Title

stata.com

reg3 — Three-stage estimation for systems of simultaneous equations

Description

`reg3` estimates a system of structural equations, where some equations contain endogenous variables among the explanatory variables. Estimation is via three-stage least squares (3SLS); see Zellner and Theil (1962). Typically, the endogenous explanatory variables are dependent variables from other equations in the system. `reg3` supports iterated GLS estimation and linear constraints.

`reg3` can also estimate systems of equations by seemingly unrelated regression estimation (SURE), multivariate regression (MVREG), and equation-by-equation ordinary least squares (OLS) or two-stage least squares (2SLS).

Nomenclature

Under 3SLS or 2SLS estimation, a *structural equation* is defined as one of the equations specified in the system. A *dependent variable* will have its usual interpretation as the left-hand-side variable in an equation with an associated disturbance term. All dependent variables are explicitly taken to be *endogenous* to the system and are treated as correlated with the disturbances in the system's equations. Unless specified in an `endog()` option, all other variables in the system are treated as *exogenous* to the system and uncorrelated with the disturbances. The exogenous variables are taken to be *instruments* for the endogenous variables.

Quick start

System of equations regressing `y1` on `x1`, `x2`, and `y2` and `y2` on `x1`, `x3`, and `y1`

```
reg3 (y1 x1 x2 y2) (y2 x1 x3 y1)
```

Same as above, but name equations `eq1` and `eq2`

```
reg3 (eq1: y1 x1 x2 y2) (eq2: y2 x1 x3 y1)
```

Same as above, but specify that `x1` is endogenous and add instrumental variable `v1`

```
reg3 (eq1: y1 x1 x2 y2) (eq2: y2 x1 x3 y1), endog(x1) exog(v1)
```

Same as above, but with iterated estimation

```
reg3 (eq1: y1 x1 x2 y2) (eq2: y2 x1 x3 y1), endog(x1) exog(v1) ireg3
```

Menu

Statistics > Endogenous covariates > Three-stage least squares

Syntax

Basic syntax

```
reg3 (depvar1 varlist1) (depvar2 varlist2) ... (depvarN varlistN) [if] [in] [weight]
```

Full syntax

```
reg3 ([eqname1:]depvar1a [depvar1b ...=] varlist1 [, noconstant])
    ([eqname2:]depvar2a [depvar2b ...=] varlist2 [, noconstant])
    ...
    ([eqnameN:]depvarNa [depvarNb ...=] varlistN [, noconstant])
    [if] [in] [weight] [, options]
```

options	Description
Model	
ireg3	iterate until estimates converge
3sls	three-stage least squares; the default
2sls	two-stage least squares
ols	ordinary least squares (OLS)
sure	seemingly unrelated regression estimation (SURE)
mvreg	sure with OLS degrees-of-freedom adjustment
corr(*correlation*)	unstructured or independent correlation structure; default is unstructured
exog(*varlist*)	exogenous variables not specified in system equations
endog(*varlist*)	additional right-hand-side endogenous variables
inst(*varlist*)	full list of exogenous variables
allexog	all right-hand-side variables are exogenous
noconstant	suppress constant from instrument list
constraints(*constraints*)	apply specified linear constraints
SE/Robust	
vce(*vcetype*)	*vcetype* may be unadjusted, robust, or cluster *clustvar*
df adj.	
small	report small-sample statistics
dfk	use small-sample adjustment
dfk2	use alternate adjustment
Reporting	
level(#)	set confidence level; default is level(95)
first	report first-stage regression
nocnsreport	do not display constraints
display_options	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling

Optimization

optimization_options	control the optimization process; seldom used
noheader	suppress display of header
notable	suppress display of coefficient table
nofooter	suppress display of footer
coeflegend	display legend instead of statistics

varlist$_1$, ..., *varlist*$_N$ and the `exog()` and the `inst()` varlist may contain factor variables; see [U] **11.4.3 Factor variables**. You must have the same levels of factor variables in all equations that have factor variables.

depvar and *varlist* may contain time-series operators; see [U] **11.4.4 Time-series varlists**.

`bootstrap`, `by`, `collect`, `fp`, `jackknife`, `rolling`, and `statsby` are allowed; see [U] **11.1.10 Prefix commands**.

Weights are not allowed with the `bootstrap` prefix; see [R] **bootstrap**.

`aweights` are not allowed with the `jackknife` prefix; see [R] **jackknife**.

`aweights` and `fweights` are allowed; see [U] **11.1.6 weight**.

`noheader`, `notable`, `nofooter`, and `coeflegend` do not appear in the dialog box.

See [U] **20 Estimation and postestimation commands** for more capabilities of estimation commands.

Explicit equation naming (*eqname*`:`) cannot be combined with multiple dependent variables in an equation specification.

Options

Model

`ireg3` causes `reg3` to iterate over the estimated disturbance covariance matrix and parameter estimates until the parameter estimates converge. Although the iteration is usually successful, there is no guarantee that it will converge to a stable point. Under SURE, this iteration converges to the maximum likelihood estimates.

`3sls` specifies the full 3SLS estimation of the system and is the default for `reg3`.

`2sls` causes `reg3` to perform equation-by-equation 2SLS on the full system of equations. This option implies `dfk`, `small`, and `corr(independent)`.

Cross-equation testing should not be performed after estimation with this option. With `2sls`, no covariance is estimated between the parameters of the equations. For cross-equation testing, use `3sls`.

`ols` causes `reg3` to perform equation-by-equation OLS on the system—even if dependent variables appear as regressors or the regressors differ for each equation; see [MV] **mvreg**. `ols` implies `allexog`, `dfk`, `small`, and `corr(independent)`; `nodfk` and `nosmall` may be specified to override `dfk` and `small`.

The covariance of the coefficients between equations is not estimated under this option, and cross-equation tests should not be performed after estimation with `ols`. For cross-equation testing, use `sure` or `3sls` (the default).

`sure` causes `reg3` to perform a SURE of the system—even if dependent variables from some equations appear as regressors in other equations; see [R] **sureg**. `sure` is a synonym for `allexog`.

`mvreg` is identical to `sure`, except that the disturbance covariance matrix is estimated with an OLS degrees-of-freedom adjustment—the `dfk` option. If the regressors are identical for all equations, the parameter point estimates will be the standard MVREG results. If any of the regressors differ, the point estimates are those for SURE with an OLS degrees-of-freedom adjustment in computing the covariance matrix. `nodfk` and `nosmall` may be specified to override `dfk` and `small`.

`corr(`*correlation*`)` specifies the assumed form of the correlation structure of the equation disturbances and is rarely requested explicitly. For the family of models fit by `reg3`, the only two allowable correlation structures are `unstructured` and `independent`. The default is `unstructured`.

This option is used almost exclusively to estimate a system of equations by 2SLS or to perform OLS regression with `reg3` on multiple equations. In these cases, the correlation is set to `independent`, forcing `reg3` to treat the covariance matrix of equation disturbances as diagonal in estimating model parameters. Thus, a set of two-stage coefficient estimates can be obtained if the system contains endogenous right-hand-side variables, or OLS regression can be imposed, even if the regressors differ across equations. Without imposing independent disturbances, `reg3` would estimate the former by 3SLS and the latter by SURE.

Any tests performed after estimation with the `independent` option will treat coefficients in different equations as having no covariance; cross-equation tests should not be used after specifying `corr(independent)`.

`exog(`*varlist*`)` specifies additional exogenous variables that are included in none of the system equations. This can occur when the system contains identities that are not estimated. If implicitly exogenous variables from the equations are listed here, `reg3` will just ignore the additional information. Specified variables will be added to the exogenous variables in the system and used in the first stage as instruments for the endogenous variables. By specifying dependent variables from the structural equations, you can use `exog()` to override their endogeneity.

`endog(`*varlist*`)` identifies variables in the system that are not dependent variables but are endogenous to the system. These variables must appear in the variable list of at least one equation in the system. Again, the need for this identification often occurs when the system contains identities. For example, a variable that is the sum of an exogenous variable and a dependent variable may appear as an explanatory variable in some equations.

`inst(`*varlist*`)` specifies a full list of all exogenous variables and may not be used with the `endog()` or `exog()` options. It must contain a full list of variables to be used as instruments for the endogenous regressors. Like `exog()`, the list may contain variables not specified in the system of equations. This option can be used to achieve the same results as the `endog()` and `exog()` options, and the choice is a matter of convenience. Any variable not specified in the *varlist* of the `inst()` option is assumed to be endogenous to the system. As with `exog()`, including the dependent variables from the structural equations will override their endogeneity.

`allexog` indicates that all right-hand-side variables are to be treated as exogenous—even if they appear as the dependent variable of another equation in the system. This option can be used to enforce a SURE or MVREG estimation even when some dependent variables appear as regressors.

`noconstant`, `constraints(`*constraints*`)`; see [R] **Estimation options**.

SE/Robust

`vce(`*vcetype*`)` specifies the type of standard error reported, which includes types that are derived from asymptotic theory (`unadjusted`), that are robust to some kinds of misspecification (`robust`), and that allow for intragroup correlation (`cluster` *clustvar*); see [R] ***vce_option***.

`vce(unadjusted)`, the default, specifies that an unadjusted (nonrobust) VCE matrix be used; this results in efficient estimates when assuming homoskedasticity.

df adj.

`small` specifies that small-sample statistics be computed. It shifts the test statistics from χ^2 and z statistics to F statistics and t statistics. This option is intended primarily to support MVREG. Although the standard errors from each equation are computed using the degrees of freedom for

the equation, the degrees of freedom for the t statistics are all taken to be those for the first equation. This approach poses no problem under MVREG because the regressors are the same across equations.

`dfk` specifies the use of an alternative divisor in computing the covariance matrix for the equation residuals. As an asymptotically justified estimator, `reg3` by default uses the number of sample observations n as a divisor. When the `dfk` option is set, a small-sample adjustment is made, and the divisor is taken to be $\sqrt{(n-k_i)(n-k_j)}$, where k_i and k_j are the number of parameters in equations i and j, respectively.

`dfk2` specifies the use of an alternative divisor in computing the covariance matrix for the equation errors. When the `dfk2` option is set, the divisor is taken to be the mean of the residual degrees of freedom from the individual equations.

Reporting

`level(#)`; see [R] **Estimation options**.

`first` requests that the first-stage regression results be displayed during estimation.

`nocnsreport`; see [R] **Estimation options**.

display_options: `noci`, `nopvalues`, `noomitted`, `vsquish`, `noemptycells`, `baselevels`, `allbaselevels`, `nofvlabel`, `fvwrap(#)`, `fvwrapon(`*style*`)`, `cformat(%`*fmt*`)`, `pformat(%`*fmt*`)`, `sformat(%`*fmt*`)`, and `nolstretch`; see [R] **Estimation options**.

Optimization

optimization_options control the iterative process that minimizes the sum of squared errors when `ireg3` is specified. These options are seldom used.

`iterate(#)` specifies the maximum number of iterations. When the number of iterations equals #, the optimizer stops and presents the current results, even if the convergence tolerance has not been reached. The default is the number set using `set maxiter`, which is 300 by default.

`trace` adds to the iteration log a display of the current parameter vector.

`log` and `nolog` specify whether to display the iteration log. The iteration log is displayed by default unless you used `set iterlog off` to suppress it; see `set iterlog` in [R] ***set iter***.

`tolerance(#)` specifies the tolerance for the coefficient vector. When the relative change in the coefficient vector from one iteration to the next is less than or equal to #, the optimization process is stopped. `tolerance(1e-6)` is the default.

The following options are available with `reg3` but are not shown in the dialog box:

`noheader` suppresses display of the header reporting the estimation method and the table of equation summary statistics.

`notable` suppresses display of the coefficient table.

`nofooter` suppresses display of the footer reporting the list of endogenous and exogenous variables in the model.

`coeflegend`; see [R] **Estimation options**.

Remarks and examples

stata.com

reg3 estimates systems of structural equations where some equations contain endogenous variables among the explanatory variables. Generally, these endogenous variables are the dependent variables of other equations in the system, though not always. The disturbance is correlated with the endogenous variables—violating the assumptions of OLS. Further, because some of the explanatory variables are the dependent variables of other equations in the system, the error terms among the equations are expected to be correlated. reg3 uses an instrumental-variables approach to produce consistent estimates and generalized least squares (GLS) to account for the correlation structure in the disturbances across the equations. Good general references on three-stage estimation include Davidson and MacKinnon (1993, 651–661) and Greene (2018, 363–365).

Three-stage least squares can be thought of as producing estimates from a three-step process.

Step 1. Develop instrumented values for all endogenous variables. These instrumented values can simply be considered as the predicted values resulting from a regression of each endogenous variable on all exogenous variables in the system. This stage is identical to the first step in 2SLS and is critical for the consistency of the parameter estimates.

Step 2. Obtain a consistent estimate for the covariance matrix of the equation disturbances. These estimates are based on the residuals from a 2SLS estimation of each structural equation.

Step 3. Perform a GLS-type estimation using the covariance matrix estimated in the second stage and with the instrumented values in place of the right-hand-side endogenous variables.

❑ Technical note

The estimation and use of the covariance matrix of disturbances in three-stage estimation is almost identical to the SURE method—sureg. As with SURE, using this covariance matrix improves the efficiency of the three-stage estimator. Even without the covariance matrix, the estimates would be consistent. (They would be 2SLS estimates.) This improvement in efficiency comes with a caveat. All the parameter estimates now depend on the consistency of the covariance matrix estimates. If one equation in the system is misspecified, the disturbance covariance estimates will be inconsistent, and the resulting coefficients will be biased and inconsistent. Alternatively, if each equation is estimated separately by 2SLS ([R] **regress**), only the coefficients in the misspecified equation are affected.

❑

❑ Technical note

If an equation is just identified, the 3SLS point estimates for that equation are identical to the 2SLS estimates. However, as with sureg, even if all equations are just identified, fitting the model via reg3 has at least one advantage over fitting each equation separately via ivregress; by using reg3, tests involving coefficients in different equations can be performed easily using test or testnl.

❑

▷ Example 1

A simple macroeconomic model relates consumption (consump) to private and government wages paid (wagepriv and wagegovt). Simultaneously, private wages depend on consumption, total government expenditures (govt), and the lagged stock of capital in the economy (capital1). Although this is not a plausible model, it does meet the criterion of being simple. This model could be written as

$$\texttt{consump} = \beta_0 + \beta_1\,\texttt{wagepriv} + \beta_2\,\texttt{wagegovt} + \epsilon_1$$

$$\texttt{wagepriv} = \beta_3 + \beta_4\,\texttt{consump} + \beta_5\,\texttt{govt} + \beta_6\,\texttt{capital1} + \epsilon_2$$

If we assume that this is the full system, consump and wagepriv will be endogenous variables, with wagegovt, govt, and capital1 exogenous. Data for the U.S. economy on these variables are taken from Klein (1950). This model can be fit with reg3 by typing

```
. use https://www.stata-press.com/data/r18/klein
. reg3 (consump wagepriv wagegovt) (wagepriv consump govt capital1)
Three-stage least-squares regression
```

Equation	Obs	Params	RMSE	"R-squared"	chi2	P>chi2
consump	22	2	1.776297	0.9388	208.02	0.0000
wagepriv	22	3	2.372443	0.8542	80.04	0.0000

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
consump						
wagepriv	.8012754	.1279329	6.26	0.000	.5505314	1.052019
wagegovt	1.029531	.3048424	3.38	0.001	.432051	1.627011
_cons	19.3559	3.583772	5.40	0.000	12.33184	26.37996
wagepriv						
consump	.4026076	.2567312	1.57	0.117	-.1005764	.9057916
govt	1.177792	.5421253	2.17	0.030	.1152461	2.240338
capital1	-.0281145	.0572111	-0.49	0.623	-.1402462	.0840173
_cons	14.63026	10.26693	1.42	0.154	-5.492552	34.75306

```
Endogenous: consump wagepriv
Exogenous:  wagegovt govt capital1
```

Without showing the 2SLS results, we note that the consumption function in this system falls under the conditions noted earlier. That is, the 2SLS and 3SLS coefficients for the equation are identical.

◁

▷ Example 2

Some of the most common simultaneous systems encountered are supply-and-demand models. A simple system could be specified as

$$\texttt{qDemand} = \beta_0 + \beta_1\,\texttt{price} + \beta_2\,\texttt{pcompete} + \beta_3\,\texttt{income} + \epsilon_1$$

$$\texttt{qSupply} = \beta_4 + \beta_5\,\texttt{price} + \beta_6\,\texttt{praw} + \epsilon_2$$

$$\texttt{Equilibrium condition: quantity} = \texttt{qDemand} = \texttt{qSupply}$$

where

quantity is the quantity of a product produced and sold,
price is the price of the product,
pcompete is the price of a competing product,
income is the average income level of consumers, and
praw is the price of raw materials used to produce the product.

In this system, price is assumed to be determined simultaneously with demand. The important statistical implications are that price is not a predetermined variable and that it is correlated with the disturbances of both equations. The system is somewhat unusual: quantity is associated with two disturbances. This fact really poses no problem because the disturbances are specified on the behavioral demand and supply equations—two separate entities. Often one of the two equations is rewritten to place price on the left-hand side, making this endogeneity explicit in the specification.

To provide a concrete illustration of the effects of simultaneous equations, we can simulate data for the above system by using known coefficients and disturbance properties. Specifically, we will simulate the data as

$$\texttt{qDemand} = 40 - 1.0\,\texttt{price} + 0.25\,\texttt{pcompete} + 0.5\,\texttt{income} + \epsilon_1$$

$$\texttt{qSupply} = 0.5\,\texttt{price} - 0.75\,\texttt{praw} + \epsilon_2$$

where

$$\epsilon_1 \sim N(0, 3.8)$$

$$\epsilon_2 \sim N(0, 2.4)$$

For comparison, we can estimate the supply and demand equations separately by OLS. The estimates for the demand equation are

```
. use https://www.stata-press.com/data/r18/supDem
. regress quantity price pcompete income
```

Source	SS	df	MS	Number of obs	=	49
				F(3, 45)	=	1.00
Model	23.1579302	3	7.71931008	Prob > F	=	0.4004
Residual	346.459313	45	7.69909584	R-squared	=	0.0627
				Adj R-squared	=	0.0002
Total	369.617243	48	7.70035923	Root MSE	=	2.7747

quantity	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
price	.1186265	.1716014	0.69	0.493	-.2269965	.4642496
pcompete	.0946416	.1200815	0.79	0.435	-.1472149	.3364981
income	.0785339	.1159867	0.68	0.502	-.1550754	.3121432
_cons	7.563261	5.019479	1.51	0.139	-2.54649	17.67301

The OLS estimates for the supply equation are

```
. regress quantity price praw
```

Source	SS	df	MS	Number of obs	=	49
				F(2, 46)	=	35.71
Model	224.819549	2	112.409774	Prob > F	=	0.0000
Residual	144.797694	46	3.14777596	R-squared	=	0.6082
				Adj R-squared	=	0.5912
Total	369.617243	48	7.70035923	Root MSE	=	1.7742

quantity	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
price	.724675	.1095657	6.61	0.000	.5041307	.9452192
praw	-.8674796	.1066114	-8.14	0.000	-1.082077	-.652882
_cons	-6.97291	3.323105	-2.10	0.041	-13.66197	-.283847

Examining the coefficients from these regressions, we note that they are not close to the known parameters used to generate the simulated data. In particular, the positive coefficient on `price` in the demand equation stands out. We constructed our simulated data to be consistent with economic theory—people demand less of a product if its price rises and more if their personal income rises. Although the `price` coefficient is statistically insignificant, the positive value contrasts starkly with what is predicted from economic price theory and the -1.0 value that we used in the simulation. Likewise, we are disappointed with the insignificance and level of the coefficient on average `income`. The supply equation has correct signs on the two main parameters, but their levels are different from the known values. In fact, the coefficient on `price` (0.724675) is different from the simulated parameter (0.5) at the 5% level of significance.

All of these problems are to be expected. We explicitly constructed a simultaneous system of equations that violated one of the assumptions of least squares. Specifically, the disturbances were correlated with one of the regressors—price.

Two-stage least squares can be used to address the correlation between regressors and disturbances. Using instruments for the endogenous variable, price, 2SLS will produce consistent estimates of the parameters in the system. Let's use ivregress (see [R] **ivregress**) to see how our simulated system behaves when fit using 2SLS.

```
. ivregress 2sls quantity (price = praw) pcompete income
Instrumental variables 2SLS regression          Number of obs   =         49
                                                Wald chi2(3)    =       8.77
                                                Prob > chi2     =     0.0326
                                                R-squared       =          .
                                                Root MSE        =     3.7333

------------------------------------------------------------------------------
    quantity | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       price |  -1.015817   .374209    -2.71   0.007    -1.749253    -.282381
    pcompete |   .3319504   .172912     1.92   0.055    -.0069508    .6708517
      income |   .5090607   .1919482    2.65   0.008     .1328491    .8852723
       _cons |   39.89988   10.77378    3.70   0.000     18.78366    61.01611
------------------------------------------------------------------------------
Endogenous: price
Exogenous:  pcompete income praw
. ivregress 2sls quantity (price = pcompete income) praw
Instrumental variables 2SLS regression          Number of obs   =         49
                                                Wald chi2(2)    =      39.25
                                                Prob > chi2     =     0.0000
                                                R-squared       =     0.5928
                                                Root MSE        =     1.7525

------------------------------------------------------------------------------
    quantity | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       price |   .5773133   .1749974     3.30   0.001     .2343247    .9203019
        praw |  -.7835496   .1312414    -5.97   0.000    -1.040778   -.5263213
       _cons |  -2.550694   5.273067    -0.48   0.629    -12.88571    7.784327
------------------------------------------------------------------------------
Endogenous: price
Exogenous:  praw pcompete income
```

We are now much happier with the estimation results. All the coefficients from both equations are close to the true parameter values for the system. In particular, the coefficients are all well within 95% confidence intervals for the parameters. The missing R^2 in the demand equation seems unusual; we will discuss that more later.

Finally, this system could be estimated using 3SLS. To demonstrate how large systems might be handled and to avoid multiline commands, we will use global macros (see [P] **macro**) to hold the specifications for our equations.

```
. global demand "(qDemand: quantity price pcompete income)"
. global supply "(qSupply: quantity price praw)"
. reg3 $demand $supply, endog(price)
```

We must specify price as endogenous because it does not appear as a dependent variable in either equation. Without this option, reg3 would assume that there are no endogenous variables in the system and produce seemingly unrelated regression (sureg) estimates. The reg3 output from our series of commands is

```
Three-stage least-squares regression
```

Equation	Obs	Params	RMSE	"R-squared"	chi2	P>chi2
qDemand	49	3	3.739686	-0.8540	8.68	0.0338
qSupply	49	2	1.752501	0.5928	39.25	0.0000

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
qDemand						
price	-1.014345	.3742036	-2.71	0.007	-1.74777	-.2809194
pcompete	.2647206	.1464194	1.81	0.071	-.0222561	.5516973
income	.5299146	.1898161	2.79	0.005	.1578819	.9019472
_cons	40.08749	10.77072	3.72	0.000	18.97726	61.19772
qSupply						
price	.5773133	.1749974	3.30	0.001	.2343247	.9203019
praw	-.7835496	.1312414	-5.97	0.000	-1.040778	-.5263213
_cons	-2.550694	5.273067	-0.48	0.629	-12.88571	7.784327

```
Endogenous: quantity price
Exogenous:  pcompete income praw
```

The use of 3SLS over 2SLS is essentially an efficiency issue. The coefficients of the demand equation from 3SLS are close to the coefficients from two-stage least squares, and those of the supply equation are identical. The latter case was mentioned earlier for systems with some exactly identified equations. However, even for the demand equation, we do not expect the coefficients to change systematically. What we do expect from three-stage least squares are more precise estimates of the parameters given the validity of our specification and `reg3`'s use of the covariances among the disturbances.

Let's summarize the results. With OLS, we got obviously biased estimates of the parameters. No amount of data would have improved the OLS estimates—they are inconsistent in the face of the violated OLS assumptions. With 2SLS, we obtained consistent estimates of the parameters, and these would have improved with more data. With 3SLS, we obtained consistent estimates of the parameters that are more efficient than those obtained by 2SLS.

◁

❑ Technical note

We noted earlier that the R^2 was missing from the two-stage estimates of the demand equation. Now we see that the R^2 is negative for the three-stage estimates of the same equation. How can we have a negative R^2?

In most estimators, other than least squares, the R^2 is no more than a summary measure of the overall in-sample predictive power of the estimator. The computational formula for R^2 is $R^2 = 1 - \text{RSS}/\text{TSS}$, where RSS is the residual sum of squares (sum of squared residuals) and TSS is the total sum of squared deviations about the mean of the dependent variable. In a standard linear model with a constant, the model from which the TSS is computed is nested within the full model from which RSS is computed—they both have a constant term based on the same data. Thus, it must be that $\text{TSS} \geq \text{RSS}$ and R^2 is constrained between 0 and 1.

For 2SLS and 3SLS, some of the regressors enter the model as instruments when the parameters are estimated. However, because our goal is to fit the structural model, the actual values, not the instruments for the endogenous right-hand-side variables, are used to determine R^2. The model residuals are computed over a different set of regressors from those used to fit the model. The two-

or three-stage estimates are no longer nested within a constant-only model of the dependent variable, and the residual sum of squares is no longer constrained to be smaller than the total sum of squares.

A negative R^2 in 3SLS should be taken for exactly what it is—an indication that the structural model predicts the dependent variable worse than a constant-only model. Is this a problem? It depends on the application. Three-stage least squares applied to our contrived supply-and-demand example produced good estimates of the known true parameters. Still, the demand equation produced an R^2 of -0.854. How do we feel about our parameter estimates? This should be determined by the estimates themselves, their associated standard errors, and the overall model significance. On this basis, negative R^2 and all, we feel pretty good about all the parameter estimates for both the supply and demand equations. Would we want to make predictions about equilibrium quantity by using the demand equation alone? Probably not. Would we want to make these quantity predictions by using the supply equation? Possibly, because based on in-sample predictions, they seem better than those from the demand equations. However, both the supply and demand estimates are based on limited information. If we are interested in predicting quantity, a reduced-form equation containing all our independent variables would usually be preferred.

❑

❑ Technical note

As a matter of syntax, we could have specified the supply-and-demand model on one line without using global macros.

```
. reg3 (quantity price pcompete income) (quantity price praw), endog(price)
Three-stage least-squares regression
```

Equation	Obs	Params	RMSE	"R-squared"	chi2	P>chi2
quantity	49	3	3.739686	-0.8540	8.68	0.0338
2quantity	49	2	1.752501	0.5928	39.25	0.0000

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
quantity						
price	-1.014345	.3742036	-2.71	0.007	-1.74777	-.2809194
pcompete	.2647206	.1464194	1.81	0.071	-.0222561	.5516973
income	.5299146	.1898161	2.79	0.005	.1578819	.9019472
_cons	40.08749	10.77072	3.72	0.000	18.97726	61.19772
2quantity						
price	.5773133	.1749974	3.30	0.001	.2343247	.9203019
praw	-.7835496	.1312414	-5.97	0.000	-1.040778	-.5263213
_cons	-2.550694	5.273067	-0.48	0.629	-12.88571	7.784327

```
Endogenous: quantity price
Exogenous:  pcompete income praw
```

However, here `reg3` has been forced to create a unique equation name for the supply equation—`2quantity`. Both the supply and demand equations could not be designated as `quantity`, so a number was prefixed to the name for the supply equation.

We could have specified

```
. reg3 (qDemand: quantity price pcompete income) (qSupply: quantity price praw),
> endog(price)
```

and obtained the same results and equation labeling as when we used global macros to hold the equation specifications.

Without explicit equation names, `reg3` always assumes that the dependent variable should be used to name equations. When each equation has a different dependent variable, this rule causes no problems and produces easily interpreted result tables. If the same dependent variable appears in more than one equation, however, `reg3` will create a unique equation name based on the dependent variable name. Because equation names must be used for cross-equation tests, you have more control in this situation if explicit names are placed on the equations.

❑

▷ Example 3: Using the full syntax of reg3

Klein's (1950) model of the U.S. economy is often used to demonstrate system estimators. It contains several common features that will serve to demonstrate the full syntax of `reg3`. The Klein model is defined by the following seven relationships:

$$\mathtt{c} = \beta_0 + \beta_1\mathtt{p} + \beta_2\mathtt{L.p} + \beta_3\mathtt{w} + \epsilon_1 \quad (1)$$
$$\mathtt{i} = \beta_4 + \beta_5\mathtt{p} + \beta_6\mathtt{L.p} + \beta_7\mathtt{L.k} + \epsilon_2 \quad (2)$$
$$\mathtt{wp} = \beta_8 + \beta_9\mathtt{y} + \beta_{10}\mathtt{L.y} + \beta_{11}\mathtt{yr} + \epsilon_3 \quad (3)$$
$$\mathtt{y} = \mathtt{c} + \mathtt{i} + \mathtt{g} \quad (4)$$
$$\mathtt{p} = \mathtt{y} - \mathtt{t} - \mathtt{wp} \quad (5)$$
$$\mathtt{k} = \mathtt{L.k} + \mathtt{i} \quad (6)$$
$$\mathtt{w} = \mathtt{wg} + \mathtt{wp} \quad (7)$$

Here we have used Stata's lag operator `L.` to represent variables that appear with a one-period lag in our model; see **[U] 13.10 Time-series operators**.

The variables in the model are listed below. Two sets of variable names are shown. The concise first name uses traditional economics mnemonics, whereas the second name provides more guidance for everyone else. The concise names serve to keep the specification of the model small (and quite understandable to economists).

Short name	Long name	Variable definition	Type
c	consump	Consumption	endogenous
p	profits	Private industry profits	endogenous
wp	wagepriv	Private wage bill	endogenous
wg	wagegovt	Government wage bill	exogenous
w	wagetot	Total wage bill	endogenous
i	invest	Investment	endogenous
k	capital	Capital stock	endogenous
y	totinc	Total income/demand	endogenous
g	govt	Government spending	exogenous
t	taxnetx	Indirect bus. taxes + net exports	exogenous
yr	year	Year—1931	exogenous

Equations (1)–(3) are behavioral and contain explicit disturbances (ϵ_1, ϵ_2, and ϵ_3). The remaining equations are identities that specify additional variables in the system and their accounting relationships with the variables in the behavioral equations. Some variables are explicitly endogenous by appearing as dependent variables in (1)–(3). Others are implicitly endogenous as linear combinations that contain other endogenous variables (for example, `w` and `p`). Still other variables are implicitly exogenous by appearing in the identities but not in the behavioral equations (for example, `wg` and `g`).

Using the concise names, we can fit Klein's model with the following command:

```
. use https://www.stata-press.com/data/r18/klein2
. reg3 (c p L.p w) (i p L.p L.k) (wp y L.y yr), endog(w p y) exog(t wg g)
Three-stage least-squares regression
```

Equation	Obs	Params	RMSE	"R-squared"	chi2	P>chi2
c	21	3	.9443305	0.9801	864.59	0.0000
i	21	3	1.446736	0.8258	162.98	0.0000
wp	21	3	.7211282	0.9863	1594.75	0.0000

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
c						
p						
--.	.1248904	.1081291	1.16	0.248	-.0870387	.3368194
L1.	.1631439	.1004382	1.62	0.104	-.0337113	.3599992
w	.790081	.0379379	20.83	0.000	.715724	.8644379
_cons	16.44079	1.304549	12.60	0.000	13.88392	18.99766
i						
p						
--.	-.0130791	.1618962	-0.08	0.936	-.3303898	.3042316
L1.	.7557238	.1529331	4.94	0.000	.4559805	1.055467
k						
L1.	-.1948482	.0325307	-5.99	0.000	-.2586072	-.1310893
_cons	28.17785	6.793768	4.15	0.000	14.86231	41.49339
wp						
y						
--.	.4004919	.0318134	12.59	0.000	.3381388	.462845
L1.	.181291	.0341588	5.31	0.000	.1143411	.2482409
yr	.149674	.0279352	5.36	0.000	.094922	.2044261
_cons	1.797216	1.115854	1.61	0.107	-.3898181	3.984251

```
Endogenous: c i wp w p y
Exogenous:  L.p L.k L.y yr t wg g
```

We used the `exog()` option to identify `t`, `wg`, and `g` as exogenous variables in the system. These variables must be identified because they are part of the system but appear directly in none of the behavioral equations. Without this option, `reg3` would not know they were part of the system and could be used as instrumental variables. The `endog()` option specifying `w`, `p`, and `y` is also required. Without this information, `reg3` would be unaware that these variables are linear combinations that include endogenous variables. We did not include `k` in the `endog()` option because only its lagged value appears in the behavioral equations.

❑ Technical note

Rather than listing additional endogenous and exogenous variables, we could specify the full list of exogenous variables in an `inst()` option,

```
. reg3 (c p L.p w) (i p L.p L.k) (wp y L.y yr), inst(g t wg yr L.p L.k L.y)
```

or equivalently,

```
. global conseqn "(c p L.p w)"
. global inveqn  "(i p L.p L.k)"
. global wageqn  "(wp y L.y yr)"
. global inlist  "g t wg yr L.p L.k L.y"
. reg3 $conseqn $inveqn $wageqn, inst($inlist)
```

Macros and explicit equations can also be mixed in the specification

```
. reg3 $conseqn (i p L.p L.k) $wageqn, endog(w p y) exog(t wg g)
```

or

```
. reg3 (c p L.p w) $inveqn (wp y L.y yr), endog(w p y) exog(t wg g)
```

Placing the equation-binding parentheses in the global macros was also arbitrary. We could have used

```
. global consump  "c p L.p w"
. global invest   "i p L.p L.k"
. global wagepriv "wp y L.y yr"
. reg3 ($consump) ($invest) ($wagepriv), endog(w p y) exog(t wg g)
```

`reg3` is tolerant of all combinations, and these commands will produce identical output.

❑

Switching to the full variable names, we can fit Klein's model with the commands below. We will use global macros to store the lists of endogenous and exogenous variables. Again, this is not necessary: these lists could have been typed directly on the command line. However, assigning the lists to local macros makes additional processing easier if alternative models are to be fit. We will also use the `ireg3` option to produce the iterated estimates.

```
. use https://www.stata-press.com/data/r18/kleinfull
. global conseqn "(consump profits L.profits wagetot)"
. global inveqn "(invest profits L.profits L.capital)"
. global wageqn "(wagepriv totinc L.totinc year)"
. global enlist "wagetot profits totinc"
. global exlist "taxnetx wagegovt govt"
. reg3 $conseqn $inveqn $wageqn, endog($enlist) exog($exlist) ireg3
Iteration 1:  Tolerance =   .3712549
Iteration 2:  Tolerance =   .1894712
Iteration 3:  Tolerance =   .1076401
  (output omitted)
Iteration 24: Tolerance =  7.049e-07
Three-stage least-squares regression, iterated
```

Equation	Obs	Params	RMSE	"R-squared"	chi2	P>chi2
consump	21	3	.9565088	0.9796	970.31	0.0000
invest	21	3	2.134327	0.6209	56.78	0.0000
wagepriv	21	3	.7782334	0.9840	1312.19	0.0000

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
consump						
profits						
--.	.1645096	.0961979	1.71	0.087	-.0240348	.3530539
L1.	.1765639	.0901001	1.96	0.050	-.0000291	.3531569
wagetot	.7658011	.0347599	22.03	0.000	.6976729	.8339294
_cons	16.55899	1.224401	13.52	0.000	14.15921	18.95877
invest						
profits						
--.	-.3565316	.2601568	-1.37	0.171	-.8664296	.1533664
L1.	1.011299	.2487745	4.07	0.000	.5237098	1.498888
capital						
L1.	-.2602	.0508694	-5.12	0.000	-.3599022	-.1604978
_cons	42.89629	10.59386	4.05	0.000	22.13271	63.65987
wagepriv						
totinc						
--.	.3747792	.0311027	12.05	0.000	.3138191	.4357394
L1.	.1936506	.0324018	5.98	0.000	.1301443	.257157
year	.1679262	.0289291	5.80	0.000	.1112263	.2246261
_cons	2.624766	1.195559	2.20	0.028	.2815124	4.968019

```
Endogenous: consump invest wagepriv wagetot profits totinc
Exogenous:  L.profits L.capital L.totinc year taxnetx wagegovt govt
```

◁

▷ Example 4: Constraints with reg3

As a simple example of constraints, (1) above may be rewritten with both wages explicitly appearing (rather than as a variable containing the sum). Using the longer variable names, we have

$$\texttt{consump} = \beta_0 + \beta_1\,\texttt{profits} + \beta_2\,\texttt{L.profits} + \beta_3\,\texttt{wagepriv} + \beta_{12}\,\texttt{wagegovt} + \epsilon_1$$

To retain the effect of the identity in (7), we need $\beta_3 = \beta_{12}$ as a constraint on the system. We obtain this result by defining the constraint in the usual way and then specifying its use in `reg3`. Because `reg3` is a system estimator, we will need to use the full equation syntax of `constraint`. The assumption that the following commands are entered after the model above has been estimated. We are simply changing the definition of the consumption equation (`consump`) and adding a constraint on two of its parameters. The rest of the model definition is carried forward.

```
. global conseqn "(consump profits L.profits wagepriv wagegovt)"
. constraint define 1 [consump]wagepriv = [consump]wagegovt
. reg3 $conseqn $inveqn $wageqn, endog($enlist) exog($exlist) constr(1) ireg3
note: additional endogenous variables not in the system have no effect and are
      ignored (wagetot).
Iteration 1:  Tolerance =   .3712547
Iteration 2:  Tolerance =    .189471
Iteration 3:  Tolerance =     .10764
  (output omitted)
Iteration 24: Tolerance =  7.049e-07
```

```
Three-stage least-squares regression, iterated
```

Equation	Obs	Params	RMSE	"R-squared"	chi2	P>chi2
consump	21	3	.9565086	0.9796	970.31	0.0000
invest	21	3	2.134326	0.6209	56.78	0.0000
wagepriv	21	3	.7782334	0.9840	1312.19	0.0000

```
 ( 1)  [consump]wagepriv - [consump]wagegovt = 0
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
consump						
profits						
--.	.1645097	.0961978	1.71	0.087	-.0240346	.353054
L1.	.1765639	.0901001	1.96	0.050	-.0000291	.3531568
wagepriv	.7658012	.0347599	22.03	0.000	.6976729	.8339294
wagegovt	.7658012	.0347599	22.03	0.000	.6976729	.8339294
_cons	16.55899	1.224401	13.52	0.000	14.1592	18.95877
invest						
profits						
--.	-.3565311	.2601567	-1.37	0.171	-.8664288	.1533666
L1.	1.011298	.2487744	4.07	0.000	.5237096	1.498887
capital						
L1.	-.2601999	.0508694	-5.12	0.000	-.359902	-.1604977
_cons	42.89626	10.59386	4.05	0.000	22.13269	63.65984
wagepriv						
totinc						
--.	.3747792	.0311027	12.05	0.000	.313819	.4357394
L1.	.1936506	.0324018	5.98	0.000	.1301443	.257157
year	.1679262	.0289291	5.80	0.000	.1112263	.2246261
_cons	2.624766	1.195559	2.20	0.028	.281512	4.968019

```
Endogenous: consump invest wagepriv wagetot profits totinc
Exogenous:  L.profits wagegovt L.capital L.totinc year taxnetx govt
```

As expected, none of the parameter or standard error estimates has changed from the previous estimates (before the seventh significant digit). We have simply decomposed the total wage variable into its two parts and constrained the coefficients on these parts. The warning about additional endogenous variables was just `reg3`'s way of letting us know that we had specified some information that was irrelevant to the estimation of the system. We had left the `wagetot` variable in our `endog` macro. It does not mean anything to the system to specify `wagetot` as endogenous because it is no longer in the system. That's fine with `reg3` and fine for our current purposes.

We can also impose constraints across the equations. For example, the admittedly meaningless constraint of requiring `profits` to have the same effect in both the consumption and investment equations could be imposed. Retaining the constraint on the wage coefficients, we would estimate this constrained system.

```
. constraint define 2 [consump]profits = [invest]profits

. reg3 $conseqn $inveqn $wageqn, endog($enlist) exog($exlist) constr(1 2) ireg3
note: additional endogenous variables not in the system have no effect and are
      ignored (wagetot).

Iteration 1:  Tolerance =   .1427927
Iteration 2:  Tolerance =    .032539
Iteration 3:  Tolerance =  .00307811
Iteration 4:  Tolerance =  .00016903
Iteration 5:  Tolerance =  .00003409
Iteration 6:  Tolerance =  7.763e-06
Iteration 7:  Tolerance =  9.240e-07

Three-stage least-squares regression, iterated
```

Equation	Obs	Params	RMSE	"R-squared"	chi2	P>chi2
consump	21	3	.9504669	0.9798	1019.54	0.0000
invest	21	3	1.247066	0.8706	144.57	0.0000
wagepriv	21	3	.7225276	0.9862	1537.45	0.0000

```
 ( 1)  [consump]wagepriv - [consump]wagegovt = 0
 ( 2)  [consump]profits - [invest]profits = 0
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
consump						
profits						
--.	.1075413	.0957767	1.12	0.262	-.0801777	.2952602
L1.	.1712756	.0912613	1.88	0.061	-.0075932	.3501444
wagepriv	.798484	.0340876	23.42	0.000	.7316734	.8652946
wagegovt	.798484	.0340876	23.42	0.000	.7316734	.8652946
_cons	16.2521	1.212157	13.41	0.000	13.87631	18.62788
invest						
profits						
--.	.1075413	.0957767	1.12	0.262	-.0801777	.2952602
L1.	.6443378	.1058682	6.09	0.000	.43684	.8518356
capital						
L1.	-.1766669	.0261889	-6.75	0.000	-.2279962	-.1253375
_cons	24.31931	5.284325	4.60	0.000	13.96222	34.6764
wagepriv						
totinc						
--.	.4014106	.0300552	13.36	0.000	.3425035	.4603177
L1.	.1775359	.0321583	5.52	0.000	.1145068	.240565
year	.1549211	.0282291	5.49	0.000	.099593	.2102492
_cons	1.959788	1.14467	1.71	0.087	-.2837242	4.203299

```
Endogenous: consump invest wagepriv wagetot profits totinc
Exogenous:  L.profits wagegovt L.capital L.totinc year taxnetx govt
```

◁

❑ Technical note

Identification in a system of simultaneous equations involves the notion that there is enough information to estimate the parameters of the model given the specified functional form. Underidentification usually manifests itself as one matrix in the 3SLS computations. The most commonly

violated order condition for 2SLS or 3SLS involves the number of endogenous and exogenous variables. There must be at least as many noncollinear exogenous variables in the remaining system as there are endogenous right-hand-side variables in an equation. This condition must hold for each structural equation in the system.

Put as a set of rules the following:

1. Count the number of right-hand-side endogenous variables in an equation and call this m_i.
2. Count the number of exogenous variables in the same equation and call this k_i.
3. Count the total number of exogenous variables in all the structural equations plus any additional variables specified in an `exog()` or `inst()` option and call this K.
4. If $m_i > (K - k_i)$ for any structural equation (i), then the system is underidentified and cannot be estimated by 3SLS.

We are also possibly in trouble if any of the exogenous variables are linearly dependent. We must have m_i linearly independent variables among the exogenous variables represented by $(K - k_i)$.

The complete conditions for identification involve rank-order conditions on several matrices. For a full treatment, see Theil (1971) or Greene (2018, 363–365). ❑

Henri Theil (1924–2000) was born in Amsterdam and awarded a PhD in 1951 by the University of Amsterdam. He researched and taught econometric theory, statistics, microeconomics, macroeconomic modeling, and economic forecasting, and policy at (now) Erasmus University Rotterdam, the University of Chicago, and the University of Florida. Theil's many specific contributions include work on 2SLS and 3SLS, inequality and concentration, and consumer demand.

Stored results

`reg3` stores the following in `e()`:

Scalars

`e(N)`	number of observations
`e(k)`	number of parameters
`e(k_eq)`	number of equations in `e(b)`
`e(mss_#)`	model sum of squares for equation #
`e(df_m#)`	model degrees of freedom for equation #
`e(rss_#)`	residual sum of squares for equation #
`e(df_r)`	residual degrees of freedom (`small`)
`e(r2_#)`	R^2 for equation #
`e(F_#)`	F statistic for equation # (`small`)
`e(rmse_#)`	root mean squared error for equation #
`e(dfk2_adj)`	divisor used with VCE when `dfk2` specified
`e(ll)`	log likelihood
`e(N_clust)`	number of clusters
`e(chi2_#)`	χ^2 for equation #
`e(p_#)`	p-value for model test for equation #
`e(cons_#)`	1 when equation # has a constant, 0 otherwise
`e(rank)`	rank of `e(V)`
`e(ic)`	number of iterations

Macros

e(cmd)	reg3
e(cmdline)	command as typed
e(depvar)	names of dependent variables
e(exog)	names of exogenous variables
e(endog)	names of endogenous variables
e(eqnames)	names of equations
e(corr)	correlation structure
e(wtype)	weight type
e(wexp)	weight expression
e(method)	3sls, 2sls, ols, sure, or mvreg
e(small)	small, if specified
e(dfk)	dfk, if specified
e(clustvar)	name of cluster variable
e(vce)	*vcetype* specified in vce()
e(vcetype)	title used to label Std. err.
e(properties)	b V
e(predict)	program used to implement predict
e(marginsok)	predictions allowed by margins
e(marginsnotok)	predictions disallowed by margins
e(marginsdefault)	default predict() specification for margins
e(asbalanced)	factor variables fvset as asbalanced
e(asobserved)	factor variables fvset as asobserved

Matrices

e(b)	coefficient vector
e(Cns)	constraints matrix
e(Sigma)	$\widehat{\boldsymbol{\Sigma}}$ matrix
e(V)	variance–covariance matrix of the estimators

Functions

e(sample)	marks estimation sample

In addition to the above, the following is stored in r():

Matrices

r(table)	matrix containing the coefficients with their standard errors, test statistics, p-values, and confidence intervals

Note that results stored in r() are updated when the command is replayed and will be replaced when any r-class command is run after the estimation command.

Methods and formulas

The most concise way to represent a system of equations for 3SLS requires thinking of the individual equations and their associated data as being stacked. reg3 does not expect the data in this format, but it is a convenient shorthand. The system could then be formulated as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_M \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1 & 0 & \dots & 0 \\ 0 & \mathbf{Z}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{Z}_M \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_M \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_M \end{bmatrix}$$

In full matrix notation, this is just

$$\mathbf{y} = \mathbf{Z}\,\mathbf{B} + \boldsymbol{\epsilon}$$

The $\mathbf{Z}$ elements in these matrices represent both the endogenous and the exogenous right-hand-side variables in the equations.

Also assume that there will be correlation between the disturbances of the equations so that

$$E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \boldsymbol{\Sigma}$$

where the disturbances are further assumed to have an expected value of 0; $E(\boldsymbol{\epsilon}) = 0$.

The first stage of 3SLS regression requires developing instrumented values for the endogenous variables in the system. These values can be derived as the predictions from a linear regression of each endogenous regressor on all exogenous variables in the system or, more succinctly, as the projection of each regressor through the projection matrix of all exogenous variables onto the regressors. Designating the set of all exogenous variables as $\mathbf{X}$ results in

$$\widehat{\mathbf{z}}_i = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}_i \quad \text{for each } i$$

Taken collectively, these $\widehat{\mathbf{Z}}$ contain the instrumented values for all the regressors. They take on the actual values for the exogenous variables and first-stage predictions for the endogenous variables. Given these instrumented variables, a generalized least squares (GLS) or Aitken (1935) estimator can be formed for the parameters of the system

$$\widehat{\mathbf{B}} = \left\{\widehat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\widehat{\mathbf{Z}}\right\}^{-1} \widehat{\mathbf{Z}}'(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})\mathbf{y}$$

All that remains is to obtain a consistent estimator for $\boldsymbol{\Sigma}$. This estimate can be formed from the residuals of 2SLS estimates of each equation in the system. Alternately, and identically, the residuals can be computed from the estimates formed by taking $\boldsymbol{\Sigma}$ to be an identity matrix. This maintains the full system of coefficients and allows constraints to be applied when the residuals are computed.

If we take $\mathbf{E}$ to be the matrix of residuals from these estimates, a consistent estimate of $\boldsymbol{\Sigma}$ is

$$\widehat{\boldsymbol{\Sigma}} = \frac{\mathbf{E}'\mathbf{E}}{n}$$

where n is the number of observations in the sample. An alternative divisor for this estimate can be obtained with the `dfk` option as outlined under options.

With the estimate of $\widehat{\boldsymbol{\Sigma}}$ placed into the GLS estimating equation,

$$\widehat{\mathbf{B}} = \left\{\widehat{\mathbf{Z}}'(\widehat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I})\widehat{\mathbf{Z}}\right\}^{-1} \widehat{\mathbf{Z}}'(\widehat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I})\mathbf{y}$$

is the 3SLS estimates of the system parameters.

The asymptotic variance–covariance matrix of the estimator is just the standard formulation for a GLS estimator

$$\mathbf{V}_{\widehat{\mathbf{B}}} = \left\{\widehat{\mathbf{Z}}'(\widehat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I})\widehat{\mathbf{Z}}\right\}^{-1}$$

Iterated 3SLS estimates can be obtained by computing the residuals from the three-stage parameter estimates, using these to formulate a new $\widehat{\boldsymbol{\Sigma}}$, and recomputing the parameter estimates. This process is repeated until the estimates $\widehat{\mathbf{B}}$ converge—if they converge. Convergence is not guaranteed. When estimating a system by SURE, these iterated estimates will be the maximum likelihood estimates for the system. The iterated solution can also be used to produce estimates that are invariant to choice of system and restriction parameterization for many linear systems under full 3SLS.

The exposition above follows the parallel developments in Greene (2018) and Davidson and MacKinnon (1993).

This command supports the Huber/White/sandwich estimator of the variance and its clustered version using `vce(robust)` and `vce(cluster` *clustvar*`)`, respectively. See [P] **_robust**, particularly *Introduction* and *Methods and formulas*.

Alexander Craig Aitken (1895–1967) was born in Dunedin, New Zealand. He attended the University of Otago on a full scholarship but left the university to enlist in the New Zealand Expeditionary Force during World War I. He would later reflect on the time he spent in active service and publish two war memoirs.

After returning from the war, he completed his studies at the University of Otago and then worked on his PhD at the University of Edinburgh under E. T. Whittaker. Although he suffered weeks of illness during the course of his postgraduate studies, he managed to write an impressive thesis on data smoothing. He then became a professor of mathematics, and later the Chair of Pure Mathematics, a post he would hold until his retirement in 1965.

Among his many accolades, he was elected a Fellow of the Royal Society of Edinburgh and a Fellow of the Royal Society of Literature, following the publication of his second war memoir. He published many papers on numerical analysis and statistics as well as a couple of books on matrices. Aitken is credited with deriving the generalized least squares estimator, which has been referred to as Aitken's generalized least squares. Aside from his many academic contributions, he had mental calculating abilities like no other. He is known to have multiplied two 9-digit numbers in less than a minute and could recite up to 707 digits of π.

References

Aitken, A. C. 1935. On least squares and linear combination of observations. *Proceedings of the Royal Society of Edinburgh* 55: 42–48. https://doi.org/10.1017/S0370164600014346.

Bewley, R. 2000. Mr. Henri Theil: An interview with the International Journal of Forecasting. *International Journal of Forecasting* 16: 1–16. https://doi.org/10.1016/S0169-2070(99)00031-X.

Davidson, R., and J. G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.

Greene, W. H. 2018. *Econometric Analysis*. 8th ed. New York: Pearson.

Klein, L. R. 1950. *Economic Fluctuations in the United States 1921–1941*. New York: Wiley.

Nichols, A. 2007. Causal inference with observational data. *Stata Journal* 7: 507–541.

Poi, B. P. 2006. Jackknife instrumental variables estimation in Stata. *Stata Journal* 6: 364–376.

Theil, H. 1971. *Principles of Econometrics*. New York: Wiley.

Zellner, A., and H. Theil. 1962. Three stage least squares: Simultaneous estimate of simultaneous equations. *Econometrica* 29: 54–78. https://doi.org/10.2307/1911287.

Also see

[R] **reg3 postestimation** — Postestimation tools for reg3

[R] **ivregress** — Single-equation instrumental-variables regression

[R] **nlsur** — Estimation of nonlinear systems of equations

[R] **regress** — Linear regression

[R] **sureg** — Zellner's seemingly unrelated regression

[MV] **mvreg** — Multivariate regression

[SEM] **Example 7** — Nonrecursive structural model

[SEM] **Intro 5** — Tour of models

[TS] **forecast** — Econometric model forecasting

[U] **20 Estimation and postestimation commands**