

Huber Correction for Two-Stage Least Squares Estimates: HREG2SLS

Mead Over, The World Bank, EMAIL aover@worldbank.org

Dean Jolliffe, The World Bank, EMAIL djolliffe@worldbank.org

Andrew Foster, University of Pennsylvania, EMAIL afoster@pop.upenn.edu

In applied microeconomic analysis, instrumental variable estimation by two-stage least squares is frequently used to estimate structural parameters when explanatory variables are endogenous. In Stata's **regress** command, instrumental variable estimation is neatly implemented by placing the list of exogenous variables in parentheses, after the list of independent variables and before the comma that demarcates the options. Because the Stata command for least squares with Huber- (or White-) corrected standard errors residuals, **hreg**, has almost the same syntax as the **regress** command, it is natural to infer that it would also accept a list of instrumental variables in parentheses before the comma as a signal to perform instrumental variable estimation before correcting the standard errors. In fact, **hreg** does not correctly interpret the variables in parentheses. (It simply ignores the parentheses and treats the list of supposed instrumental variables as if they were additional members of the list of independent variables.)

Hreg2sls is an altered version of **hreg** which does recognize the set of variables in parentheses as a set of instrumental variables. It is identical to **hreg** in all respects except that it allows instrumental variable estimation.

Syntax

```
hreg2sls [depvar [varlist1 [(varlist2)]] [weight] [if exp] [in exp] [in range]]  
    [, group(varname) level(#) regress_options ]
```

Examples

A typical use of **hreg2sls** will follow this pattern:

```
hreg2sls y1 y2 x1 x2 x3 (x1 x2 x3 z1 z2), group(cluster)
```

where y1 and y2 are endogenous variables, x1-x3 are exogenous variables, z1-z2 are the excluded instruments, and cluster is a variable designating the first stage of a two-stage sample design (i.e, town or city in a household survey). If residuals in the same region are correlated or residual variances differ systematically by region then a 2SLS procedure such as **reg** that assumes homoscedasticity and independence will in general produce inconsistent standard errors.

For another example of **hreg2sls** consider a slightly modified version of the model used in the Stata manual to describe two-stage least squares estimation:

$$\mathbf{hsngval} = \alpha_0 + \alpha_1 \mathbf{faminc} + \alpha_2 \mathbf{pcturban} + \varepsilon$$

$$\mathbf{rent} = \beta_0 + \beta_1 \mathbf{hsngval} + \beta_2 \mathbf{pcturban} + \upsilon$$

Hsngval is median value of housing in each state, **rent** is the state-level, median monthly rent, **faminc** is the median value of family income, and **pcturban** is the percentage of the state population living in urban areas. The data are found in the **hsng.dta** file distributed with Stata. The only difference between this example and the one used in the Stata manual is that the region dummy variables have been dropped from the **hsngval** equation. For this example it is assumed that the inter-region variation of the residuals is different from the intra-region variation, which results in a heteroscedastic error structure. Use of the group option in **hreg2sls** will correct the estimated standard errors for this form of heteroscedasticity. Below are the two-stage least squares estimates of this model, and then following are the two-stage least squares estimates with Huber-corrected standard errors.

. reg rent hsngval pcturban (pcturban faminc)

(2SLS)						
Source	SS	df	MS	Number of obs = 50		
Model	17681.4852	2	8840.74262	F(2, 47)	= 24.18	
Residual	43561.6348	47	926.843293	Prob > F	= 0.0000	
Total	61243.12	49	1249.85959	R-squared	= 0.2887	
				Adj R-squared	= 0.2584	
				Root MSE	= 30.444	
rent	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hsngval	.0031938	.0006401	4.990	0.000	.0019062	.0044815
pcturban	-.5064118	.4966869	-1.020	0.313	-1.505617	.4927933
_cons	113.8143	21.17164	5.376	0.000	71.22248	156.4062

. hreg2sls rent hsngval pcturban (pcturban faminc), group(region)

Regression with Huber standard errors (2SLS)

Number of obs = 50
R-square = 0.2887
Adj R-square = 0.2584
Root MSE = 30.4441

Grouping variable: region

rent	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
hsngval	.0031938	.0004785	6.674	0.000	.0022311	.0041565
pcturban	-.5064118	.7125682	-0.711	0.481	-1.939914	.9270905
_cons	113.8143	21.43369	5.310	0.000	70.6953	156.9334

Methods and Formulas

The Huber variance covariance matrix for ordinary least squares estimates of β in the linear expression $y_i = \beta'x_i + u_i$ is:

$$\text{var}(\beta) = \left(\sum_i \frac{x_i x_i'}{n} \right)^{-1} \left(\sum_j \frac{\hat{u}_j^2 x_j x_j'}{n} \right) \left(\sum_i \frac{x_i x_i'}{n} \right)^{-1} \quad (1)$$

where \hat{u}_j is the estimated residual for observation j . The formula corresponding to equation (1) for two-stage least squares estimation (see, for example, White 1984 p. 141) is obtained by replacing the vector x_i in (1) with its predicted value from the first stage regressions, \hat{x}_i :

$$\text{var}(\beta) = \left(\sum_i \frac{\hat{x}_i \hat{x}_i'}{n} \right)^{-1} \left(\sum_j \frac{\hat{u}_j^2 \hat{x}_j \hat{x}_j'}{n} \right) \left(\sum_i \frac{\hat{x}_i \hat{x}_i'}{n} \right)^{-1} \quad (2)$$

Extension of these formula to the case of clustered data is straightforward as illustrated (for equation 1) in the Stata manual in [5s] under Huber.

Hreg2sls takes advantage of the similarity between equations (2) and (1) by replacing each of the x variables in the data set by its respective predicted value and then calling Stata's Huber engine, **_huber**. The **preserve** command is used to ensure that the x variables are restored to their original values upon termination of the procedure.

References

Huber, P.J. 1967. The behavior of maximum likelihood estimates under non-standard conditions. *Proceeding of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1: 221-233.

White, Halbert 1984. *Asymptotic Theory for Econometricians*, New York: Academic Press.