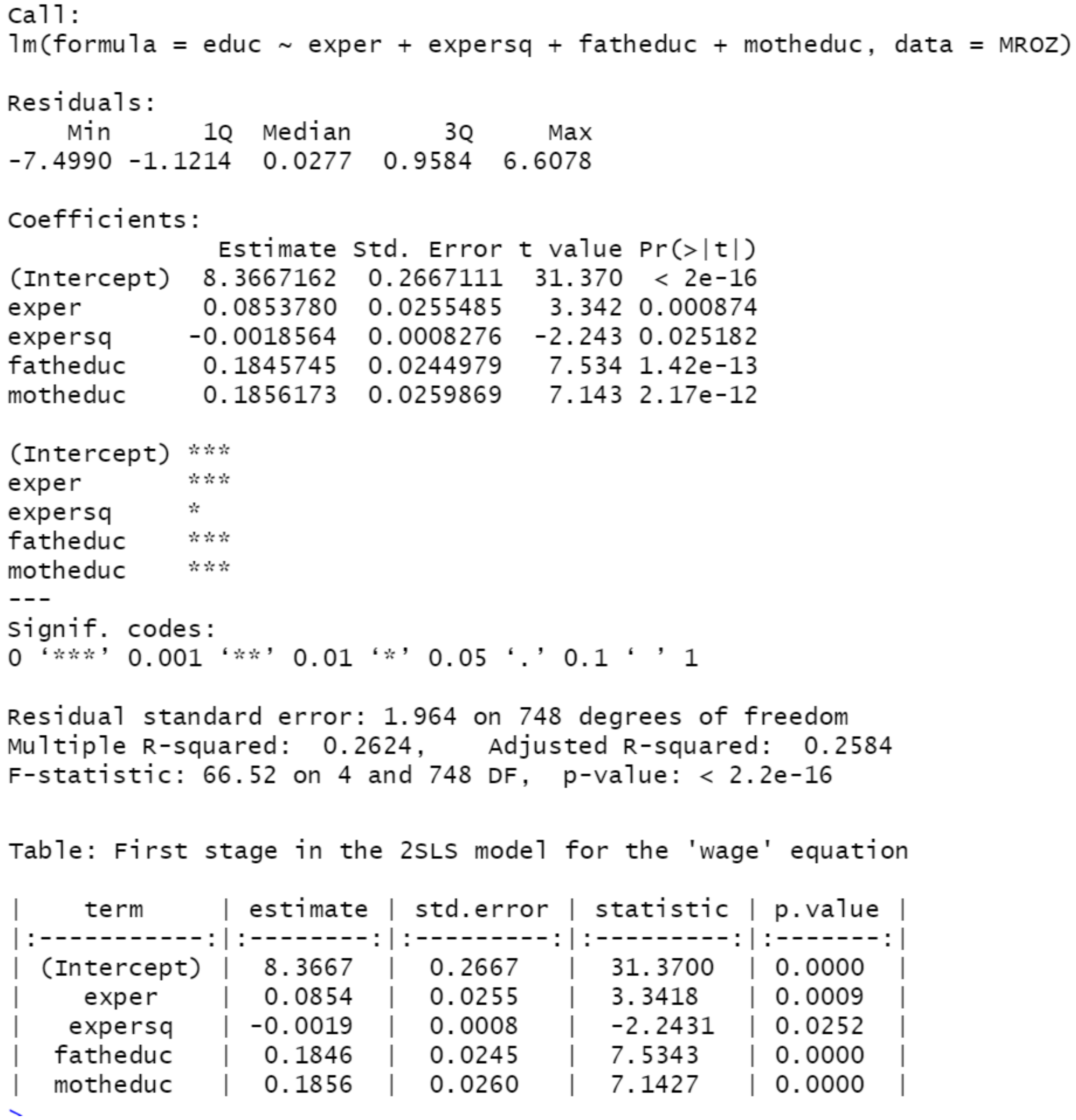
INFERENCE DRAWN The Wu-Hausman test for endogeneity barely rejects the null hypothesis that the variable of concern is uncorrelated with the error term, indicating that *educ* is marginally endogenous.

3.2 FIRST STAGE MODEL FOR 2SLS

This regression model is run with the endogenous variable $educ$ on instrument variables $fatheduc$ and $motheduc$

$$educ = \delta + \gamma_2 exper + \gamma_3 expersq + \gamma_4 fatheduc + \gamma_5 motheduc$$

The results of the regression followed by it's summary is as follows:



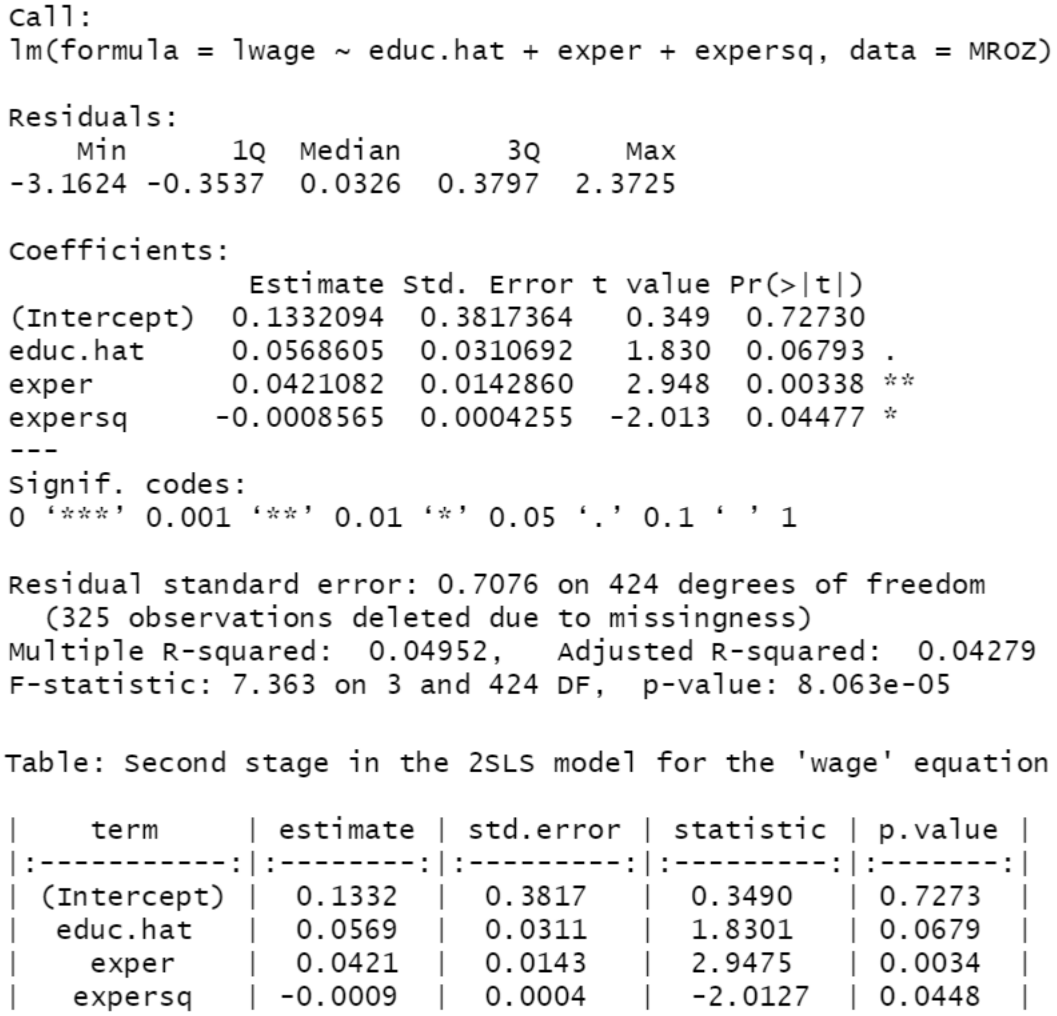
```
Call:
lm(formula = educ ~ exper + expersq + fatheduc + motheduc, data = MROZ)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.4990 -1.1214  0.0277  0.9584  6.6078 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.3667162  0.2667111  31.370  < 2e-16
exper        0.0853780  0.0255485   3.342 0.000874
expersq     -0.0018564  0.0008276  -2.243 0.025182
fatheduc     0.1845745  0.0244979   7.534 1.42e-13
motheduc     0.1856173  0.0259869   7.143 2.17e-12

(Intercept) ***
exper       ***
expersq     *  
fatheduc    ***
motheduc    ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.964 on 748 degrees of freedom
Multiple R-squared:  0.2624,	Adjusted R-squared:  0.2584 
F-statistic: 66.52 on 4 and 748 DF,  p-value: < 2.2e-16


Table: First stage in the 2SLS model for the 'wage' equation

|    term     | estimate | std.error | statistic | p.value |
|:-----------:|:--------:|:---------:|:---------:|:-------:|
| (Intercept) |  8.3667  |  0.2667   |  31.3700  | 0.0000  |
|    exper    |  0.0854  |  0.0255   |  3.3418   | 0.0009  |
|   expersq   | -0.0019  |  0.0008   |  -2.2431  | 0.0252  |
|  fatheduc   |  0.1846  |  0.0245   |  7.5343   | 0.0000  |
|  motheduc   |  0.1856  |  0.0260   |  7.1427   | 0.0000  |
```

3.3 SECOND STAGE MODEL FOR 2SLS

before runnign this model, we calculated the predicted values for $educ$ (denoted by $educ.hat$), which replaces the former in the equation.

$$educ.hat = \psi + \sigma_2 exper + \sigma_3 expersq + \sigma_4 fatheduc + \sigma_5 motheduc$$

The results of the regression followed by it's summary is as follows:



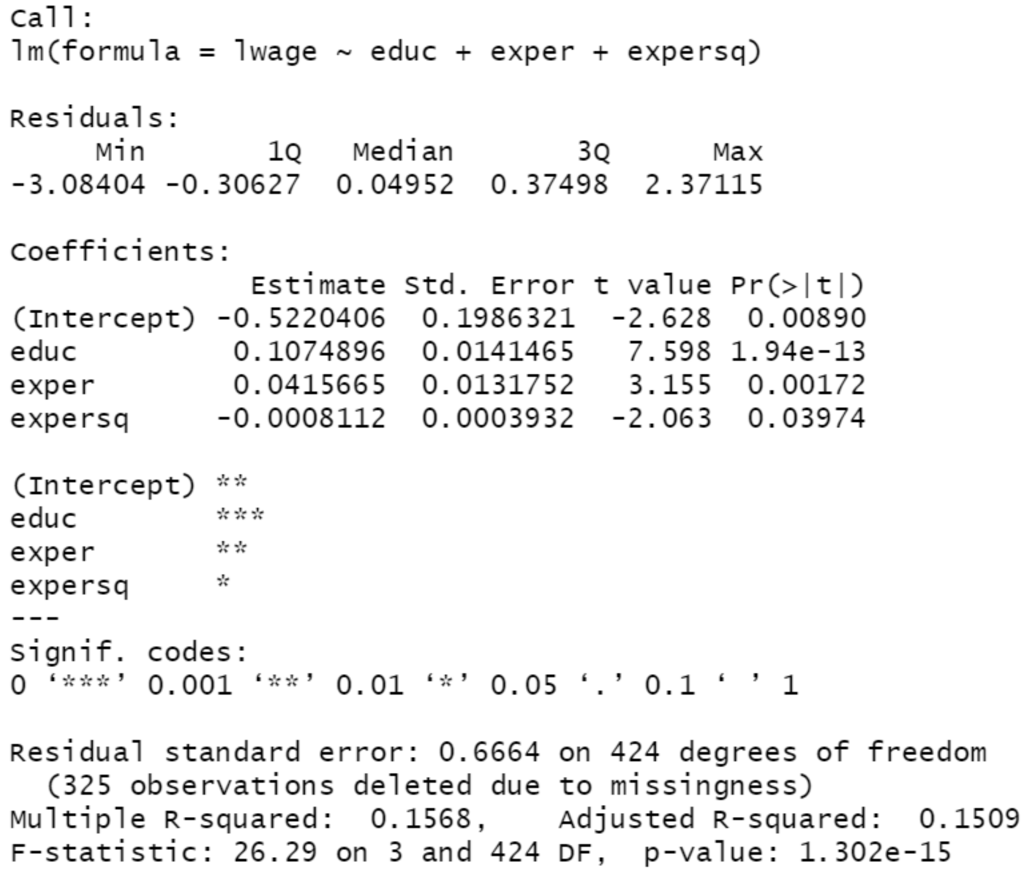
```
Call:
lm(formula = lwage ~ educ.hat + exper + expersq, data = MROZ)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.1624 -0.3537  0.0326  0.3797  2.3725 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  0.1332094  0.3817364   0.349  0.72730   
educ.hat     0.0568605  0.0310692   1.830  0.06793 . 
exper        0.0421082  0.0142860   2.948  0.00338 **
expersq     -0.0008565  0.0004255  -2.013  0.04477 * 
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7076 on 424 degrees of freedom
  (325 observations deleted due to missingness)
Multiple R-squared:  0.04952,   Adjusted R-squared:  0.04279 
F-statistic: 7.363 on 3 and 424 DF,  p-value: 8.063e-05


Table: Second stage in the 2SLS model for the 'wage' equation

|    term     | estimate | std.error | statistic | p.value |
|:-----------:|:--------:|:---------:|:---------:|:-------:|
| (Intercept) |  0.1332  |  0.3817   |  0.3490   | 0.7273  |
|  educ.hat   |  0.0569  |  0.0311   |  1.8301   | 0.0679  |
|    exper    |  0.0421  |  0.0143   |  2.9475   | 0.0034  |
|   expersq   | -0.0009  |  0.0004   |  -2.0127  | 0.0448  |
```

3.4 Hausman Test for Endogeneity of Regressors

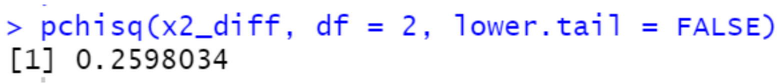
The Hausman Test (also called the Hausman specification test) detects endogenous regressors (predictor variables) in a regression model. For this test, we run a simple OLS regression model as follows:

$$lwage = c_0 + b_1 educ + b_2 exper + b_3 expersq$$

The results of the regression followed by it's summary is as follows:



```
Call:
lm(formula = lwage ~ educ + exper + expersq)

Residuals:
     Min       1Q   Median       3Q      Max 
-3.08404 -0.30627  0.04952  0.37498  2.37115 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.5220406  0.1986321  -2.628  0.00890
educ         0.1074896  0.0141465   7.598 1.94e-13
exper        0.0415665  0.0131752   3.155  0.00172
expersq     -0.0008112  0.0003932  -2.063  0.03974

(Intercept) ** 
educ        ***
exper       ** 
expersq     *  
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6664 on 424 degrees of freedom
  (325 observations deleted due to missingness)
Multiple R-squared:  0.1568,	Adjusted R-squared:  0.1509 
F-statistic: 26.29 on 3 and 424 DF,  p-value: 1.302e-15
```



```
> pchisq(x2_diff, df = 2, lower.tail = FALSE)
[1] 0.2598034
```

4 R Code

The R code for this project has been attached below:

```
library(AER)
library(knitr) ## For Making neat table using kable()
library(broom) ## For making neat table using tidy()

########### Importing data
library(readr)
MROZ <- read_csv("MROZ.csv")
View(MROZ)
attach(MROZ)
summary(MROZ)
###########

# 2SLS Model
model <- ivreg(lwage ~ educ + exper + expersq | . - educ + fatheduc + motheduc
                data = MROZ)
summary(model, diagnostics = TRUE)

############
#(Wu-)Hausman test for endogeneity: barely rejects the null that the variable
#concern is uncorrelated with the error term, indicating that  educ  is
#marginally endogenous
############

# 2SLS - first stage
# Regression of endogenous variable educ on instruments fatheduc and motheduc
tsls1 <- lm(educ ~ exper + expersq + fatheduc + motheduc, MROZ)
summary(tsls1,)
kable(tidy(tsls1), digits = 4, align = 'c',caption =
            "First stage in the 2SLS model for the 'wage' equation")

# Predicted values for educ_hat
educ.hat<-fitted.values(tsls1)

# 2SLS - second stage
# Replace educ with predicted value educ_hat
tsls2 <- lm(lwage ~ educ.hat+ exper + expersq, MROZ)
summary(tsls2)
kable(tidy(tsls2), digits = 4, align = 'c', caption =
             "Second stage in the 2SLS model for the 'wage' equation")
```

```
###########
#Hausman test for endogeneity  of regressors
olsreg <- lm(lwage ~ educ + exper + expersq)
summary(olsreg)

cf_diff <- coef(model) - coef(olsreg)
vc_diff <- vcov(model) - vcov(olsreg)

x2_diff <- as.vector(t(cf_diff) %*% solve(vc_diff) %*% cf_diff)
pchisq(x2_diff, df = 2, lower.tail = FALSE)
```