hausman_taylor version 1.0

Allin Cottrell

December 18, 2025

1 Introduction

This package estimates a panel-data model using the method of Hausman and Taylor (1981). As is well known,

- the standard fixed-effects estimator cannot handle time-invariant variables (since nothing remains of such variables after sweeping out the individual means), while
- the random-effects estimator cannot handle (on pain of inconsistency) regressors that are correlated with the unobserved individual effects.[1]

The Hausman–Taylor estimator bridges this gap: it permits estimation of a model that includes both time-invariant terms and regressors that are correlated with the individual effects. The "price of admission" (more formally, the condition for identification) is that there must be at least as many time-varying exogenous regressors—"exogenous" in the sense of being uncorrelated with the individual effects—as there are time-invariant regressors that are suspected of endogeneity, that is, of being correlated with the individual effects.

2 The model

Let $i = 1, \ldots, N$ index individuals and $t = 1, \ldots, T$ index time. The model is

$$y_{it} = \beta_0 + x'_{1it}\beta_1 + x'_{2it}\beta_2 + z'_{1i}\gamma_1 + z'_{2i}\gamma_2 + u_i + \epsilon_{it} \tag{1}$$

where x_1 and x_2 are time-varying and z_1 and z_2 are time-invariant. The variables x_1 and z_1 are exogenous (uncorrelated with the individual effects, u_i) while x_2 and z_2 are assumed to be so correlated. All of the regressors are assumed to be uncorrelated with ϵ_{it}.

In general, x_{1it}, x_{2it}, z_{1i} and z_{2i} are vectors of length k_1, k_2, g_1 and g_2, respectively, subject to the identification requirement $k_1 \geq g_2$.

The algorithm for the Hausman–Taylor estimator—for a balanced panel in which the time-series length, T, is the same for all individuals—is commonly given as follows:

1. Regress $\tilde{y} = (y_{it} - \bar{y}_i)$ on $\tilde{x}_1 = (x_{1it} - \bar{x}_{1i})$ and $\tilde{x}_2 = (x_{2it} - \bar{x}_{2i})$ to obtain initial estimates of β_1 and β_2. Use the residuals from this fixed-effects regression, e_{it}, to estimate the "within" error variance σ^2_ϵ.
2. Perform an IV regression of the stacked individual means of e_{it} on z_1 and z_2, using as instruments z_1 and x_1. Use the residual variance from this regression, s^2_2, to estimate σ^2_u as $s^2_2 - \hat{\sigma}^2_\epsilon/T$, and calculate the GLS coefficient

$$\theta = 1 - \left(\frac{\hat{\sigma}^2_\epsilon}{\hat{\sigma}^2_\epsilon + T\hat{\sigma}^2_u} \right)^{0.5} \tag{2}$$

[1] See the chapter titled "Panel data" in the *Gretl User's Guide* for an extended discussion of these points.

3. Let $w_{it} \equiv (x_{1it}, x_{2it}, z_{1i}, z_{2i})$. Run an IV regression of $y^*_{it} = (y_{it} - \theta\bar{y}_i)$ on $w^*_{it} = (w_{it} - \theta\bar{w}_i)$, using as instruments $\tilde{x}_1$, $\tilde{x}_2$, $\bar{x}_1$ and z_1.

The final step can also be described thus: regress the quasi-demeaned dependent variable on the quasi-demeaned regressors, taking as instruments the fully-demeaned time-varying regressors, the individual means of the exogenous time-varying terms, and the levels of the exogenous time-invariant terms. As Hausman and Taylor (1981, p. 1393) remark, "Making use of time-varying variables in two ways—to estimate their own coefficients and to serve as instruments for endogenous time-invariant variables—allows identification and efficient estimation of both β and γ."

In an unbalanced panel the time-series length, T_i, differs across individuals. In that case steps 2 and 3 above have to be modified slightly. First, the calculation of $\hat{\sigma}^2_u$ uses the harmonic mean of the T_is in place of a common T. Second, the value of θ differs across individuals:

$$\theta_i = 1 - \left(\frac{\hat{\sigma}^2_\epsilon}{\hat{\sigma}^2_\epsilon + T_i \hat{\sigma}^2_u} \right)^{0.5} \tag{3}$$

Section 7 below takes up a further issue pertaining to the unbalanced case.

3 The hausman_taylor function

The signature of this function is

```
bundle hausman_taylor (series y "dependent variable",
                       list Lexo "exogenous regressors",
                       list Lndo "endogenous regressors",
                       int verbosity[0:2:1],
                       bool as_stata[0])
```

The series `y` is the dependent variable, and the lists `Lexo` and `Lndo` correspond, respectively, to x_1 plus z_1 and x_2 plus z_2 in equation (1). This function does not undertake to judge which regressors are exogenous and which endogenous—you must decide the partition between `Lexo` and `Lndo`—but it can easily determine which regressors are time-varying and which invariant.

The `verbosity` parameter accepts values 0, 1 or 2: a value of 0 means that nothing is printed; 1 means that the Hausman–Taylor estimates are printed; 2 means that in addition the results of the preliminary regressions (described in Section 2) are printed. The default value is 1.

The `as_stata` boolean can be used to produce results comparable with Stata's `xthtaylor` command; an explanation of this option is given on page 7.

This function returns a bundle containing the items shown in Table 1. Most of the contents should be fairly self-explanatory, but the following comments may be useful.

- In the case of an unbalanced panel, the GLS coefficients θ_i will differ across individuals; the `theta` value is then the mean of the θ_is.
- The `Wald`, `Htest` and `Stest` matrices, if present, are each row vectors containing test statistic, degrees of freedom and P-value pertaining to the Wald, Hausman and Sargan tests, respectively. The Wald test uses the coefficient vector and covariance matrix to test the null hypothesis that only the constant truly has a non-zero coefficient. The Hausman and Sargan tests are available only if the specification is overidentified ($k_1 > g_2$). They both test the null hypothesis of correct specification. The Hausman test is based on a vector of contrasts, the difference between the fixed-effects and Hausman–Taylor estimates of the coefficients on the time-varying regressors. The Sargan test is based on the explained sum of squares from a regression of the Hausman–Taylor residuals on all of the instruments. Small P-values on these tests cast doubt on the consistency of the estimator.

name	*type*	*description*
depvar	string	name of the dependent variable
parnames	strings	names of regressors
Lexo	list	exogenous regressors
Lndo	list	endogenous regressors
coeff	matrix	regression coefficients
stderr	matrix	standard errors
vcv	matrix	variance–covariance matrix
s_e	scalar	square root of "within" variance, $\hat{\sigma}_\epsilon^2$
s_u	scalar	square root of $\hat{\sigma}_u^2$
ncoeff	scalar	total number of coefficients
nobs	scalar	total number of observations used
effn	scalar	number of units included
Tmin	scalar	minimum T_i value
Tmax	scalar	maximum T_i value
theta	scalar	GLS coefficient, θ
Wald	matrix	Wald test results
Htest	matrix	Hausman test results
Stest	matrix	Sargan test results
yhat	series	fitted values
uhat	series	residuals
rsq	scalar	$\text{corr}(y, \hat{y})^2$

Table 1: Items in hausman_taylor bundle

4 Sample script

The sample script for this package uses data on (log) wages and several covariates for 595 individuals observed annually from 1976 to 1982, taken from the US Panel Study of Income Dynamics. These data were originally employed by Cornwell and Rupert (1988) to assess various instrumental-variable estimators for panel data including Hausman–Taylor. They were revisited by Baltagi and Khanti-Akom (1990) and in chapter 7 of Baltagi (2005). The script replicates both sets of estimates; partial output is shown in Listing 4.

In each case we may take it that the endogenous regressor of primary interest is `ed` (education level). The specifications differ in their treatment of two pairs of regressors: in Cornwell and Rupert `wks` (weeks worked) and `ms` (marital status) are taken to be exogenous while `occ` (blue-collar dummy) and `ind` (manufacturing dummy) are endogenous; Baltagi reverses this, assuming that `wks` and `ms` are endogenous, `occ` and and `ind` exogenous. Baltagi finds a slightly smaller, but more sharply estimated, return to education. The Hausman and Sargan specification tests favor Baltagi's specification.

5 Graphical interface

Assuming you have said OK to this feature when installing the package, an entry-point for hausman_taylor can be found under the Panel sub-menu of gretl's Model menu: the label is Hausman-Taylor. The dialog that appears is shown in Figure 1.

6 Ancillary printing function

The ancillary public function `ht_print()` is provided to "pretty-print" the results contained in the bundle provided by `hausman_taylor()`; `ht_print()` takes a pointer to the bundle as its sole argument.

7 More on the unbalanced case

A noteworthy aspect of the Hausman–Taylor estimator is the treatment of x_1^*—that is, quasi-demeaned x_1—in the final IV regression. If the panel is balanced x_1^* is effectively treated as exogenous. It does not appear explicitly among the instruments, but we have the exact linear relationship

$$x_{1it}^* \equiv (x_{1it} - \theta\bar{x}_{1i}) = (x_{1it} - \bar{x}_{1i}) + (1-\theta)\bar{x}_{1i} = \tilde{x}_{1it} + (1-\theta)\bar{x}_{1i}$$

so that x_1^* is "perfectly instrumented" by $\tilde{x}_1$ and $\bar{x}_1$. This is as it should be. By assumption x_{1it} is independent of u_i, and therefore so is $\bar{x}_{1i}$. The transformation $x_{1it}^* = x_{1it} - \theta\bar{x}_{1i}$ clearly does not introduce any dependence on u_i, so x_1^* ought to be treated as exogenous. It is not included as an instrument simply because it would be redundant, given the point made above.

Now consider the unbalanced case. It is standard to calculate σ_u^2 as $s_2^2 - \sigma_\epsilon^2/\bar{T}$, where $\bar{T}$ is the harmonic mean of the T_is. And θ varies by individual according to (3) above. This means there is no longer an exact linear relationship between x_1^* and the instruments $\tilde{x}_1$ and $\bar{x}_1$, which raises the question, should x_1^* be added to the set of instruments in the final IV step of Hausman–Taylor?

The alternative—including $\tilde{x}_1$ and $\bar{x}_1$ as instruments, but not x_1^*—amounts to treating x_1^* as endogenous, but there is no reason for this. When the panel is unbalanced x_{1it}^* is defined by

$$x_{1it}^* = x_{1it} - \theta_i\bar{x}_{1i}$$

The substitution of θ_i for the common θ in the balanced case doesn't make any relevant difference to the status of x_1^*. The only way in which individual-specific information enters θ_i is via the number of observations, T_i, and there is no reason to believe that T_i should be correlated with u_i.

Cornwell and Rupert specification

```
Hausman-Taylor estimates for lwage
using 4165 observations (n = 595, T = 7)

             coefficient    std. error       z      p-value
  ------------------------------------------------------------
  const       2.88442       0.852777       3.382    0.0007   ***
  wks         0.000909009   0.000598818    1.518    0.1290
  south       0.00713766    0.0325480      0.2193   0.8264
  smsa       -0.0417623     0.0194019     -2.152    0.0314   **
  ms         -0.0363440     0.0188575     -1.927    0.0539   *
  exper       0.112972      0.00246967    45.74     0.0000   ***
  exper2     -0.000419119   5.45872e-05   -7.678    1.62e-14 ***
  occ        -0.0213946     0.0137801     -1.553    0.1205
  ind         0.0188416     0.0154404      1.220    0.2224
  union       0.0303548     0.0148964      2.038    0.0416   **
  fem        -0.136847      0.127280      -1.075    0.2823
  blk        -0.281829      0.176627      -1.596    0.1106
  ed          0.140525      0.0658715      2.133    0.0329   **

  sigma_u = 0.94172543
  sigma_e = 0.15180272
  theta   = 0.93918626

Hausman test: chi-square(3) = 14.5555 [0.0022]
Sargan test:  chi-square(3) = 14.8759 [0.0019]
```

Baltagi's specification

```
Hausman-Taylor estimates for lwage
using 4165 observations (n = 595, T = 7)

             coefficient    std. error       z      p-value
  ------------------------------------------------------------
  const       2.91273       0.283652      10.27     9.76e-25 ***
  occ        -0.0207047     0.0137809     -1.502    0.1330
  south       0.00743984    0.0319550      0.2328   0.8159
  smsa       -0.0418334     0.0189581     -2.207    0.0273   **
  ind         0.0136039     0.0152374      0.8928   0.3720
  exper       0.113133      0.00247095    45.79     0.0000   ***
  exper2     -0.000418865   5.45981e-05   -7.672    1.70e-14 ***
  wks         0.000837403   0.000599732    1.396    0.1626
  ms         -0.0298507     0.0189800     -1.573    0.1158
  union       0.0327714     0.0149084      2.198    0.0279   **
  fem        -0.130924      0.126659      -1.034    0.3013
  blk        -0.285748      0.155702      -1.835    0.0665   *
  ed          0.137944      0.0212485      6.492    8.47e-11 ***

  sigma_u = 0.94180300
  sigma_e = 0.15180272
  theta   = 0.93919126

Hausman test: chi-square(3) = 5.25773 [0.1539]
Sargan test:  chi-square(3) = 5.22910 [0.1558]
```

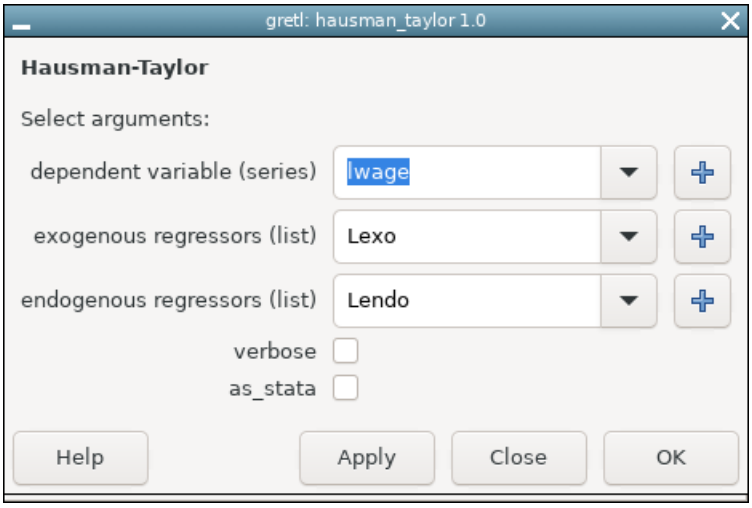

Figure 1: Specify arguments for `hausman_taylor`

We conclude that failing to include x_1^* as an instrument in the unbalanced case will degrade the efficiency of the estimator. Yet this is what is done in Stata's xthtaylor command and in R's plm package.

If the argument above is correct, it should be possible to show via simulation the degradation of the efficiency of Hausman–Taylor when quasi-demeaned x_1 is treated as endogenous in the final IV regression, given unbalanced data. Conversely, if the argument above is wrong then presumably simulation should produce evidence of inconsistency when quasi-demeaned x_1 is added as an instrument.

To explore this we ran a simulation of the following form.

1. For $K = 5000$ iterations, generate a random dataset with a known set of parameter values and a correlation structure that respects the Hausman–Taylor assumptions. Randomly assign missing values to some proportion of the observations.
2. For each dataset, run the Hausman–Taylor procedure both ways (respectively omitting and including x_1^* as an instrument in the final stage) and record the parameter estimates.
3. Calculate the mean and standard deviation of $\hat{\theta} - \theta$ for each parameter θ.

Specifically, we constructed datasets containing one variable in each of the categories x_1, x_2, z_1 and z_2, using the parameter values $\beta_0 = \beta_1 = \beta_2 = \gamma_1 = \gamma_2 = 1$. The panel comprised $T = 10$ observations for each of N individuals. The series were constructed as follows:

$$
\begin{aligned}
u_i &= N(0,1) \\
x_{1it} &= N(0,1) \\
x_{2it} &= N(0,1) + au_i \\
z_{1i} &= N(0,1) \\
z_{2i} &= N(0,1) + au_i + b\bar{x}_{1i} \\
\epsilon_{it} &= N(0,1)
\end{aligned}
$$

with $a = 0.3$ and $b = 0.5$. The formula for x_{2it} ensures that x_2 is endogenous; and that for z_{2i} ensures both the endogeneity of z_2 and its correlation with x_1, which is wanted so that x_1 can serve as an instrument for z_2. In the case of the time-invariant variables, random sequences of length N were generated and the value for each individual was entered at all T observations.

After constructing the data a uniform random series, v, was generated on $[0, 1)$ and the value of the dependent variable was set to "missing" at observations for which $v_{it} < 0.04$, giving an expectation of 4 percent unusable observations, hence unbalancing the panel.

One further point: since σ_u^2 is estimated indirectly, in finite samples it may happen that $\hat{\sigma}_u^2 \leq 0$, in which case the standard procedure is to set $\theta = 0$. This erases the distinction we're interested in, between the two variants of the Hausman–Taylor estimator. It's therefore necessary to calibrate the simulation so that a non-positive $\hat{\sigma}_u^2$ doesn't arise too often.[2]

Figure 2 shows results for the mean error of estimation, relevant to assessing consistency. Points are shown for four values of N: 20, 50, 100 and 200. The results differ only marginally for β_1 and β_2, while for β_0, γ_1 and γ_2 the results are better when x_1^* included as an instrument. There's no evidence of inconsistency in the latter case.

Figure 3 shows the standard deviation of estimate minus parameter, relevant to assessing efficiency; results are again given for four values of N. The relative performance of the variants strongly supports the contention made above. While performance with respect to β_1 and β_2 is virtually identical, the β_0 and γ estimates show *much* greater variance when x_1^* is not included as an instrument; indeed, the variance of the γ estimates is such that they may be useless in practice.

We are now in a position to explain the `as_stata` option for the `hausman_taylor` function mentioned on page 2. This option has no effect for balanced panels, but in the unbalanced case it means "Do what Stata does—that is omit x_1^* as an instrument in the final regression—even though we reckon it's not the right thing to do."

As a practical point, it should be noted that the inclusion (in the unbalanced case) of x_1^* as an instrument alongside $\tilde{x}_1$ and $\bar{x}_1$ in the final Hausman–Taylor regression may produce near-singularity of the instrument matrix, depending on the dataset (this was evident in the simulations). However, with modern econometric software this does not pose a serious problem, since redundant instruments will be dropped automatically.

References

Baltagi, B. H. (2005) *Econometric Analysis of Panel Data, 3e*, Chichester: Wiley.

Baltagi, B. H. and S. Khanti-Akom (1990) 'On efficient estimation with panel data: An empirical comparison of instrumental variables estimators', *Journal of Applied Econometrics* 5: 401–406.

Cornwell, C. and P. Rupert (1988) 'Efficient estimation with panel data: An empirical comparison of instrumental variables estimators', *Journal of Applied Econometrics* 3: 149–155.

Hausman, J. A. and W. E. Taylor (1981) 'Panel data and unobservable individual effects', *Econometrica* 49: 1377–1398.

[2] In addition we discarded iterations in which $\hat{\sigma}_u^2$ was non-positive, continuing until the specified number of replications was reached with non-zero θ.

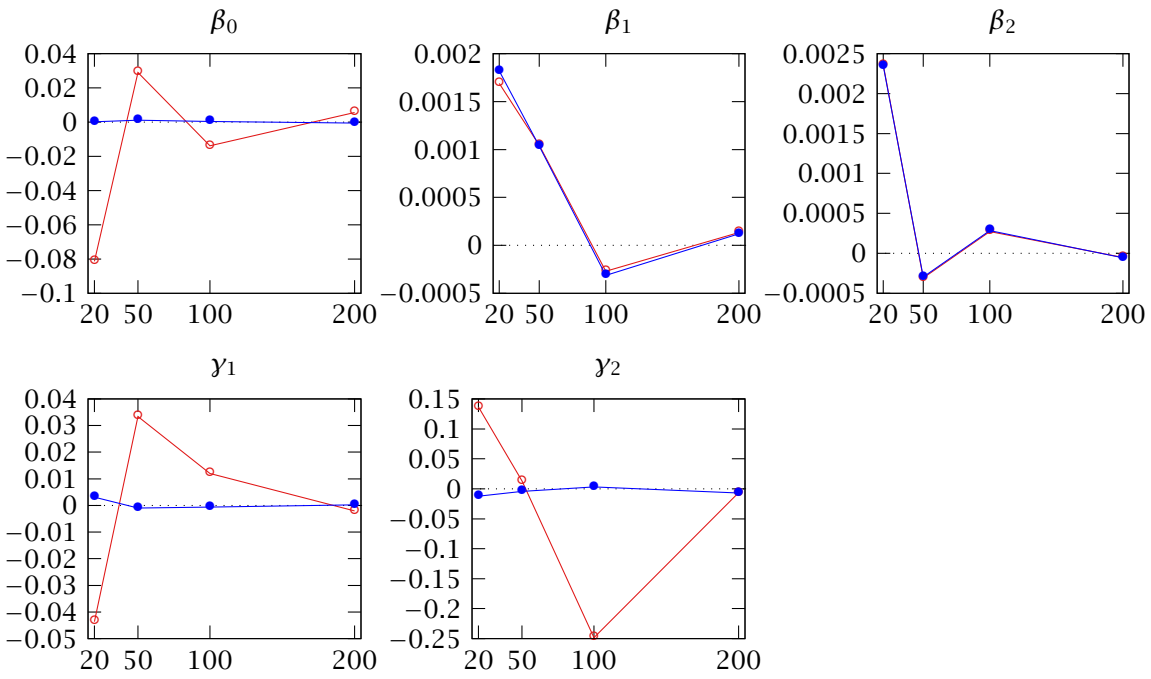

Figure 2: Mean error of estimates (y-axis) against number of individuals, N, in sample (x-axis). Red: x_1^* excluded from the set of instruments; blue, x_1^* included. $T = 10$ and 5000 replications.

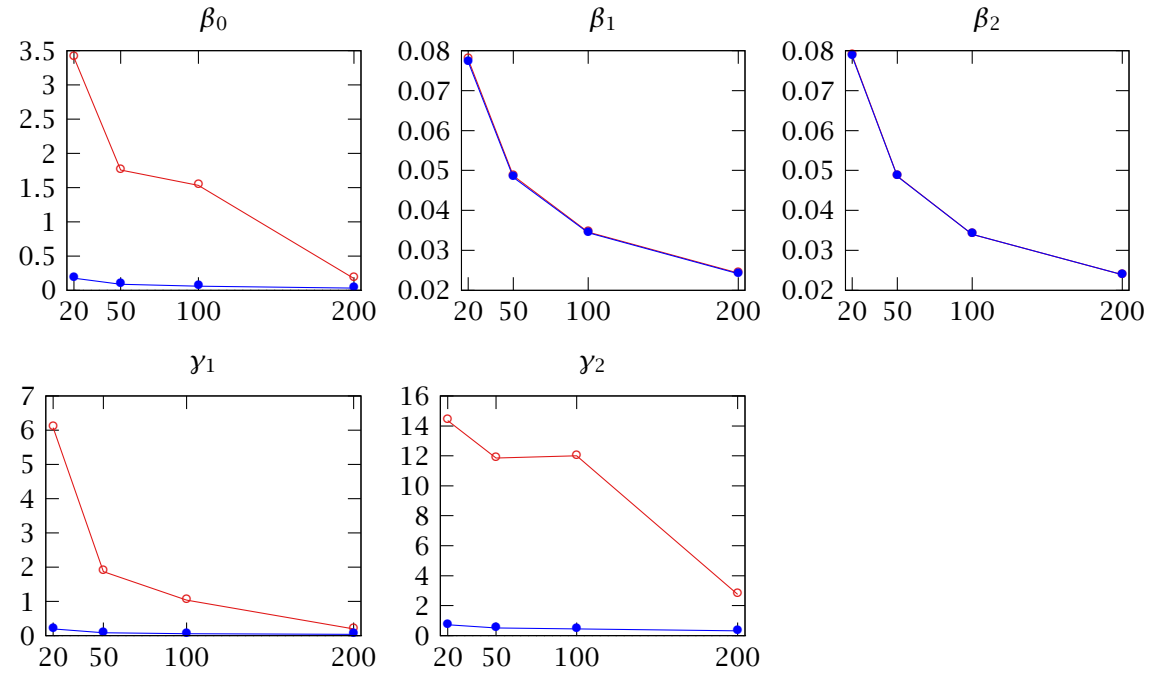

Figure 3: Standard error of estimates (y-axis) against number of individuals, N, in sample (x-axis). Red: x_1^* excluded from the set of instruments; blue x_1^* included. $T = 10$ and 5000 replications.