$$\mathrm{cov}(Y_{t-1}, v_t) = -\gamma\sigma_u^2 \tag{19.23}$$

As we have shown before, if a regressor is correlated with the error term, the OLS estimators are not only biased but are also inconsistent, regardless of the sample size.

To summarize, in all the four cases we have considered there is a strong possibility that the regressor(s) is (are) not only stochastic but also correlated with the error term. As a result, the OLS estimators are biased as well inconsistent. This suggests that we either abandon OLS or find a suitable alternative(s) which will produce estimators that are at least consistent. One of the alternatives prominently suggested in the literature is the method of instrumental variable(s), which we now discuss.

19.4 The method of instrumental variables

The main problem with the use of OLS in regression models that contain one or more regressors that are correlated with the error term is that the OLS estimators are biased as well as inconsistent. Can we find a "substitute" or "proxy" variables for the suspect stochastic regressors such that the proxy variables produce consistent estimators of the true (population) regression coefficients? If we can do that successfully, such variables are called **instrumental variables** or simply **instruments**. How do we find such instruments? How do we know they are good instruments? Are there formal ways to find out if the chosen instrument is indeed a good instrument?

To answer these questions, let us start with the simple linear regression given in Eq. (19.2). Suppose in this regression that regressor X is stochastic and that it is correlated with the error term u. Suppose a variable Z is a candidate instrument for X. To be a valid instrument, Z must satisfy the following criteria:

1 *Instrument relevance*: That is, Z must be correlated, positively or negatively, with the stochastic variable for which it acts as an instrument, variable X in the present case. The greater the extent of correlation between the two variables, the better is the instrument. Symbolically,

$$\mathrm{cov}\ (X_i, Z_i) \neq 0 \tag{19.24}$$

2 *Instrument exogeneity*: Z must not be correlated with the error term u. That is,

$$\mathrm{cov}\ (Z_i, u_i) = 0 \tag{19.25}$$

3 *Not a regressor in its own right*. That is, it does not belong in the original model. If it does, the original model must be misspecified.

Before we proceed further, it may be noted that if we have a multiple regression with several regressors and some of them are correlated with the error term, we must find an instrument for each of the stochastic regressors. *In other words, there must be at least as many instruments as the number of stochastic regressors in the model.* But we will have more to say about this later.

As you can see, all these conditions may be hard to satisfy at the same time. So it is not easy to find good instruments in every application. That is why sometimes the idea

If we solve Eqs. (19.35) and (19.36) simultaneously, treating Gini as exogenous (a kind of instrument), we obtain:

$$\text{Enforcement Spending}_i = C_1 + C_2 Gini_i + u_{3i} \tag{19.37}$$

$$\text{Crime Rate}_i = D_1 + D_2 Gini_i + u_{4i} \tag{19.38}$$

where the coefficients in these equations are (nonlinear) combinations of the coefficients in Eqs. (19.35) and (19.36). Also, the error terms in these equations are (nonlinear) combinations of the error terms in Eqs. (19.35) and (19.36).

Equations (19.37) and (19.38) are known as **reduced form equations** in the language of simultaneous equation models.[26] Compared with the reduced form equations, Eqs. (19.35) and (19.36) are called the **structural equations**. In reduced form equations only exogenous or predetermined (i.e. lagged endogenous or lagged exogenous) variables appear on the right-hand side of the equations.

The coefficients of the reduced form equations are called the **reduced form coefficients**, whereas those in the structural equations are called the **structural coefficients**.

We can estimate reduced form equations by OLS. Once the reduced form coefficients are estimated, we may be able to estimate one or all of the structural coefficients. If we can estimate all the structural coefficients from the reduced form coefficients, we say the structural equations are **identified**; that is, we can obtain unique estimates of the structural coefficients. If this is not possible for one or more structural equations, we say that the equation(s) is (are) **unidentified**. If we obtain more than one estimate for one or more parameters of a structural equation, we say that equation is **overidentified**.

It may be noted that the method of obtaining the structural coefficients from the reduced form coefficients is known as the method of **indirect least squares** – we first estimate the reduced form coefficients and then try to retrieve the structural coefficients.

Shortly, we will discuss the method of **two-stage least squares (2SLS)** and show how it aids in finding instrumental variables.

Toward that purpose we now consider a numerical example.

IV

19.7 A numerical example: earnings and educational attainment of youth in the USA

The National Longitudinal Survey of Youth 1979 (NLSY79) is a repeated survey of a nationally representative sample of young males and females between ages 14 to 21 in 1979. From 1979 until 1994 the survey was conducted annually, but since then it is conducted bi-annually. Originally the core sample consisted of 3,003 males and 3,108 females.

The NLSY cross-section data is provided in 22 subsets, each subset consisting of randomly drawn sample of 540 observations: 270 males and 270 females.[27] Data are collected on a variety of socio-economic conditions and is quite extensive. The major

[26] For a detailed discussion of simultaneous equations, see Gujarati/Porter, *op cit.*, Chapters 18, 19 and 20. As noted elsewhere, this topic is no longer as prominent as it was in the 1960s and 1970s.

[27] The data used here can be obtained from http://www.bls.gov/nls/. Some of the data can be downloaded and more extensive data can be purchased.

categories of data obtained pertain to gender, ethnicity, age, years of schooling, highest educational qualification, marital status, faith, family background (mother's and father's education and number of siblings), place of living, earning, hours, years of work experience, type of employment (government, private sector, self-employed), and the region of the country (North central, North eastern, Southern and Western).

We will use some of these data for 2002 (sample subset number 22) to develop an earnings function. Following the tradition established by Jacob Mincer, we consider the following earnings function:[28]

$$\begin{aligned}\ln \mathit{Earn}_i = B_1 + B_2 S_i + B_3\ \mathit{Wexp}_i + B_4\ \mathit{Gender}_i + \\ B_5\ \mathit{Ethblack}_i + B_6\ \mathit{Ethhisp}_i + u_i\end{aligned} \qquad (19.39)$$

where ln *Earn* = log of hourly earnings in \$, S = years of schooling (highest grade completed in 2002), *Wexp* = total out-of-school work experience in years as of the 2002 interview, *Gender* = 1 for female and 0 for men, *Ethblack* = 1 for blacks, *Ethhis* = 1 for Hispanic; non-black and non-Hispanic being the left-out, or reference, category.

As you can see, some variables are quantitative and some are dummy variables. *A priori*, based on prior empirical evidence, we expect $B_2 > 0$, $B_3 > 0$, $B_4 < 0$; $B_5 < 0$, and $B_6 < 0$.

For the purpose of this chapter our concern is with the education variable S in the above model. If (native) ability and education are correlated, we should include both variables in the model. However, the ability variable is difficult to measure directly. As a result, it may be subsumed in the error term. But in that case the education variable may be correlated with the error term, thereby making education an endogenous or stochastic regressor. From our discussion of the consequences of stochastic regressor(s) it would seem that if we estimate Eq. (19.39) by OLS the coefficient of S will be biased as well as inconsistent. This is so because we may not be able to find the true impact of education on earnings that does not net out the effect of ability. Naturally, we would like to find a suitable instrument or instruments for years of schooling so that we can obtain consistent estimate of its coefficient.

Before we search for the instrument(s), let us estimate Eq. (19.39) by OLS for comparative purposes. The regression results using *Stata 10* are given in Table 19.4.

All the estimated coefficients have the expected signs and under the classical assumptions all the coefficients are statistically highly significant, the sole exception being the dummy coefficient for Hispanics.

These results show that compared to male workers, female workers on average earn less than their male counterpart, *ceteris paribus*. The average hourly earnings of black workers is lower than that of non-black non-Hispanic workers, *ceteris paribus*, which is the base category. Qualitatively, the sign of the Hispanic coefficient is negative, but the coefficient is statistically insignificant.

Noting that the regression model is log-lin, we have to interpret the coefficients of quantitative and qualitative (i.e. dummy) variables carefully (see Chapter 2 on functional forms). For quantitative variables, schooling and work experience, the estimated coefficients represent **semi-elasticities**. Thus, if schooling increases by a year, the average hourly earnings go up by about 13%, *ceteris paribus*. Similarly, if work

[28] Jacob Mincer, *Schooling, Experience, and Earnings*, Columbia University Press, 1974. See also James J. Hickman, Lance J. Lochner and Petra E. Todd, *Fifty Years of Mincer Earnings Functions*, National Bureau of Economic Research, Working Paper No. 9732, May 2003.

Table 19.4 Earnings function, USA, 2000 data set.

```
regress lEarnings s female wexp ethblack ethhisp,robust
Linear regression      Number of obs = 540
                  F(5, 534) = 50.25
                  Prob > F = 0.0000
                  R-squared = 0.3633
                  Root MSE = .50515
```

		Robust				
lEarnings	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
S	.1263493	.0097476	12.96	0.000	.1072009	.1454976
female	−.3014132	.0442441	−6.81	0.000	−.3883269	−.2144994
wexp	.0327931	.0050435	6.50	0.000	.0228856	.0427005
ethblack	−.2060033	.062988	−3.27	0.001	−.3297381	−.0822686
ethhisp	−.0997888	.088881	−1.12	0.262	−.2743881	.0748105
_cons	.6843875	.1870832	3.66	0.000	.3168782	1.051897

Note: Regress is *Stata*'s command for OLS regression. This command is followed first by the dependent variables and then the regressors. Sometimes additional options are given, such as *robust*, which computes robust standard errors – in the present case standard errors corrected for heteroscedasticity, a topic we have discussed in the chapter on heteroscedasticity.

experience goes up by 1 year, the average hourly earnings go up by about 3.2%, *ceteris paribus*.

To obtain the semi-elasticity of a dummy variable, we first take the anti-log of the dummy coefficient, subtract 1 from it, and multiply the difference by 100%. Following this procedure, for the female dummy coefficient we obtain a value of about 0.7397, which suggests that females on average earn about 26% less than the male workers. The semi-elasticities for black and Hispanic workers are about 0.81 and 0.90, respectively. This suggests that black and Hispanic workers on average earn less than the base category by about 19% and 10%, although the semi-elasticity for Hispanics is not statistically different from the base category.

As we have discussed, since the education variable does not necessarily take into account ability, it may be correlated with the error term, thus rendering it a stochastic regressor. If we can find a suitable instrument for schooling that satisfies the three requirements that we specified for a suitable instrument, we can use it and estimate the earnings function by the IV method. The question is what may be a proper instrument? This question is difficult to answer categorically. What we can do is to try one or more proxies and compare the OLS results given in Table 19.4 and see how far the OLS results are biased, if any.

In the data we have information on mother's and father's education (as measured by years of schooling), number of siblings, and the ASVAB verbal (word knowledge) and mathematics (arithmetic reasoning) scores.

In choosing a proxy or proxies we must bear in mind that such proxies must be uncorrelated with the error term but must be correlated (presumably highly) with the stochastic regressor(s) and must *not* be a candidate in their own right as regressors – in the latter case, the model used in the analysis will suffer from model specification errors. It is not always easy to accomplish these entire objectives in every case. So very often it is a matter of trial and error, supplemented by judgment or "feel" for the subject under study.

However, there are diagnostic tests which can tell us if the chosen proxy or proxies are appropriate, tests which we will consider shortly. The data gives information on mother's schooling (*Sm*), which we will use as the instrument for participant's schooling. The thinking here is that *S* and *Sm* are correlated, a reasonable assumption. For our data the correlation between the two is about 0.40. We have to assume that *Sm* is uncorrelated with the error term. We also assume that *Sm* does not belong in the participant's earning function, which seems reasonable.

We accept for the time being the validity of *Sm* as an instrument, which will be tested after we present the details of IV estimation.

To use *Sm* as the instrument for *S* and estimate the earnings function, we proceed in two stages:

> **Stage 1:** We regress the suspected endogenous variable (*S*) on the chosen instrument (*Sm)* and the other regressors in the original model and obtain the estimated value of *S* from this regression; call it *S*-hat.
>
> **Step 2:** We then run the earnings regression on the regressors included in the original model but replace the education variable by its value estimated from the Step 1 regression.

This method of estimating the parameters of the model of interest is appropriately called the method of **two-stage least squares (2SLS)**, for we apply OLS twice. Therefore *the IV method is also known as 2SLS.*

Let us illustrate this method (Table 19.5). Using the estimated *S*-hat value from this regression, we obtain the second stage regression 2SLS (Table 19.6).

Note that in this (log) earnings function, unlike the one reported in Table 19.4, we use *S-hat* (estimated from the first-stage of 2SLS) instead of *S* as the regressor. However, *the standard errors reported in Table 19.6 are not correct* because they are based on the incorrect estimator of the variance of the error term, u_i. The formula to correct

Table 19.5 First stage of 2SLS with *Sm* as instrument.

```
regress s female wexp ethblack ethhisp sm
```

Source	SS	df	MS	Number of obs = 540
				F(5, 534) = 35.06
Model	822.26493	5	164.452986	Prob > F = 0.0000
Residual	2504.73322	534	4.69051165	R-squared = 0.2471
				Adj R-squared = 0.2401
Total	3326.99815	539	6.17253831	Root MSE = 2.1658

s	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
female	–.0276157	.1913033	–0.14	0.885	–.4034151	.3481837
wexp	–.1247765	.0203948	–6.12	0.000	–.1648403	–.0847127
ethblack	–.9180353	.2978136	–3.08	0.002	–1.503065	–.3330054
ethhisp	.4566623	.4464066	1.02	0.307	–.420266	1.333591
Sm	.3936096	.0378126	10.41	0.000	.3193298	.4678893
_cons	11.31124	.6172187	18.33	0.000	10.09876	12.52371

Table 19.6 Second stage of 2SLS of the earnings function.

regress lEarnings s_hat female wexp ethblack ethhisp

Source	SS	df	MS	Number of obs = 540
				F(5, 534) = 24.26
Model	39.6153236	5	7.92306472	Prob > F = 0.0000
Residual	174.395062	534	.326582514	R-squared = 0.1851
				Adj R-squared = 0.1775
Total	214.010386	539	.397050809	Root MSE = .57147

lEarnings	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S_hat	.140068	.0253488	5.53	0.000	.0902724	.1898636
female	−.2997973	.0505153	−5.93	0.000	−.3990304	−.2005642
wexp	.0347099	.0064313	5.40	0.000	.0220762	.0473437
ethblack	−.1872501	.0851267	−2.20	0.028	−.3544744	−.0200258
ethhisp	−.0858509	.1146507	−0.75	0.454	−.3110726	.1393708
_cons	.4607716	.4257416	1.08	0.280	−.3755621	1.297105

the estimated standard errors is rather involved. So it is better to use software like *Stata* or *Eviews* that not only correct the standard errors, but also obtain the 2SLS estimates without explicitly going through the cumbersome two-step procedure.

To do this, we can use the **ivreg** (instrumental variable regression) command of *Stata*. Using this command, we obtain the results in Table 19.7.

Observe that the estimated coefficients in the preceding two tables are the same, but the standard errors are different. As pointed out, we should rely on the standard errors reported in Table 19.7. Also notice that with the ivreg command we need only one table, instead of two, as in the case of the rote application of 2SLS.

19.8 Hypothesis testing under IV estimation

Now that we have estimated the earnings function using the IV method, how do we test hypotheses about an individual regression coefficient (like the *t* test in CLRM) and hypotheses about several coefficients collectively (like the *F* test of CLRM)? For the time being, assume that the instrument we have chosen (*Sm*) is the appropriate instrument for schooling, although we will provide a test to find out if this is indeed correct in the following section.

As Davidson and MacKinnon note, "Because the finite sample distributions of IV estimators are almost never known, exact tests of hypotheses based on such estimators are almost never available".[29]

However, in large samples it can be shown the IV estimator is approximately normally distributed with mean and variance as shown in Eq. (19.30). Therefore, instead of using the standard *t* test, we use the *z* test (i.e. the standard normal distribution) as shown in Table 19.7. The *z* values in this table are all individually highly statistically significant, save the coefficient of Hispanic.

[29] Davidson and MacKinnon, *op cit.*, pp. 330–5.

Table 19.7 One step estimates of the earnings function (with robust standard errors).

. ivregress 2sls lEarnings female wexp ethblack ethhisp (S = Sm),robust
(Instrumental variables (2SLS) regression Number of obs = 540
Wald chi2(5) = 138.45
Prob > chi2 = 0.0000
R-squared = 0.3606
Root MSE = .50338

		Robust				
lEarnings	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
s	.140068	.0217263	6.45	0.000	.0974852	.1826508
female	–.2997973	.043731	–6.86	0.000	–.3855085	–.2140861
wexp	.0347099	.0055105	6.30	0.000	.0239095	.0455103
ethblack	–.1872501	.0634787	–2.95	0.003	–.3116661	–.0628342
ethhisp	–.0858509	.0949229	–0.90	0.366	–.2718963	.1001945
_cons	.4607717	.3560759	1.29	0.196	–.2371241	1.158668

Instrumented: S
Instruments: female wexp ethblack ethhisp sm

To test joint hypotheses of two or more coefficients, instead of using the classical F test we use the Wald test, which is a large sample test. The Wald statistic follows the *chi-square* statistic with degrees of freedom equal to the number of regressors estimated: 5 in Table 19.7. The null hypothesis, as in the usual F test, is that all the regressor coefficients are zero simultaneously, that is, collectively none of the regressors have any bearing on (log) earnings. In our example the chi-square value is about 138 and the probability of obtaining such a chi-square value or greater is practically nil.

In other words, collectively all the regressors have important impact on hourly earnings.

A caution on the use of R^2 in IV estimation

Although we have presented the R^2 for the IV regressions given in the preceding two tables, it does not have the same interpretation as in the classical linear regression model and sometimes it can actually be negative. Hence the reported R^2 in IV regressions should be taken with a grain of salt.[30]

Diagnostic testing

Having presented the basics of IV estimation, we now consider several questions regarding the IV methodology. Because of their importance in practice, we discuss these questions sequentially.

A How do we know that a regressor is truly endogenous?

B How do we find out if an instrument is weak or strong?

[30] The conventionally computed coefficient of determination is defined as $R^2 = 1 - RSS/TSS$, but in case of IV estimation RSS can be greater than TSS, making R^2 negative.

C What happens if we introduce several instruments for a stochastic regressor? And how do we test the validity of all the instrument?

D How do we estimate a model when there is more than one stochastic regressor?

In what follows we answer these questions sequentially.

19.9 Test of endogeneity of a regressor

We have been working on the assumption that *S* in our example is endogenous. But we can test this assumption explicitly by using one of the variants of the Hausman test. This test is relatively simple, and involves two steps:

Step 1: We regress the endogenous *S* on all the (nonstochastic) regressors in the earnings function plus the instrumental variable(s) and obtain residuals from this regression; call it *S*-hat.

Step 2: We then regress lEarnings on all the regressors, including the (stochastic) *S* and the residuals from Step I. If in this regression the *t* value of the residuals variable is statistically significant, we conclude that *S* endogenous or stochastic. If it is not, then there is no need for IV estimation, for in that case *S* is its own instrument.

Returning to our example, we obtain the results in Table 19.8.

The results of the second step regression are as given in Table 19.9.

Since the coefficient of *shat* is not statistically significant, it would seem that schooling is not an endogenous variable. But we should not take these results at face value because we have cross-sectional data and heteroscedasticity is usually a problem in such data. Therefore we need to find heteroscedasticity-corrected standard error, such as the HAC standard errors discussed the chapter on heteroscedasticity.

Table 19.8 Hausman test of endogeneity of schooling: first step result.

```
regress s female wexp ethblack ethhisp sm

Source         SS          df        MS          Number of obs = 540
                                                 F( 5, 534) = 35.06
Model       822.26493      5      164.452986     Prob > F = 0.0000
Residual    2504.73322     534    4.69051165     R-squared = 0.2471
                                                 Adj R-squared = 0.2401
Total       3326.99815     539    6.17253831     Root MSE = 2.1658

S           Coef.        Std. Err.    t        P>|t|    [95% Conf. Interval]
female      -.0276157    .1913033     -0.14    0.885    -.4034151    .3481837
wexp        -.1247765    .0203948     -6.12    0.000    -.1648403    -.0847127
ethblack    -.9180353    .2978136     -3.08    0.002    -1.503065    -.3330054
ethhisp     .4566623     .4464066     1.02     0.307    -.420266     1.333591
sm          .3936096     .0378126     10.41    0.000    .3193298     .4678893
_cons       11.31124     .6172187     18.33    0.000    10.09876     12.52371

. predict shat,residuals
```

Table 19.9 Hausman test of endogeneity of schooling: second step results.

```
egress lEarnings s female wexp ethblack ethhisp shat
```

Source	SS	df	MS	
				Number of obs = 540
				F(6, 533) = 50.80
Model	77.8586985	6	12.9764498	Prob > F = 0.0000
Residual	136.151687	533	.255444066	R-squared = 0.3638
				Adj R-squared = 0.3566
Total	214.010386	539	.397050809	Root MSE = .50541

lEarnings	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	.140068	.0224186	6.25	0.000	.0960283	.1841077
female	–.2997973	.044676	–6.71	0.000	–.38756	–.2120346
wexp	.0347099	.0056879	6.10	0.000	.0235365	.0458834
ethblack	–.1872501	.0752865	–2.49	0.013	–.3351448	–.0393554
ethhisp	–.0858509	.1013977	–0.85	0.398	–.2850391	.1133373
shat	–.0165025	.0245882	–0.67	0.502	–.0648041	.0317992
_cons	.4607717	.3765282	1.22	0.222	–.2788895	1.200433

Table 19.10 Hausman endogeneity test with robust standard errors.

```
regress lEarnings s female wexp shat,vce(robust)
Linear regression                          Number of obs = 540
F( 4, 535) = 59.14
Prob > F = 0.0000
R-squared = 0.3562
Root MSE = .50747
```

		Robust				
lEarnings	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	.1642758	.0209439	7.84	0.000	.1231334	.2054183
female	–.3002845	.0443442	–6.77	0.000	–.3873947	–.2131744
wexp	.0390386	.0053869	7.25	0.000	.0284565	.0496207
shat	–.0407103	.022955	–1.77	0.077	–.0858034	.0043828
_cons	.0311987	.3380748	0.09	0.927	–.6329182	.6953156

We can use the robust standard error command in *Stata* to obtain the heteroscedasticity-corrected standard errors, which are given in Table 19.10.

Now the coefficient of the *shat* variable is statistically significant, at about the 8% level, indicating that education (schooling) seems to be endogenous.

19.10 How to find whether an instrument is weak or strong

If an instrument used in the analysis is weak in the sense that it is poorly correlated with the stochastic regressor for which it is an instrument, the IV estimator can be severely biased and its sampling distribution is not approximately normal, even in large samples. As a consequence, the IV standard errors and the confidence intervals based on them are highly misleading, leading to hypotheses tests that are unreliable.

To see why this is the case, refer to Eq. (19.30). If ρ_{xz} in this equation is zero, the variance of the IV estimator is infinite. If ρ_{xz} is not exactly zero, but very low (the case of a weak instrument), the IV estimator is not normally distributed, even in large samples. But how do we decide in a given case whether an instrument is weak?

In the case of a single endogenous regressor a rule of thumb says that an F statistic of less than 10 in the first step of the Hausman test suggests that the chosen instrument is weak. If it is greater than 10, it probably is not a weak instrument.[31] In the case of a single (stochastic) regressor, this rule translates into a a t value of about 3.2 because of the relationship between the F and t statistics, namely, that $F_{1,k} = t_k^2$, where for the F statistic has 1 df in the numerator and k df in the denominator.

On that score, in our example *Sm* (mother's schooling) seems to be a strong instrument for *S* because the value of the F statistic in the first stage of the two-stage procedure is about 35, which exceeds the threshold value of 10. But this rule of thumb, like most rules of thumb, should not be used blindly.

19.11 The case of multiple instruments

Since there are competing instruments, education may be correlated with more than one instrumental variable. To allow for this possibility, we can include more than one instrument in the IV regression. This is often done with the aid of **two-stage least squares** (**2SLS**) that we just discussed.

Step 1: We regress the suspected variable on all the instruments, and obtain the estimated value of the regressor.

Step 2: We then run the earnings regression on the regressors included in the original model but replace the education variable by its value estimated from the Step 1 regression.

We can replace this two-step procedure by a single step by invoking Stata's *ivreg* command by including several instruments simultaneously, as the following example demonstrates.

For our earnings regression, in addition to mother's education (*Sm*), we can include father's schooling (*Sf*), and the number of siblings as instruments in the regression of earnings on education (*S*), gender (*female* = 1), years of work experience (*wexp*), ethnicity (dummies for black and Hispanics).

Step 1: Regress schooling (*S*) on all the original (nonstochastic) regressors and the instruments. From this regression we obtain the estimated value of *S*, say, $\hat{s}$.

Step 2: We now regress earnings on gender, wexp, ethnic dummies, and $\hat{s}$, the latter estimated from Step 1.

See Table 19.11. Compared to a single instrument in Table 19.7, when we introduced multiple instruments, the coefficient of *S* (education) has gone up a bit, but it is still

[31] Why 10? The slightly technical answer for this can be found in James H. Stock and Mark W. Watson, *Introduction to Econometrics*, 2nd edn, Pearson/Addison Wesley, Boston, 2007, p. 466. If the F statistic exceeds 10, it suggests that the small sample bias of the IV estimate is less than 10% of the OLS bias. Remember that in cases of stochastic regressor(s) OLS is biased in small as well as large samples.

Table 19.11 Earnings function with several instruments.

```
ivreg lEarnings female wexp ethblack ethhisp (S=sm sf siblings),robust
Instrumental variables (2SLS) regression          Number of obs = 540
F( 5, 534) = 26.63
Prob > F = 0.0000
R-squared = 0.3492
Root MSE = .51071
```

		Robust				
lEarnings	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
s	.1579691	.0216708	7.29	0.000	.1153986	.2005396
female	–.2976888	.0441663	–6.74	0.000	–.3844499	–.2109278
wexp	.0372111	.005846	6.37	0.000	.0257271	.0486951
ethblack	–.1627797	.0625499	–2.60	0.010	–.2856538	–.0399056
ethhisp	–.0676639	.098886	–0.68	0.494	–.2619172	.1265893
_cons	.1689836	.3621567	0.47	0.641	–.542443	.8804101

```
Instrumented: S
Instruments: female wexp ethblack ethhisp sm sf siblings
```

significantly higher than the OLS regression. But notice again that the relative standard error of this coefficient is higher than its OLS counterpart, again reminding us that IV estimators may be less efficient.

Testing the validity of surplus instruments

Earlier we stated that the number of instruments must be at least equal to the number of stochastic regressor. So, technically for our earnings regression one instrument will suffice, as in Table 19.7 where we used *Sm* (mother's education) as an instrument. In Table 19.11 we have three instruments, two more than the absolute minimum. How do we know that they are valid in the sense they are correlated with education but are not correlated with the error term? In simple terms, are they relevant?

Before we provide an answer to this question, it is worth mentioning the following:

1 If the number of instruments (*I*) equals the number of endogenous regressors, say *K*, we say that the regression coefficients are **exactly identified**, that is, we can obtain unique estimates of them.

2 If the number of instruments (*I*) exceeds the number of regressors, *K*, the regression coefficients are **overidentified**, in which case we may obtain more than one estimate of one or more of the regressors.

3 If the number of instruments is less than the number of endogenous regressors, the regression coefficients are **underidentified**, that is, we cannot obtain unique values of the regression coefficients.[32]

[32] The topic of identification is usually discussed in the context of simultaneous equation models. For details, see Gujarati/Porter, *op cit.*, Chapters 18, 19 and 20.

In the present example, if we use three instruments (*Sm*, *Sf*, *siblings*), we have extra or surplus instruments. How do we find out the validity of the extra instrum We can proceed as follows:[33]

1 Obtain the IV estimates of the earnings regression coefficients including al (exogenous) variables in the model plus all the instruments, three in the pre case.

2 Obtain residuals from this regression; call them *Res*.

3 Regress *Res* on all the original regressors, including the instruments, and ob the R^2 value from this regression.

4 Multiply the R^2 value obtained in Step 3 by the sample size (n = 540). Tha obtain nR^2. If all the surplus instruments are valid, it can be shown that nR^2 ~ that is nR^2 follows the chi-square distribution with m df, where m is the numb surplus instruments; two in our case.

5 If the estimated chi-square value exceeds the critical chi-square value, say, the level, we conclude that at least one surplus instrument is *not* valid.

We have already given the IV estimates of the earnings regression including three instruments in Table 19.11. From this regression we obtained the following gression as per Step 3 above. The results are given in Table 19.12.

We need not worry about the coefficients in this table. The important entity her R^2, which is 0.0171. Multiplying this by the sample size of 540, we obtain nR^2 = 9.2 The chi-square 1% significance value for 2 df is about 9.21. So the compu chi-square value is highly significant, which suggests that at least one surp

Table 19.12 Test of surplus instruments.

regress Res female wexp ethblack ethhisp sm sf siblings

Source	SS	df	MS
Model	2.38452516	7	.340646452
Residual	136.894637	532	.257320746
Total	139.279162	539	.258402898

Number of obs = 540
F(7, 532) = 1.32
Prob > F = 0.2366
R-squared = 0.0171
Adj R-squared = 0.0042
Root MSE = .50727

Res	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
female	−.0067906	.0449329	−0.15	0.880	−.0950584	.0814771
wexp	−.0001472	.0047783	−0.03	0.975	−.0095339	.0092396
ethblack	−.0034204	.0708567	−0.05	0.962	−.1426136	.1357728
ethhisp	−.0197119	.1048323	−0.19	0.851	−.225648	.1862241
sm	−.0206955	.0110384	−1.87	0.061	−.0423797	.0009887
sf	.0215956	.0082347	2.62	0.009	.0054191	.0377721
siblings	.0178537	.0110478	1.62	0.107	−.0038489	.0395563
_cons	−.0636028	.1585944	−0.40	0.689	−.3751508	.2479452

[33] This discussion is based on R. Carter Hill, William E. Griffiths and Guay C. Lim, *Principles Econometrics*, 3rd edn, John Wiley & Sons, New York, 2008, pp. 289–90.

instrument is not valid. We could throw away two of the three instruments, as we need just one to identify (i.e. estimate) the parameters. Of course, it is not a good idea to throw away instruments. There are procedures in the literature to use weighted least-squares to obtain consistent IV estimates. We leave the reader to discover more about this in the references (see the Stock and Watson text for additional details).

19.12 Regression involving more than one endogenous regressor

So far we have concentrated on a single endogenous regressor. How do we deal with a situation of two or more stochastic regressors? Suppose in our earnings regression we think that the regressor work experience (*wexp*) is also stochastic. Now we have two stochastic regressors, education (*S*) and *wexp*. We can use 2SLS method to handle this case.

Just as one instrument (*Sm*) sufficed to identify the impact of education on earnings, we need another instrument for *wexp*. We have a variable, *age*, in our data. So we can use it to proxy *wexp*. We can treat age as truly exogenous. To estimate the earnings regression with two stochastic regressors, we proceed as follows:

Stage 1: We regress each endogenous regressor on all exogenous variables and obtain the estimated values of these regressors.

Stage 2: We estimate the earnings function using all exogenous variables and the estimated values of the endogenous regressors from Stage 1.

Actually, we do not have to go trough this two-stage procedure, for packages like *Stata* can do this in one step. The results are given in Table 19.13.

This regression shows that the return to education per incremental year is about 13.4%, *ceteris paribus*. The regressors female and ethblack are individually highly significant, as before, but the work experience variable is not statistically significant.

Table 19.13 IV Estimation with two endogenous regressors.

```
. ivregress 2sls lEarnings female ethblack ethhisp (s wexp = sm age)
Instrumental variables (2SLS) regression Number of obs = 540
Wald chi2(5) = 139.51
Prob > chi2 = 0.0000
R-squared = 0.3440
Root MSE = .50987
```

lEarnings	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
s	.1338489	.0229647	5.83	0.000	.0888389	.1788589
wexp	.0151816	.0158332	0.96	0.338	−.0158509	.0462141
female	−.3378409	.0535152	−6.31	0.000	−.4427287	−.2329531
ethblack	−.215774	.0787299	−2.74	0.006	−.3700818	−.0614663
ethhisp	−.1252153	.1063871	−1.18	0.239	−.3337301	.0832995
_cons	.8959276	.4964128	1.80	0.071	−.0770236	1.868879

```
Instrumented: s wexp
Instruments: female ethblack ethhisp sm age
```

We have argued that IV estimation will give consistent estimates in case a regresso has serious measurement errors, even though the estimates thus obtained are inefficient. But if measurement errors are absent OLS and IV estimates are both consistent in which case we should choose OLS because it is more efficient. Thus it behooves u to find out if the instruments chosen for consideration are valid.

A test developed by Durbin, Wu and Hausman (DWH), but popularly known as the **Hausman test**, is one that is used in applied econometrics to test the validity of instruments.[34]

Although the mathematics of the test is involved, the basic idea behind the DWH test is quite simple. We compare the differences between OLS and IV coefficients of al the variables in the model, and obtain, say, $m = (b^{\text{OLS}} - b^{\text{IV}})$. Under the null hypothesis that $m = 0$, it can be shown that m is distributed as the chi-square distribution with degrees of freedom equal to the number of coefficients compared. If m turns out to be zero, it would suggest that the (stochastic) regressor is not correlated with the error term and we can use OLS in lieu of IV, because OLS estimators are more efficient.

The results of the DWH test based on *Stata* are given in Table 19.14. In this table the column (b) gives the estimates of the model under IV (earniv) and column (B) gives the estimates obtained by OLS (earnols). The next column gives the difference between the two sets of coefficients (m) and the last column gives the standard error of the difference between the two estimates.

Table 19.14 The DWH test of instrument validity for the earnings function.

hausman earniv earnols1, constant

	Coefficients			
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	earniv	earnols	Difference	S.E.
educ	.1431384	.1082223	.0349161	.0273283
female	−.2833126	−.2701109	−.0132017	.0121462
wexp	.0349416	.029851	.0050906	.0040397
ethblack	−.1279853	−.1165788	−.0114065	.0138142
ethhisp	−.0506336	−.0516381	.0010045	.0141161
asvab02	.0044979	.0093281	−.0048302	.0037962
_cons	.1715716	.483885	−.3123135	.2454617

b = consistent under Ho and Ha; obtained from ivreg
B = inconsistent under Ha, efficient under Ho; obtained from regress
Test: Ho: difference in coefficients not systematic
chi2(7) = (b−B)'[(V_b−V_B)^(−1)](b−B)
= 1.63
Prob>chi2 = 0.9774

[34] See Jerry Hausman, Specification tests in econometrics, *Econometrica*, vol. 46, no. 6, 1978, pp. 1251–71; James Durbin, Errors in variables, *Review of the International Statistical Institute*, vol. 22, no. 1, 1954, pp. 23–32, and Wu, De-Min, Alternative tests of independence between stochastic regressors and disturbances, *Econometrica*, vol. 41, no. 4, 1073, 733–50. See also A. Nakamura and M. Nakamura, On the relationship among several specification error tests presented by Durbin, Wu, and Hausman, *Econometrica*, vol. 49, November 1981, pp. 1583–8.