ECO 372: Introduction to Econometrics

Spring 2025

Lecture 8: Endogeneity and Emergence of Instrumental Variable Regressions-II

Sakib Bin Amin, Ph.D.

Associate Professor in Economics
Director, Accreditation Project Team (APT)

Outline

Our objectives for this lecture will be to learn

- The Two Stage Least Squares (TSLS) Estimator
- Derivation of TSLS Estimator
- Endogeneity Detection
- Instrument Validity Check
- Overidentification Restriction
- Inference Using TSLS
- A Brief Discussion on Generalised Method of Moments (GMM)
- Example of TSLS
- Applications of IV Estimation

Conditions for Valid Instruments

- A valid instrument needs to satisfy two conditions:

  1. Instrument relevance: $corr(\, Z_i | X_i) \neq 0$ [the magnitude must be is very high]

  2. Instrument exogeneity: $corr(\, Z_i | u_i) = 0$ [if not zero then it must be very close to zero]

- The first assumption requires that there is some association between the instrument and the variable being instrumented.

- The second assumption requires that instrument is exogenous. Then that part of the variation of X_i captured by the instrumental variable is exogenous.

General Framework of IV Models

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \psi_1 R_{1i} + \cdots + \psi_j R_{ji} + u_i$$

i=runs over all observations, i={1, 2, 3,…,N}

Y_i=dependent variable

X_i= independent variable

R_{ji}= control cariables

β_0= intercept of the regression line

$\beta_1, \dots, \beta_k$= slopes/predictors [any or all can be endogenous]

$\psi_1, \dots, \psi_j$= slopes of exogenous controls

u_i= error term

$Z_{1i}, \dots, Z_{mi}$= Instruments

General Framework of IV Models

- If the number of instruments equals (e.g.,skills) the number of endogenous regressors (e. g., education): exactly identified model.

- If the number of instruments greater than the number of endogenous regressors: overidentified model.

- If the number of instruments less than the number of endogenous regressors: underidentified model.

- Having (at least) one instrument for any single endogenous regressor is essential. Otherwise, computation is not possible.

The Two Stage Least Squares (TSLS) Estimator

- If the instrument Z satisfies the conditions of instrument relevance and exogeneity, coefficient β_1 can be estimated using an IV estimator called Two Stage Least Squares (TSLS).
- As the name suggests, the TSLS estimator is done in two stages.
- First stage decomposes X into two components: a problematic component that may be correlated with the regression error and another, problem-free component that is uncorrelated with the error.
- The second stage uses the problem-free component to estimate β_1.
- The first stage begins with a regression linking X and Z:

$$X_i = \eta_0 + \eta_1 Z_i + u_i$$

- This regression provides the needed decomposition of X_i. $\eta_0 + \eta_1 Z_i$ is problem free part.

The Two Stage Least Squares (TSLS) Estimator

- The idea behind TSLS is to use the problem-free component and disregard error component.

- The first stage of TSLS uses the predicted value from the OLS regression:

$$\hat{X}_i = \hat{\eta}_0 + \hat{\eta}_1 Z_i$$

- $\hat{\eta}_0$ and $\hat{\eta}_1$ are the OLS estimates and $\hat{X}_i$ is the predicted value.

- The second stage of TSLS is easy. Regress Y_i on $\hat{X}_i$ using OLS:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$$

- The resulting estimators from the second-stage regression are the TSLS estimators $\hat{\beta}_0^{TSLS}$ and $\hat{\beta}_1^{TSLS}$.

Derivation of the TSLS Estimator

- Recall OLS derivation:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

- This expression can be further simplified in terms of variance and covariance:

$$\hat{\beta}_1 = \frac{S_{XY}}{S_X^2} = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

- $\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})$ is just sample variance and $\sum_{i=1}^{N}(X_i - \bar{X})^2$ is variance of X_i.

Derivation of the TSLS Estimator

- The first stage of TSLS is to regress $X_i$on the instrument Z_i by OLS and then compute the OLS predicted value of X_i.

- The second stage is to regress Y_i on $\hat{X}_i$ by OLS.

- So, TSLS estimator expressed in terms of the predicted value:

$$\hat{\beta}_1^{TSLS} = \frac{S_{\hat{X}Y}}{S_{\hat{X}}^2}$$

- We know $\hat{X}_i = \hat{\eta}_0 + \hat{\eta}_1 Z_i$

- Because $\hat{X}_i$ is predicted, sample variances and covariances imply that: $S_{\hat{X}Y} = \hat{\eta}_1 S_{ZY}$ and $S_{\hat{X}}^2 = \hat{\eta}_1^2 S_Z^2$

Derivation of the TSLS Estimator

$S_{\hat{X}Y}$ $=cov\left(\hat{X}_i, Y_i\right)$

$=cov(\hat{\eta}_0 + \hat{\eta}_1 Z_i, Y_i)$ $[\hat{X}_i = \hat{\eta}_0 + \hat{\eta}_1 Z_i]$

$=cov(\hat{\eta}_0, Y_i)+cov(\hat{\eta}_1 Z_i, Y_i)$

$=0 + cov(\hat{\eta}_1 Z_i, Y_i)$ [rule: covariance of a constant and a random variable is zero]

$=\hat{\eta}_1 cov(Z_i, Y_i)$ [rule: attached coefficients comes outside of the operator]

$=\hat{\eta}_1 S_{ZY}$

Similarly, we can work our way out and get $S_{\hat{X}}^2 = \hat{\eta}_1^2 S_Z^2$

$S_{\hat{X}}^2 = Var(\hat{X}_i)$. The rest is straight forward. Remember, variance of a constant is zero.

Derivation of the TSLS Estimator

- Hence, the TSLS estimator can be written as: $\hat{\beta}_1^{TSLS} = \frac{s_{XY}}{s_{\hat{X}}^2} = \frac{s_{ZY}}{\hat{\eta}_1 s_Z^2}$

- We can derive $\hat{\eta}_1$ as $\frac{s_{ZX}}{s_Z^2}$. Because $\hat{\eta}_1$ is the slope of the first stage OLS.

- Plug the value in the $\hat{\beta}_1^{TSLS}$ formula. Hence,

$$\hat{\beta}_1^{TSLS} = \frac{S_{ZY}}{S_{ZX}^2}$$

- We can go back to our known representation of OLS:

$$\hat{\beta}_1^{TSLS} = \frac{\sum_{i=1}^{N}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(Z_i - \bar{Z})(X_i - \bar{X})}$$

Derivation of the TSLS Estimator

- $\hat{\beta}_1^{TSLS}$ that we have just seen can be derived in a different way:

- We can say $\hat{\beta}_1^{TSLS} = \frac{\sum_{i=1}^{N}(Z_i-\bar{Z})(Y_i-\bar{Y})}{\sum_{i=1}^{N}(Z_i-\bar{Z})(X_i-\bar{X})} = \frac{Cov(Z_i,Y_i)}{Cov\ (Z_i,X_i)}$

$$Cov(Z_i, Y_i) = Cov(Z_i, \beta_0 + \beta_1 X_i + u_i)$$

$$= \beta_1 Cov(Z_i, X_i) + Cov(Z_i, u_i)$$

$$= \beta_1 Cov(Z_i, X_i) + 0$$

- $Cov(Z_i, u_i)$ =0 because there should not be any covariance between Z_i and u_i
- From here, we can easily find $\hat{\beta}_1^{TSLS} = \frac{Cov(Z_i,Y_i)}{Cov\ (Z_i,X_i)}$

Endogeneity Detection

- How to confirm whether the regressor is endogenous or exogenous?
- Well, there are two ways: intuitively and computationally.
- It is strongly suggested that the variable should be tested in both ways.
- After intuitive justification, one should test whether, given the data that intuition statistically holds or not.
- The Durbin (1954), Wu-Hausman (Wu 1974; Hausman 1978), and Wooldridge's (1995) robust score test are some examples.
- In all cases, if the test statistic is significant, then the variables being tested must be treated as endogenous.

Instrument Validity Check

- Whether IV regression is useful in a given application hinges on whether the instruments are valid.
- Invalid instruments produce meaningless results.
- It therefore is essential to assess whether a given set of instruments is valid in a particular application.
- Instruments that explain little of the variation in X are called weak instruments.
- If the instruments are weak, then the normal distribution provides a poor approximation to the sampling distribution of the TSLS estimator, even if the sample size is large.
- A Rule of Thumb for Checking for Weak Instruments:
  - The first-stage F-statistics. If $Z_{1i}, \dots, Z_{mi}$=0 (as null hypothesis) is proved, then instruments are weak.
  - First-stage F-statistics less than 10 for single endogenous regressor, then instrument(s) is weak.

Overidentification Restriction

- Overidentification is when we have more instruments than endogenous variables.
- Sometimes using multiple instruments may lead to non-zero correlation with the TSLS residual and other variables like controls.
- Therefore, we need to constantly check whether our chosen instruments are valid.
- We need to restrict inclusion of instruments if we find above conditions is true. This can be done by dropping instruments one by one and checking whether we are in the safe zone.
- The null hypothesis here states that residuals of the TSLS (depending on Zs) and Rs (i.e. Controls), should be uncorrelated.
  - This becomes very difficult when you apply in real world survey data due to the underlying data dynamics.
- Anderson and Rubin's (1950) chi-squared test, Hansen's (1982) J-statistic, Basmann's F-test, etc. are some of the examples of overidentification restriction.

Inference Using TSLS

- Statistical inference proceeds in the usual way.

- The justification is as usual.

- In large samples, the sampling distribution of the IV/TSLS estimator is normal.

- Inference (e.g., hypothesis tests and confidence intervals) proceeds in the usual way.

- The standard errors from the second-stage OLS regression are not valid, because they do not take account of the act that the first-stage is also estimated.

- So it is necessary to use a dedicated regression package that carries out TSLS with correct standard errors and hence t-ratios rather than do two separate OLS regressions manually.

A Brief Discussion on Generalised Method of Moments (GMM)

- Is there any other estimator that can be used instead of TSLS?
- The answer is Yes!
- We can use the Generalised Method of Moments or GMM estimator.
- Its an old estimator, older than OLS but very effective due to its characteristics.
- One advantage of this estimator is that its so general that it can be shaped into another type of estimator with higher asymptotic consistency and allowing simple formulation.
- Therefore, it is said that most of the estimators found in the existing literature are special cases of GMM.
- That is why, it is often referred as the "mother of all traditional estimators".

A Brief Discussion on Generalised Method of Moments (GMM)

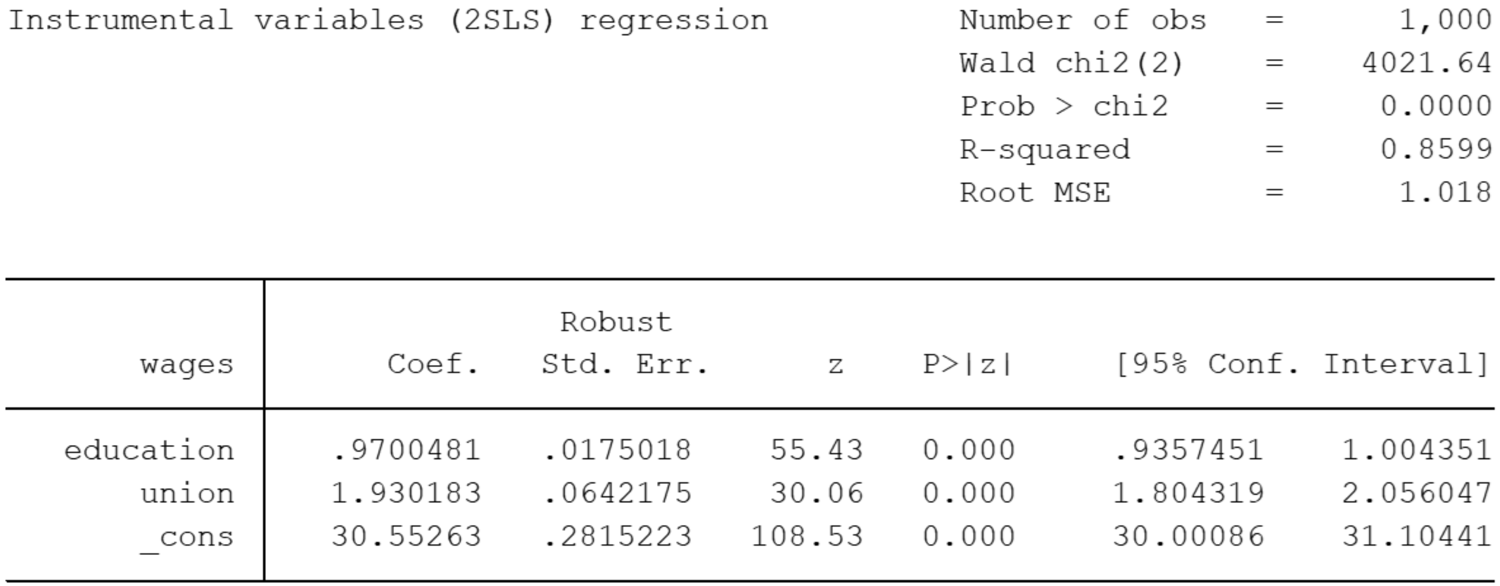
- Instrumental Variable-Generalised Method of Moments (IV-GMM) can be effective for multiple reasons.
- The IV-GMM can deal with unobserved heterogeneity, omitted variable biasness, measurement errors along with endogeneity problems.
- It also provides reasonable inference about the estimated variables as well as performs more efficiently than conventional TSLS under heterogeneous variance condition by using appropriate weights.
- Use of multiple endogenous variables is found to be more simple in IV-GMM framework.
- However, one pitfall is it needs a large sample. You will need a large sample to yield consistent and unbiased estimates.

Example of TSLS

- Let’s take this following model for IV estimation with TSLS:

$$Wage_i = \alpha_0 + \beta_1 Union_i + \beta_2 Edu_i + u_i$$

- Education is endogenous variable. We want to it instrument with father’s and mother’s education levels (our Zs) because literature supports parents education levels tend to shape education of the children.



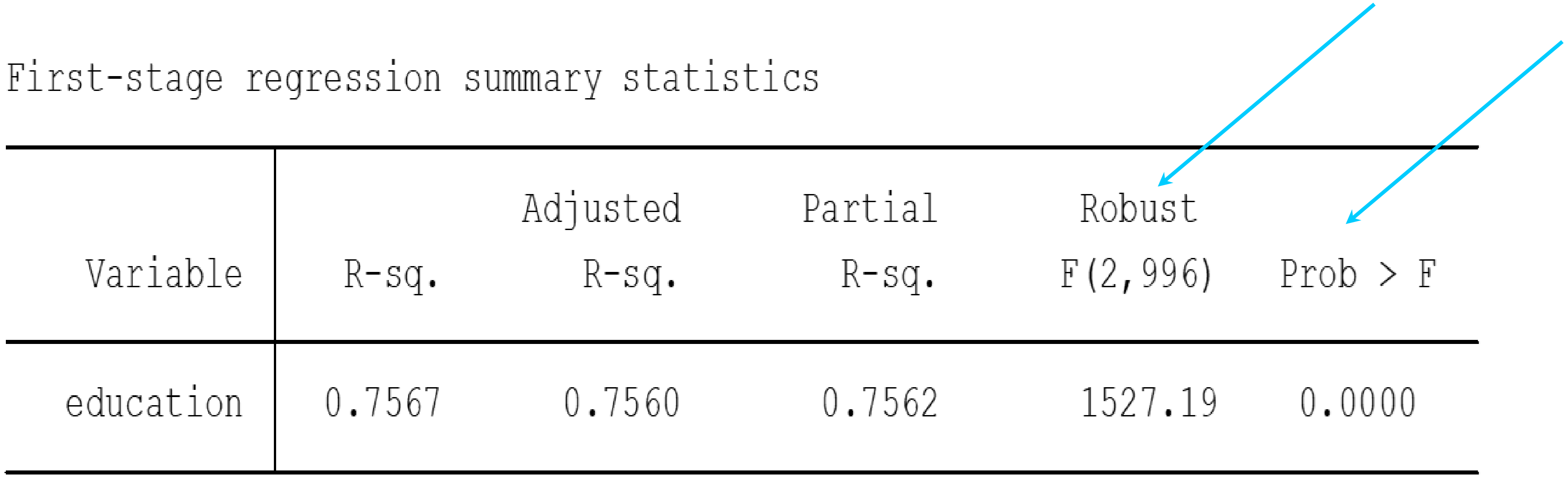
```
Instrumental variables (2SLS) regression          Number of obs     =      1,000
                                                  Wald chi2(2)      =    4021.64
                                                  Prob > chi2       =     0.0000
                                                  R-squared         =     0.8599
                                                  Root MSE          =      1.018

------------------------------------------------------------------------------
             |               Robust
       wages |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   .9700481   .0175018    55.43   0.000     .9357451    1.004351
       union |   1.930183   .0642175    30.06   0.000     1.804319    2.056047
       _cons |   30.55263   .2815223   108.53   0.000     30.00086    31.10441
------------------------------------------------------------------------------
```

Example of TSLS

- Instruments are valid and not weak.

- We can confirm that F-statistics from first-stage is significant. It also meets that greater than 10 rule of thumb when single endogenous regressor is considered in the model.



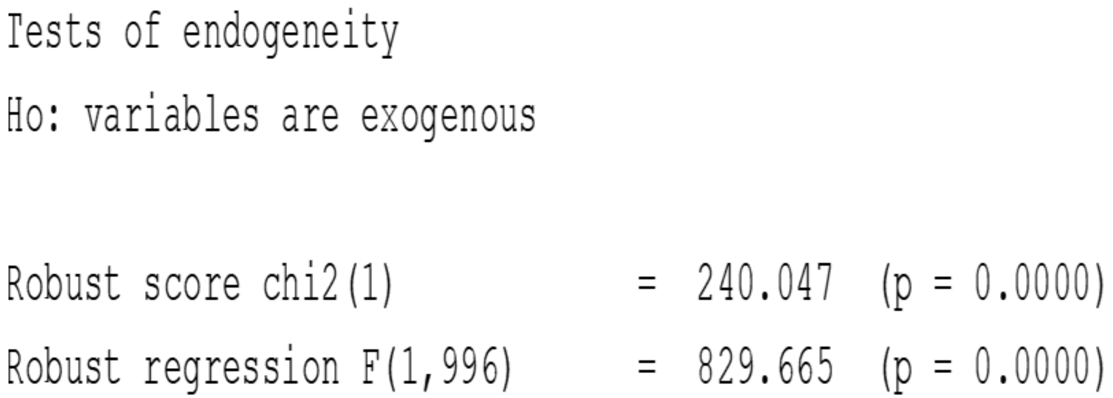
```
First-stage regression summary statistics

                             Adjusted     Partial     Robust
    Variable    R-sq.         R-sq.        R-sq.     F(2,996)   Prob > F

   education   0.7567        0.7560       0.7562     1527.19     0.0000
```

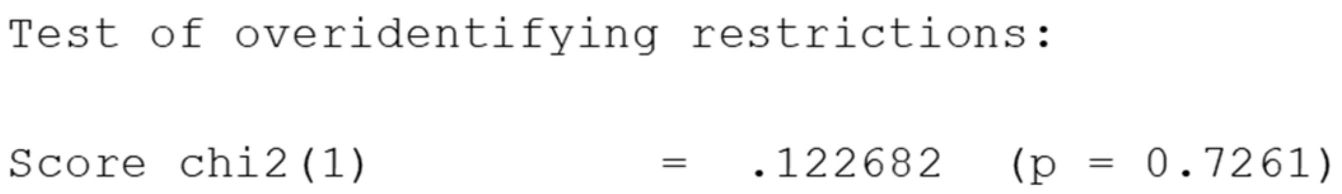
Example of TSLS

- Now let’s check whether education is actually an endogenous variable.
  - See, we cant reject the null, means education is indeed an endogenous variable.
- Similarly, we can check whether there is any issue with overidentification
  - Test can not reject null, means there is no issue with overidentification



```
Tests of endogeneity
Ho: variables are exogenous

Robust score chi2(1)          =  240.047  (p = 0.0000)
Robust regression F(1,996)    =  829.665  (p = 0.0000)
```



```
Test of overidentifying restrictions:

Score chi2(1)           =   .122682   (p = 0.7261)
```

Applications of IV Estimation

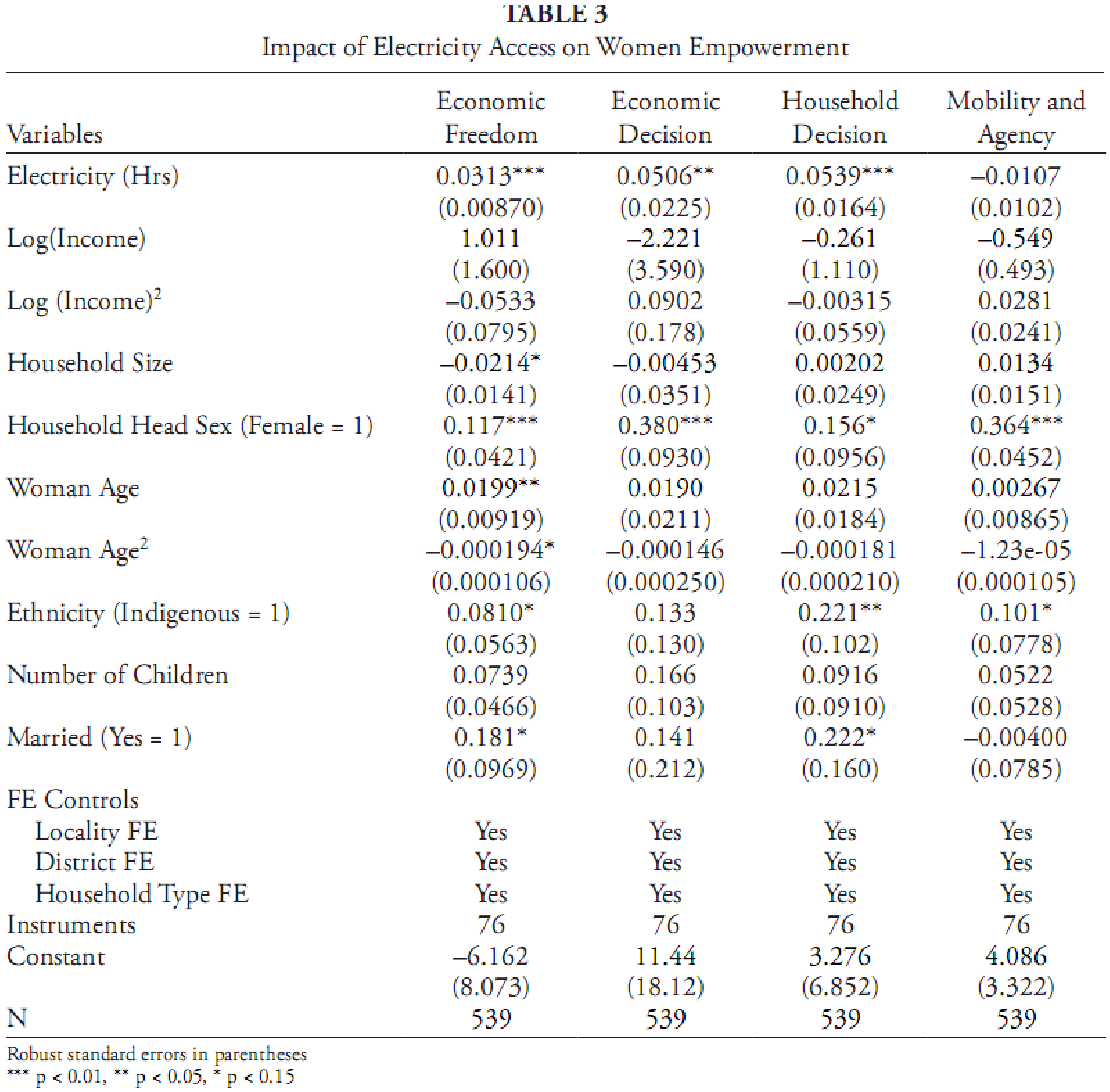


TABLE 3

Impact of Electricity Access on Women Empowerment

Variables	Economic Freedom	Economic Decision	Household Decision	Mobility and Agency
Electricity (Hrs)	0.0313***	0.0506**	0.0539***	–0.0107
	(0.00870)	(0.0225)	(0.0164)	(0.0102)
Log(Income)	1.011	–2.221	–0.261	–0.549
	(1.600)	(3.590)	(1.110)	(0.493)
Log (Income)^2	–0.0533	0.0902	–0.00315	0.0281
	(0.0795)	(0.178)	(0.0559)	(0.0241)
Household Size	–0.0214*	–0.00453	0.00202	0.0134
	(0.0141)	(0.0351)	(0.0249)	(0.0151)
Household Head Sex (Female = 1)	0.117***	0.380***	0.156*	0.364***
	(0.0421)	(0.0930)	(0.0956)	(0.0452)
Woman Age	0.0199**	0.0190	0.0215	0.00267
	(0.00919)	(0.0211)	(0.0184)	(0.00865)
Woman Age^2	–0.000194*	–0.000146	–0.000181	–1.23e-05
	(0.000106)	(0.000250)	(0.000210)	(0.000105)
Ethnicity (Indigenous = 1)	0.0810*	0.133	0.221**	0.101*
	(0.0563)	(0.130)	(0.102)	(0.0778)
Number of Children	0.0739	0.166	0.0916	0.0522
	(0.0466)	(0.103)	(0.0910)	(0.0528)
Married (Yes = 1)	0.181*	0.141	0.222*	–0.00400
	(0.0969)	(0.212)	(0.160)	(0.0785)
FE Controls				
Locality FE	Yes	Yes	Yes	Yes
District FE	Yes	Yes	Yes	Yes
Household Type FE	Yes	Yes	Yes	Yes
Instruments	76	76	76	76
Constant	–6.162	11.44	3.276	4.086
	(8.073)	(18.12)	(6.852)	(3.322)
N	539	539	539	539

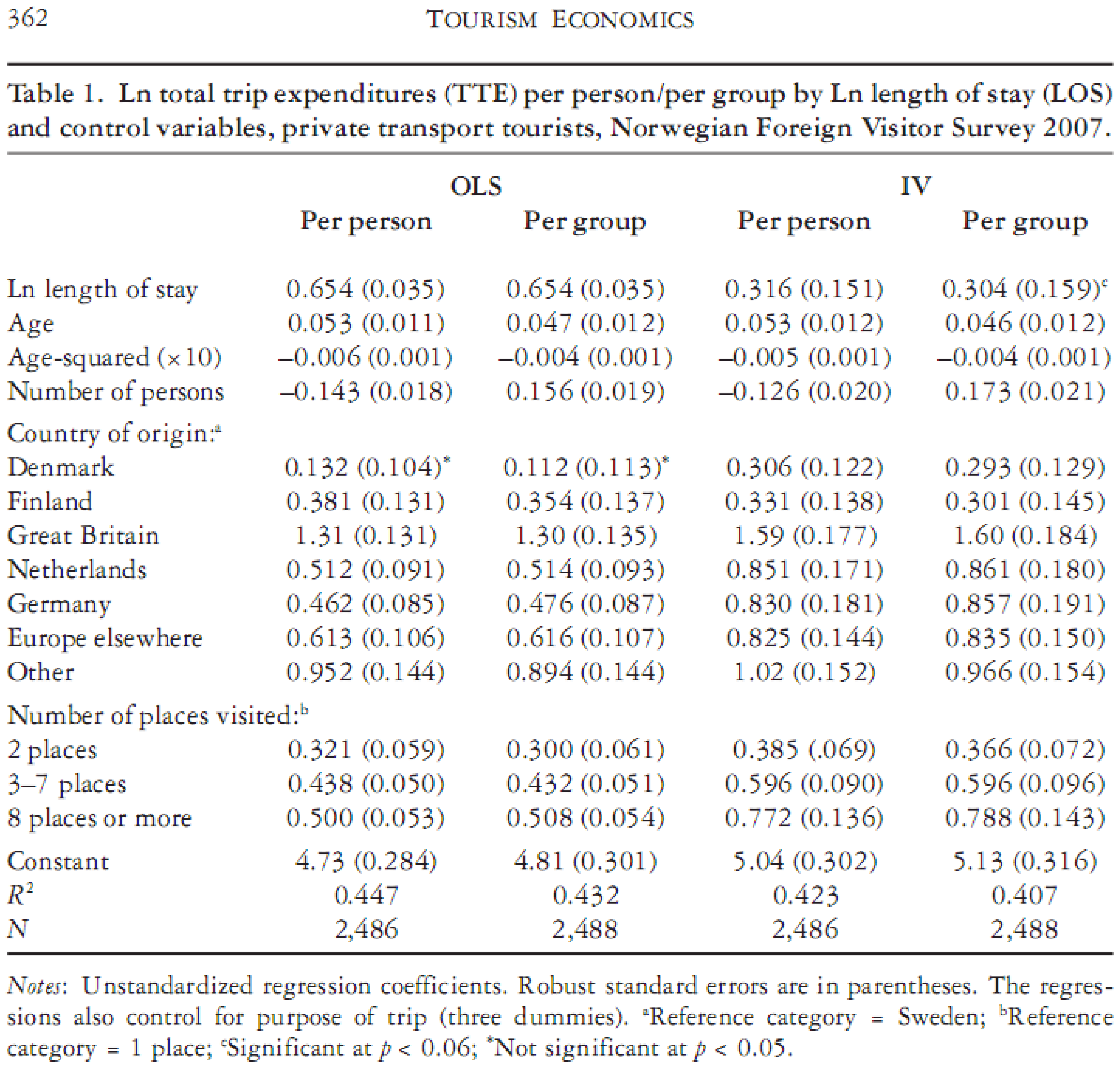
Robust standard errors in parentheses

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.15$

- ➢In these models electricity access (in hours) is endogenous.
- ➢Instruments are geographic terrain, area-specific household behaviour, certain norms and culture, income opportunities, local education facilities, tourism market, inter-district migration.
- ➢The number of valid instruments is 76 in these models.

Amin, S. B., Jamasb, T., Khan, F., & Nepal, R. (2024). Electricity access, gender disparity, and renewable energy adoption dynamics: The case of mountain areas of Bangladesh. *Economics of Energy & Environmental Policy*, 13 (1), DOI: 10.5547/2160-5890.13.1.sami

Application of IV Estimations



362 TOURISM ECONOMICS

Table 1. Ln total trip expenditures (TTE) per person/per group by Ln length of stay (LOS) and control variables, private transport tourists, Norwegian Foreign Visitor Survey 2007.

	OLS		IV	
	Per person	Per group	Per person	Per group
Ln length of stay	0.654 (0.035)	0.654 (0.035)	0.316 (0.151)	0.304 (0.159)[c]
Age	0.053 (0.011)	0.047 (0.012)	0.053 (0.012)	0.046 (0.012)
Age-squared (×10)	–0.006 (0.001)	–0.004 (0.001)	–0.005 (0.001)	–0.004 (0.001)
Number of persons	–0.143 (0.018)	0.156 (0.019)	–0.126 (0.020)	0.173 (0.021)
Country of origin:[a]				
Denmark	0.132 (0.104)*	0.112 (0.113)*	0.306 (0.122)	0.293 (0.129)
Finland	0.381 (0.131)	0.354 (0.137)	0.331 (0.138)	0.301 (0.145)
Great Britain	1.31 (0.131)	1.30 (0.135)	1.59 (0.177)	1.60 (0.184)
Netherlands	0.512 (0.091)	0.514 (0.093)	0.851 (0.171)	0.861 (0.180)
Germany	0.462 (0.085)	0.476 (0.087)	0.830 (0.181)	0.857 (0.191)
Europe elsewhere	0.613 (0.106)	0.616 (0.107)	0.825 (0.144)	0.835 (0.150)
Other	0.952 (0.144)	0.894 (0.144)	1.02 (0.152)	0.966 (0.154)
Number of places visited:[b]				
2 places	0.321 (0.059)	0.300 (0.061)	0.385 (.069)	0.366 (0.072)
3–7 places	0.438 (0.050)	0.432 (0.051)	0.596 (0.090)	0.596 (0.096)
8 places or more	0.500 (0.053)	0.508 (0.054)	0.772 (0.136)	0.788 (0.143)
Constant	4.73 (0.284)	4.81 (0.301)	5.04 (0.302)	5.13 (0.316)
R^2	0.447	0.432	0.423	0.407
N	2,486	2,488	2,486	2,488

Notes: Unstandardized regression coefficients. Robust standard errors are in parentheses. The regressions also control for purpose of trip (three dummies). [a]Reference category = Sweden; [b]Reference category = 1 place; [c]Significant at $p < 0.06$; *Not significant at $p < 0.05$.

Thrane, C. (2015). On the relationship between length of stay and total trip expenditures: a case study of instrumental variable (IV) regression analysis. *Tourism Economics*, 21(2), 357-367.

STATA Codes for IV Estimations

```
*TSLS
webuse educwages
ivregress 2sls wages union (education= meducation feducation), vce(r)
        estat firststage
        estat endogenous
        estat overid

*IV-GMM
ivregress gmm wages union (education= meducation feducation), vce(r)
        estat firststage
        estat endogenous
        estat overid

**note that there are many packages which you can use to run IV estimations in STATA other than “ivregress”.
For instance “ivreg2”**
```