Instrumental Variables

Martin J. Conyon

October 2015

Lancaster University

Preamble

- These notes are based (mainly) on Jeffrey Wooldridge's text "Introductory Econometrics: A Modern Approach".
- See Chapter 15, Instrumental Variables and Two Stage Least Squares.

Motivation - omitted variables

- Suppose we have an omitted variable bias problem – unobserved heterogeneity – How can we address this issue statistically?

- We could a) ignore it, b) find a proxy for the missing variable, c) assume effects are constant over time & then eliminate it by taking first differences in a panel of data. Another approach is to use Instrumental Variables (IV)

- A standard example from labor economics is how to estimate the causal effect of schooling on future earnings. The big economic question is 'what are the returns to eduction'?

Motivation - omitted variables

- The problem is that there are missing variables from the earnings equation. The 'true' model differs from the 'estimated' model.

- Consider the 'true' model:
  $\log(wage) = \beta_o + \beta_1 educ + \beta_2 abil + u$

- The term β_1 measure the marginal effect of education and β_2 the effect of ability.

- If there is another variable called 'IQ' that is observable and a perfect proxy for ability (*abil*), then we can use OLS and regress log(wage) on education and IQ. If not, we use IV. Ignoring relevant explanatory variable and using OLS leads to a bias.

Omitted variable bias

- In general, OLS leads to biased parameter estimates if relevant variables are missing & correlated with included regressors. Suppose one estimates erroneously the linear model:

$$y_i = X_i\beta + \epsilon_i, \qquad i = 1, \dots, \tag{1}$$

but the underlying true model is:

$$y_i = X_i\beta + Z_i\delta + \epsilon_i, \qquad i = 1, \dots, \tag{2}$$

The OLS estimator, ignoring the matrix Z is:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3}$$

In general this estimator is biased and inconsistent.

Omitted variable bias

- Substitute the 'true' model $Y_i = X_i\beta + Z_i\delta + \epsilon_i$ in $\hat{\beta}$ surpress i for convenience.

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'(X\beta + Z\delta + \epsilon)\\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'Z\delta + (X'X)^{-1}X'\epsilon\\ &= \beta + (X'X)^{-1}X'Z\delta + (X'X)^{-1}X'\epsilon\end{aligned}$$

Taking expectations and noting $(X'X)^{-1}X'E(\epsilon) = 0$

$$\begin{aligned}\hat{\beta} &= \beta + (X'X)^{-1}X'Z\delta + (X'X)^{-1}X'E(\epsilon)\\ &= \beta + (X'X)^{-1}X'Z\delta\\ &= \beta + bias\\ &= \beta + Cov(X,Z) \times \delta\end{aligned}$$

Instrumental variables

- Back to the original example, suppose there is no proxy variable for the ability term and we estimate:

$$\log(wage) = \beta_0 + \beta_1 educ + u \tag{4}$$

- Now the *abil* term is in the error u. If eqn 4. If estimated by OLS then β_1 is biased and inconsistent if we think that a) ability is correlated with earnings b) ability is correlated with education.

- We can still use eqn 4 if we can find an instrumental variable for *educ*. Consider the estimating equation:

$$\log(wage) = \beta_0 + \beta_1 x + u \tag{5}$$

Instrument requirements

- Suppose that the terms x and u are correlated (e.g. via ability):

$$Cov(x, u) \neq 0 \tag{6}$$

- To obtain consistent estimates of β_0 and β_1 we need more information. Suppose there is an observable variable z that satisfies two assumptions:

$$Cov(z, u) = 0 \tag{7}$$

$$Cov(z, x) \neq 0 \tag{8}$$

- This says that z is uncorrelated with the error term u, and z is correlated with variable of interest x. Then we say that z is an **instrumental variable** or **instrument** for x

Instrument requirements

- **Instrument exogeneity**: Equation (7) requires $Cov(z, u) = 0$
- This means that z has no effect on y (controlling for x) and z is uncorrelated with the omitted variables.
- This assumption cannot be tested since u_i is not observable – We have to assume $Cov(z, u) = 0$ and in writing economics justify it.

- **Instrument relevance**: Equation (8) requires $Cov(z, x) \neq 0$. The instrument z is correlated with x.
- It says the variable z is relevant for explaining variation in x
- The assumption $Cov(z, x) \neq 0$ *can* be tested

Instrument requirements

- We can test the validity of the **instrument relevance** assumption by performing the regression:

$$x = \pi_o + \pi_1 z + v \tag{9}$$

- Because $\pi_1 = Cov(z, x)/Var(z)$ then the assumption that $Cov(z, x) \neq 0$ holds iff $\pi_1 \neq 0$. We simply test

$$H_0 : \pi_1 = 0 \tag{10}$$

- So, in the example wage equation (4) we require the instrument to a) be uncorrelated with ability and other omitted factors in u, and b) correlated with education. (Q: is the last digit of your SSN a good instrument?)

Identification

- We now want to show that the availability of an instrument, z, can be used to consistently estimate Eqn 5: $\log(wage) = \beta_0 + \beta_1 x + u$
- The conditions in Eqn (7) $[Cov(z, u) = 0]$ and Eqn (8) $[Cov(z, x) \neq 0]$ serve to identify the population parameter β_1
- Identification in this context means writing β_1 in terms of population moments that can be estimated using sample data.

Identification

- Using Eqn (5) the covariance between the instrument z and outcome y can be written as:

$$Cov(z, y) = \beta_1 Cov(z, x) + Cov(z, u) \tag{11}$$

- Since $Cov(z, u) = 0$ (exogeneity) and $Cov(z, x) \neq 0$ (relevance) then β_1:

$$\beta_1 = \frac{Cov(z, y)}{Cov(z, x)} \tag{12}$$

Identification

- Given a sample of data, the population parameter $\beta_1 = \frac{Cov(z,y)}{Cov(z,x)}$ is estimated from the sample analog.

- The **Instrumental Variable (IV) estimator** of β_1 can be written as:

$$\hat{\beta}_{IV} = \hat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\displaystyle\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \tag{13}$$

- The IV estimator of β_0 is $\bar{y} - \hat{\beta}_1\bar{x}$. When $z = x$ the OLS estimator of β_1 is equal to the IV estimator

Statistical inference with the IV estimator

- In large samples the IV estimator has an approximate normal distribution. Assuming homoskedasticity, we further assume that

$$E(u^2/z) = \sigma^2 = Var(u) \tag{14}$$

- Under assumptions of Eqns (7), (8), (14) the asymptotic variance of $\hat{\beta}_1$ is given as:

$$\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2} \tag{15}$$

- where σ^2 is the population variance of u, σ_x^2 is the variance of x, and $\rho_{x,z}^2$ is the square of the population correlation between x and z. The asymptotic variance decreases at rate $\frac{1}{n}$.

Statistical inference with the IV estimator

- We can get an estimate of σ^2 using the IV residuals $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, i = 1...n$ where $\hat{\beta}_0$ and $\hat{\beta}_1$ are IV estimates. Then

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2 \tag{16}$$

- The asymptotic standard error of $\hat{\beta}_1$ is given as the square root of the asymptotic variance. The variance is:

$$\frac{\hat{\sigma}^2}{SST_x \times R^2_{x,z}} \tag{17}$$

where SST_x is the total sum of squares of x and $R^2_{x,z}$ is the R-squared. The estimated SEs can be used to construct t-statistics or confidence intervals.

Statistical inference with the IV estimator

- Equations (15) and (17) allow a comparison of asymptotic variances of IV and OLS.

- If we assume the Gauss-Markov axioms hold then the variance of OLS estimator is σ^2/SST_x, while for the IV estimator it is $\sigma^2/(SST_x \times R^2_{x,z})$

- Because $R^2_{x,z} < 1$ the IV variance is always larger than OLS variance, when OLS is valid. When $R^2_{x,z}$ is small this can lead to large IV sampling variance.

Statistical inference with the IV estimator

- There is a cost to using IV then which depends on the correlation of the instrument z with x. The higher the correlation of z and x then $R^2_{x,z}$ is closer to 1.

- Better instruments lead to lower (better) estimates of the standard errors for statistical inference.

- When x and u are uncorrelated, IV estimation comes at a large cost – the IV asymptotic variance is always larger. This is part of a problem called **weak instruments** to which we return later.

Examples

- Lets look at some real examples (see explanation in Wooldridge)
- First, estimate the returns to education for married women. Use the mroz.dta stata data file.
- For OLS estimate regress log wage on education (educ) and print the results. Next for the IV model first regress 'educ' on fathers education (fatheduc) in the first state; IV second-stage regress log wage on education using IV.

Example: returns to education for married women

```
. reg lwage edu
```

Source	SS	df	MS
Model	26.3264193	1	26.3264193
Residual	197.001022	426	.462443713
Total	223.327441	427	.523015084

Number of obs =	428
F(1, 426) =	56.93
Prob > F =	0.0000
R-squared =	0.1179
Adj R-squared =	0.1158
Root MSE =	.68003

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1086487	.0143998	7.55	0.000	.0803451	.1369523
_cons	-.1851968	.1852259	-1.00	0.318	-.5492673	.1788736

Estimated in Stata 12.0 using data set: mroz.dta

Example: returns to education for married women

```
. reg educ fatheduc if lwage~=.

      Source |       SS       df       MS              Number of obs =     428
-------------+------------------------------           F(  1,   426) =   88.84
       Model |  384.841983     1  384.841983           Prob > F      =  0.0000
    Residual |  1845.35428   426  4.33181756           R-squared     =  0.1726
-------------+------------------------------           Adj R-squared =  0.1706
       Total |  2230.19626   427  5.22294206           Root MSE      =  2.0813

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    fatheduc |   .2694416   .0285863     9.43   0.000     .2132538    .3256295
       _cons |   10.23705   .2759363    37.10   0.000     9.694685    10.77942
------------------------------------------------------------------------------
```

Estimated in Stata 12.0 using data set: mroz.dta

Example: returns to education for married women

```
. ivregress 2sls lwage (educ=fatheduc)

Instrumental variables (2SLS) regression          Number of obs =      428
                                                  Wald chi2(1)  =     2.85
                                                  Prob > chi2   =   0.0914
                                                  R-squared     =   0.0934
                                                  Root MSE      =   .68778
```

lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.0591735	.0350596	1.69	0.091	-.009542	.127889
_cons	.4411034	.4450583	0.99	0.322	-.4311947	1.313402

Estimated in Stata 12.0 using data set: mroz.dta

Example: returns to education for married women

- OLS leads to 10.9% estimate of the returns to education; IV leaders to 5.9% return – half the OLS estimate in this sample & with these instruments.
- Estimates are for just one sample. Not clear which is 'truer' estimate.
- The standard error on the IV returns estimate is 1.5 times the size of the OLS estimate. IV confidence interval contains the OLS estimate.

Example: returns to education for men

- Estimate returns to education for men.
- Use the number of siblings as an instrumental variable.
- Notice that the returns to education are higher in the IV model.

Example: returns to education for men

```
. reg lwage educ // OLS of log wage on education

      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  1,   933) =  100.70
       Model |  16.1377042     1  16.1377042           Prob > F      =  0.0000
    Residual |  149.518579   933  .160255712           R-squared     =  0.0974
-------------+------------------------------           Adj R-squared =  0.0964
       Total |  165.656283   934  .177362188           Root MSE      =  .40032

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0598392   .0059631    10.03   0.000     .0481366    .0715418
       _cons |   5.973063   .0813737    73.40   0.000     5.813366    6.132759
------------------------------------------------------------------------------
```

Estimated in Stata 12.0 using data set: wage2.dta

Example: returns to education for men

. reg educ sibs // regress education on number of siblings

Source	SS	df	MS
Model	258.055048	1	258.055048
Residual	4248.7642	933	4.55387374
Total	4506.81925	934	4.82528828

Number of obs =	935
F(1, 933) =	56.67
Prob > F =	0.0000
R-squared =	0.0573
Adj R-squared =	0.0562
Root MSE =	2.134

educ	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sibs	-.2279164	.0302768	-7.53	0.000	-.287335	-.1684979
_cons	14.13879	.1131382	124.97	0.000	13.91676	14.36083

Estimated in Stata 12.0 using data set: wage2.dta

Example: returns to education for men

```
. ivregress 2sls lwage (educ=sibs) // IV regression, treating education as endogenous

Instrumental variables (2SLS) regression               Number of obs =      935
                                                       Wald chi2(1)  =    21.63
                                                       Prob > chi2   =   0.0000
                                                       R-squared     =        .
                                                       Root MSE      =   .42284

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1224326   .0263224     4.65   0.000     .0708417    .1740235
       _cons |   5.130026   .3547911    14.46   0.000     4.434648    5.825404
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   sibs
```

Estimated in Stata 12.0 using data set: wage2.dta

Weak Instruments

- IV estimation leads to **consistent** estimates (under assumptions $Cov(z, u) = 0$ and $Cov(z, x) \neq 0$).
- IV estimates have **larger** standard errors if z and x are only weakly correlated.
- The problem is more serious. IV estimators can have a **large asymptotic bias** even if z and x are **moderately** correlated. To see this we can write the probability limit of the IV estimator:

$$\text{plim}\hat{\beta}_{1,IV} = \beta_1 + \frac{Corr(z, u)}{Corr(z, x)} \times \frac{\sigma_u}{\sigma_x} \tag{18}$$

- Where σ_u and σ_x are standard deviations of u and x in the population.

Weak instruments

1. If both $Corr(z, u)$ and $Corr(z, x)$ are **small** then the IV estimator is not consistent. The weak instrument $Corr(z, x)$ is swamped by equally small correlation between the equation error and the instrument $Corr(z, x)$.

2. Even if we want to focus on consistency it may not be better to use IV if the instruments are weak.

3. When x and z are hardly correlated, or not correlated at all, things are especially bad whether or not z is uncorrelated with u.

Example: smoking and birthweight

- Consider estimating the effect of smoking on child birth-weight. The simple linear regression model is:

$$log(bweight) = \beta_0 + \beta_1 packs + u \tag{19}$$

- Where *bweight* is birth-weight and *packs* is the number of packs of cigarettes smoked by mother per days and z. We are concerned that u and *packs* are correlated, because *packs* is correlated with other health factors.

- A potential instrument for *packs* is the cigarette price, *cigprice* (we assume *cigprice* and u are uncorrelated)

Example: smoking and birthweight

- For *cigprice* to be relevant instrument it must be correlated to *packs*. If cigarettes were a normal good then *cigprice* and *packs* are negatively correlated (why?).
- The following regression output shows that *cigprice* and *packs* are not correlated (why?).
- Because of this we should not use it as a regression variable. What happens if we do? We get a huge coefficient estimate in the birthweight equation; a large standard error; and the wrong sign!

Example: smoking and birthweight

```
. reg packs cigprice

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  1,  1386) =    0.13
       Model |  .011648626     1  .011648626           Prob > F      =  0.7179
    Residual |  123.684481  1386  .089238442           R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.0006
       Total |  123.696129  1387  .089182501           Root MSE      =  .29873

------------------------------------------------------------------------------
       packs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cigprice |   .0002829    .000783     0.36   0.718    -.0012531    .0018188
       _cons |   .0674257   .1025384     0.66   0.511    -.1337215    .2685728
------------------------------------------------------------------------------
```

Estimated in Stata 12.0 using data set: bwght.dta

Example: smoking and birthweight

```
. ivregress 2sls lbwght (packs=cigprice),

Instrumental variables (2SLS) regression          Number of obs =    1388
                                                  Wald chi2(1)  =    0.12
                                                  Prob > chi2   =  0.7310
                                                  R-squared     =       .
                                                  Root MSE      =  .93818
```

lbwght	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
packs	2.988676	8.692619	0.34	0.731	-14.04854	20.0259
_cons	4.448136	.9075006	4.90	0.000	2.669468	6.226805

Estimated in Stata 12.0 using data set: bwght.dta

IV Estimation: Multiple Regression

- The IV Estimator for the simple 2-variable linear regression is easily extended to the multiple regression case.
- Consider the **structural equation**:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_1 z_1 + u_1 \tag{20}$$

- *Explanatory variables:* y_2 *and* z_1*;* Endogenous variables: y_1 and y_2 ; and *Exogenous variable*: z_1
- Estimating Eqn (20) by OLS leads to biased and inconsistent estimates of all variables.

IV estimation: multiple regression

- To estimate by IV we need to find another **exogenous** variable, call it z_2.
- We make the following exogeneity assumptions for IV estimation

$$E(u_1) = 0, Cov(z_1, u_1) = 0, Cov(z_2, u_1) = 0 \qquad (21)$$

IV estimation: multiple regression

- Use methods of moments to derive IV estimators $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ by solving the normal equations:

$$\begin{aligned}
\sum_{i=1}^{n}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) &= 0 \\
\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) &= 0 \\
\sum_{i=1}^{n} z_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) &= 0
\end{aligned} \tag{22}$$

IV estimation: multiple regression

- Like before the instrument z_2 needs to be correlated with y_2. Write the endogenous variables as a **reduced form** function of the exogenous variables:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2 \tag{23}$$

- Where $E(v_2) = 0$, $Cov(z_1, v_2) = 0$, $Cov(z_2, v_2) = 0$ and π_j are unknown parameters.
- The key identification condition is that:

$$\pi_2 \neq 0 \tag{24}$$

- Meaning that after controlling for z_1 the variable y_2 is still correlated with z_2. Without loss of generality, this result is extended to many exogenous variables (Z_i).

IV estimation: multiple regression

- Write the **structural equation** as:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 ... + \beta_k z_{k-1} + u_1 \tag{25}$$

- Suppose y_2 is correlated with u_1. Let there be a z_k not in Eqn (25) that is exogenous, so we assume

$$E(u_1) = 0, Cov(z_j, u_1) = 0, j = 1, ..., k. \tag{26}$$

- Eqn (26) implies $z_1 ... z_k$ are exogenous. The reduced form for the endogenous y_2 is:

$$y_2 = \pi_0 + \pi_1 z_1 + ... + \pi_{k-1} z_{k-1} + \pi_k z_k + v_2 \tag{27}$$

- And for identification we require (at least some):

$$\pi_k \neq 0 \tag{28}$$

Two Stage Least Squares

- Up to now we have assumed that there is one endogenous variable y_2 and one instrumental variable.
- Often it is the case that we have more than one instrument for the endogenous variable.
- Multiple instrumental variables are easily incorporated into this framework. We can also add tests of a) edogeneity and b) Overidentifying restrictions.

Two Stage Least Squares

- Consider the case of a single endogenous variable. Suppose there are 2 exogenous variables excluded from Eqn (20), z_2 & z_3, and one endogenous variable, y_2

- The assumption that z_2 & z_3 do **not** appear in Eqn (20) and are uncorrelated with u_1 are called the **exclusion restrictions**.

- We could use either z_1 or z_2 as an instrument, but neither estimator is likely to be efficient. Instead, to find the optimal IV write the reduced form equation:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2 \tag{29}$$

Two Stage Least Squares

- Where
  $E(v_2) = 0, Cov(z_1, v_2) = 0, Cov(z_2, v_2) = 0, Cov(z_3, v_3) = 0$
  and π_j are unknown parameters.

- We require $\pi_1 \neq 0$ or $\pi_2 \neq 0$ for the IV not to be perfectly correlated with z_1

- Form the instrument

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3 \tag{30}$$

- With multiple instruments the IV estimator is also called the **Two Stage Least Squares (2SLS) estimator.**

Two Stage Least Squares

- Adding more exogenous variables z_3, z_4 changes things very little and the model is easily generalized.
- Most econometrics packages (Stata, for example) contain 2SLS estimators (in the case of Stata it is the ivregress command)
- This means you don't have to perform the two stage regressions yourself. In fact, you should not as the standard errors are incorrect.
- Adding more endogenous variables is also easily accommodated. With two endogenous variables you need at least two exogenous variables.

Endogeneity Tests

- 2SLS estimator is less efficient than OLS when explanatory variables are exogenous. Therefore important to test for endogeneity to see if 2SLS is really necessary.

- Hausman (1978) suggests directly comparing the OLS and 2SLS estimates and determining whether the estimates are statistically significant. OLS and 2SLS are both consistent if all the variables are exogenous. If they turn out to be different we can conclude that the variable y_2 is, in fact, endogenous.

Endogeneity Tests

- Estimate reduced form for y_2 for all exogenous variables and retrieve residuals, $\hat{v}_2$

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2 \tag{31}$$

- Augment the structural equation with $\hat{v}_2$

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \beta_4 \hat{v}_2 + error \tag{32}$$

- Test $H_0 : \beta_4 = 0$. If significantly different, conclude endogeniety.

Overidentification Tests

- The model with a single endogenous variable is said to be overidentified when the number of instruments is greater than one, $M > 1$, and there are $M - 1$ overidentifying restrictions.
- This means that if each z_i has some correlation with the endogenous variable x_i, then we have $M - 1$ more exogenous variables than needed to identify the parameters .
- In this case we can test whether the additional instruments are valid in the sense that they are uncorrelated with u_i.

Overidentification Tests

- Hausman (1978) suggested comparing the 2SLS estimator using all instruments to 2SLS using a subset that just identifies equation. If all instruments are valid, the estimates should differ only as a result of sampling error.

- Perform the following simple regression based procedure, under homoskedasticity,
    1. Run 2SLS using all instruments
    2. Obtain the residuals $\widehat{u}$
    3. Run OLS $\widehat{u}$ on all exogenous variables
    4. Construct the test statistic NR_u^2, from the OLS regression.
- Under the Null $H_o : E(Z'u) = 0; NR_u^2 \sim \chi^2_{M-1}$

Heteroskedasticity Tests

- For both OLS and 2SLS heteroskedasticity does not affect the consistency of the estimators.

- However, we can test $H_o : E(u^2|X) = \sigma^2$ against the alternative that $E(u^2|X)$ depends on X in some way.

- Use Breusch and Pagan (1979) or White (1980) to test and then correct if necessary.

- Finally, one can perform a Ramset RESET test for functional form non-linearities if desired.

2SLS

- For both OLS and 2SLS heteroskedasticity does not affect the consistency of the estimators.

- However, we can test $H_o : E(u^2|X) = \sigma^2$ against the alternative that $E(u^2|X)$ depends on X in some way.

- Use Breusch and Pagan (1979) or White (1980) to test and then correct if necessary.

- Finally, one can perform a Ramset RESET test for functional form non-linearities if desired.