Title

> **hausman —** Hausman specification test

Description

hausman performs Hausman's (1978) specification test.

Quick start

Hausman test for stored models consistent and efficient

    hausman consistent efficient

As above, but compare fixed-effects and random-effects linear regression models

    hausman fixed random, sigmamore

Endogeneity test after ivprobit and probit with estimates stored in iv and noiv

    hausman iv noiv, equations(1:1)

Test of independence of irrelevant alternatives for model with all alternatives all and model with omitted alternative omitted

    hausman omitted all, alleqs constant

Menu

Statistics > Postestimation

Syntax

> hausman *name-consistent* [*name-efficient*] [, *options*]

options	Description
Main	
<u>c</u>onstant	include estimated intercepts in comparison; default is to exclude
<u>all</u>eqs	use all equations to perform test; default is first equation only
<u>sk</u>ipeqs(*eqlist*)	skip specified equations when performing test
<u>eq</u>uations(*matchlist*)	associate/compare the specified (by number) pairs of equations
force	force performance of test, even though assumptions are not met
df(*#*)	use *#* degrees of freedom
<u>sigmam</u>ore	base both (co)variance matrices on disturbance variance estimate from efficient estimator
<u>sigmal</u>ess	base both (co)variance matrices on disturbance variance estimate from consistent estimator
Advanced	
<u>tc</u>onsistent(*string*)	consistent estimator column header
<u>te</u>fficient(*string*)	efficient estimator column header

where *name-consistent* and *name-efficient* are names under which estimation results were stored via estimates store; see [R] **estimates store**.

A period (.) may be used to refer to the last estimation results, even if these were not already stored.

Not specifying *name-efficient* is equivalent to specifying the last estimation results as ".".

Options

> **Main**

constant specifies that the estimated intercept(s) be included in the model comparison; by default, they are excluded. The default behavior is appropriate for models in which the constant does not have a common interpretation across the two models.

alleqs specifies that all the equations in the models be used to perform the Hausman test; by default, only the first equation is used.

skipeqs(*eqlist*) specifies in *eqlist* the names of equations to be excluded from the test. Equation numbers are not allowed in this context, because the equation names, along with the variable names, are used to identify common coefficients.

equations(*matchlist*) specifies, by number, the pairs of equations that are to be compared.

The *matchlist* in equations() should follow the syntax

$$\#_c \text{:} \#_e \; \left[,\#_c \text{:} \#_e \left[,\dots \right] \right]$$

where $\#_c$ ($\#_e$) is an equation number of the always-consistent (efficient under H_0) estimator. For instance, equations(1:1), equations(1:1, 2:2), or equations(1:2).

If equations() is not specified, then equations are matched on equation names.

equations() handles the situation in which one estimator uses equation names and the other does not. For instance, equations(1:2) means that equation 1 of the always-consistent estimator is to be tested against equation 2 of the efficient estimator. equations(1:1, 2:2) means that equation 1 is to be tested against equation 1 and that equation 2 is to be tested against equation 2. If equations() is specified, the alleqs and skipeqs options are ignored.

force specifies that the Hausman test be performed, even though the assumptions of the Hausman test seem not to be met, for example, because the estimators were pweighted or the data were clustered.

df(#) specifies the degrees of freedom for the Hausman test. The default is the matrix rank of the variance of the difference between the coefficients of the two estimators.

sigmamore and sigmaless specify that the two covariance matrices used in the test be based on a common estimate of disturbance variance (σ^2).

sigmamore specifies that the covariance matrices be based on the estimated disturbance variance from the efficient estimator. This option provides a proper estimate of the contrast variance for so-called tests of exogeneity and overidentification in instrumental-variables regression.

sigmaless specifies that the covariance matrices be based on the estimated disturbance variance from the consistent estimator.

These options can be specified only when both estimators store e(sigma) or e(rmse), or with the xtreg command. e(sigma_e) is stored after the xtreg command with the fe or mle option. e(rmse) is stored after the xtreg command with the re option.

sigmamore or sigmaless are recommended when comparing fixed-effects and random-effects linear regression because they are much less likely to produce a non–positive-definite-differenced covariance matrix (although the tests are asymptotically equivalent whether or not one of the options is specified).

> Advanced

tconsistent(*string*) and tefficient(*string*) are formatting options. They allow you to specify the headers of the columns of coefficients that default to the names of the models. These options will be of interest primarily to programmers.

Remarks and examples

stata.com

hausman is a general implementation of Hausman's (1978) specification test, which compares an estimator $\widehat{\theta}_1$ that is known to be consistent with an estimator $\widehat{\theta}_2$ that is efficient under the assumption being tested. The null hypothesis is that the estimator $\widehat{\theta}_2$ is indeed an efficient (and consistent) estimator of the true parameters. If this is the case, there should be no systematic difference between the two estimators. If there exists a systematic difference in the estimates, you have reason to doubt the assumptions on which the efficient estimator is based.

The assumption of efficiency is violated if the estimator is pweighted or the data are clustered, so hausman cannot be used. The test can be forced by specifying the force option with hausman. For an alternative to using hausman in these cases, see [R] **suest**.

To use hausman, you

```
    . (compute the always-consistent estimator)
    . estimates store name-consistent
    . (compute the estimator that is efficient under H_0)
    . hausman name-consistent .
```

Alternatively, you can turn this around:

> . (*compute the estimator that is efficient under H_0*)
> . `estimates store` *name-efficient*
> . (*fit the less-efficient model*)
> . (*compute the always-consistent estimator*)
> . `hausman . ` *name-efficient*

You can, of course, also compute and store both the always-consistent and efficient-under-H_0 estimators and perform the Hausman test with

> . `hausman` *name-consistent name-efficient*

▷ Example 1

We are studying the factors that affect the wages of young women in the United States between 1970 and 1988, and we have a panel-data sample of individual women over that time span.

```
. use http://www.stata-press.com/data/r14/nlswork4
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)
. describe
Contains data from http://www.stata-press.com/data/r14/nlswork4.dta
  obs:        28,534                          National Longitudinal Survey.
                                                Young Women 14-26 years of age
                                                in 1968
 vars:             6                          29 Jan 2014 16:35
 size:       370,942
```

variable name	storage type	display format	value label	variable label
idcode	int	%8.0g		NLS ID
year	byte	%8.0g		interview year
age	byte	%8.0g		age in current year
msp	byte	%8.0g		1 if married, spouse present
ttl_exp	float	%9.0g		total work experience
ln_wage	float	%9.0g		ln(wage/GNP deflator)

```
Sorted by: idcode  year
```

We believe that a random-effects specification is appropriate for individual-level effects in our model. We fit a fixed-effects model that will capture all temporally constant individual-level effects.

```
. xtreg ln_wage age msp ttl_exp, fe
```

```
Fixed-effects (within) regression              Number of obs      =      28,494
Group variable: idcode                         Number of groups   =       4,710

R-sq:                                          Obs per group:
    within  = 0.1373                                         min =           1
    between = 0.2571                                         avg =         6.0
    overall = 0.1800                                         max =          15

                                               F(3,23781)         =     1262.01
corr(u_i, Xb)  = 0.1476                         Prob > F           =      0.0000
```

ln_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
age	-.005485	.000837	-6.55	0.000	-.0071256	-.0038443
msp	.0033427	.0054868	0.61	0.542	-.0074118	.0140971
ttl_exp	.0383604	.0012416	30.90	0.000	.0359268	.0407941
_cons	1.593953	.0177538	89.78	0.000	1.559154	1.628752

sigma_u	.37674223		
sigma_e	.29751014		
rho	.61591044	(fraction of variance due to u_i)	

```
F test that all u_i=0: F(4709, 23781) = 7.76                 Prob > F = 0.0000
```

We assume that this model is consistent for the true parameters and store the results by using `estimates store` under a name, `fixed`:

```
. estimates store fixed
```

Now we fit a random-effects model as a fully efficient specification of the individual effects under the assumption that they are random and follow a normal distribution. We then compare these estimates with the previously stored results by using the `hausman` command.

```
. xtreg ln_wage age msp ttl_exp, re
```

```
Random-effects GLS regression                  Number of obs      =      28,494
Group variable: idcode                         Number of groups   =       4,710

R-sq:                                          Obs per group:
    within  = 0.1373                                         min =           1
    between = 0.2552                                         avg =         6.0
    overall = 0.1797                                         max =          15

                                               Wald chi2(3)       =     5100.33
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =      0.0000
```

ln_wage	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
age	-.0069749	.0006882	-10.13	0.000	-.0083238	-.0056259
msp	.0046594	.0051012	0.91	0.361	-.0053387	.0146575
ttl_exp	.0429635	.0010169	42.25	0.000	.0409704	.0449567
_cons	1.609916	.0159176	101.14	0.000	1.578718	1.641114

sigma_u	.32648519		
sigma_e	.29751014		
rho	.54633481	(fraction of variance due to u_i)	

```
. hausman fixed ., sigmamore
```

	Coefficients			
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	fixed	.	Difference	S.E.
age	-.005485	-.0069749	.0014899	.0004803
msp	.0033427	.0046594	-.0013167	.0020596
ttl_exp	.0383604	.0429635	-.0046031	.0007181

```
                    b = consistent under Ho and Ha; obtained from xtreg
             B = inconsistent under Ha, efficient under Ho; obtained from xtreg
    Test:  Ho:  difference in coefficients not systematic

            chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                    =        260.40
          Prob>chi2 =      0.0000
```

Under the current specification, our initial hypothesis that the individual-level effects are adequately modeled by a random-effects model is resoundingly rejected. This result is based on the rest of our model specification, and random effects might be appropriate for some alternate model of wages.

◁

Jerry Allen Hausman (1946–) is an American economist and econometrician. He was born in West Virginia and went on to study economics at Brown and Oxford. He joined the MIT faculty in 1972 and continues to teach there. He currently researches new goods and their effects on consumer welfare and its measurement in the Consumer Price Index along with regulation and competition in the telecommunications industry.

Hausman is best known for his many contributions to econometrics. In 1978, he published his now famous paper giving the Hausman specification test. The work remains one of the most widely cited econometrics papers. He has also done extensive work in applied microeconomics pertaining to governments role in the economy, including antitrust regulation, public finance, and taxation.

In 1980, Hausman received the Frisch Medal, a biennial award from the Econometric Society recognizing exceptional applied work, for his paper with David Wise on attrition bias. In 1985, he won the John Bates Clark Award from the American Economics Association, which is given for outstanding contributions to economics by an economist under 40 years of age. In 2012, the *Advances in Econometrics* book series devoted an entire volume to Hausman and his contributions to econometrics.

▷ Example 2

A stringent assumption of multinomial and conditional logit models is that outcome categories for the model have the property of independence of irrelevant alternatives (IIA). Stated simply, this assumption requires that the inclusion or exclusion of categories does not affect the relative risks associated with the regressors in the remaining categories.

One classic example of a situation in which this assumption would be violated involves the choice of transportation mode; see McFadden (1974). For simplicity, postulate a transportation model with the four possible outcomes: rides a train to work, takes a bus to work, drives the Ford to work, and drives the Chevrolet to work. Clearly, "drives the Ford" is a closer substitute to "drives the Chevrolet" than it is to "rides a train" (at least for most people). This means that excluding "drives the Ford" from the model could be expected to affect the relative risks of the remaining options and that the model would not obey the IIA assumption.

Using the data presented in [R] **mlogit**, we will use a simplified model to test for IIA. The choice of insurance type among indemnity, prepaid, and uninsured is modeled as a function of age and gender. The indemnity category is allowed to be the base category, and the model including all three outcomes is fit. The results are then stored under the name allcats.

```
. use http://www.stata-press.com/data/r14/sysdsn3
(Health insurance data)

. mlogit insure age male

Iteration 0:   log likelihood = -555.85446
Iteration 1:   log likelihood = -551.32973
Iteration 2:   log likelihood = -551.32802
Iteration 3:   log likelihood = -551.32802

Multinomial logistic regression                  Number of obs   =        615
                                                 LR chi2(4)      =       9.05
                                                 Prob > chi2     =     0.0598
Log likelihood = -551.32802                      Pseudo R2       =     0.0081
```

insure	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
Indemnity	(base outcome)					
Prepaid						
age	-.0100251	.0060181	-1.67	0.096	-.0218204	.0017702
male	.5095747	.1977893	2.58	0.010	.1219147	.8972346
_cons	.2633838	.2787575	0.94	0.345	-.2829708	.8097383
Uninsure						
age	-.0051925	.0113821	-0.46	0.648	-.0275011	.0171161
male	.4748547	.3618462	1.31	0.189	-.2343508	1.18406
_cons	-1.756843	.5309602	-3.31	0.001	-2.797506	-.7161803

```
. estimates store allcats
```

Under the IIA assumption, we would expect no systematic change in the coefficients if we excluded one of the outcomes from the model. (For an extensive discussion, see Hausman and McFadden [1984].) We reestimate the parameters, excluding the uninsured outcome, and perform a Hausman test against the fully efficient full model.

```
. mlogit insure age male if insure != "Uninsure":insure

Iteration 0:   log likelihood =  -394.8693
Iteration 1:   log likelihood =  -390.4871
Iteration 2:   log likelihood = -390.48643
Iteration 3:   log likelihood = -390.48643

Multinomial logistic regression                  Number of obs   =        570
                                                 LR chi2(2)      =       8.77
                                                 Prob > chi2     =     0.0125
Log likelihood = -390.48643                      Pseudo R2       =     0.0111
```

insure	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
Indemnity	(base outcome)					
Prepaid						
age	-.0101521	.0060049	-1.69	0.091	-.0219214	.0016173
male	.5144003	.1981735	2.60	0.009	.1259874	.9028133
_cons	.2678043	.2775563	0.96	0.335	-.276196	.8118046

```
. hausman . allcats, alleqs constant
```

	── Coefficients ── (b) .	(B) allcats	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
age	-.0101521	-.0100251	-.0001269	.
male	.5144003	.5095747	.0048256	.0123338
_cons	.2678043	.2633838	.0044205	.

```
                       b = consistent under Ho and Ha; obtained from mlogit
            B = inconsistent under Ha, efficient under Ho; obtained from mlogit
      Test:  Ho:  difference in coefficients not systematic

                  chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        0.08
               Prob>chi2 =      0.9944
               (V_b-V_B is not positive definite)
```

The syntax of the if condition on the mlogit command simply identified the "Uninsured" category with the insure value label; see [U] **12.6.3 Value labels**. On examining the output from hausman, we see that there is no evidence that the IIA assumption has been violated.

Because the Hausman test is a standardized comparison of model coefficients, using it with mlogit requires that the base outcome be the same in both competing models. In particular, if the most-frequent category (the default base outcome) is being removed to test for IIA, you must use the baseoutcome() option in mlogit to manually set the base outcome to something else. Or you can use the equation() option of the hausman command to align the equations of the two models.

Having the missing values for the square root of the diagonal of the covariance matrix of the differences is not comforting, but it is also not surprising. This covariance matrix is guaranteed to be positive definite only asymptotically (it is a consequence of the assumption that one of the estimators is efficient), and assurances are not made about the diagonal elements. Negative values along the diagonal are possible, and the fourth column of the table is provided mainly for descriptive use.

We can also perform the Hausman IIA test against the remaining alternative in the model:

```
. mlogit insure age male if insure != "Prepaid":insure

Iteration 0:   log likelihood = -132.59913
Iteration 1:   log likelihood = -131.78009
Iteration 2:   log likelihood = -131.76808
Iteration 3:   log likelihood = -131.76807
```

Multinomial logistic regression				Number of obs	=	338
				LR chi2(2)	=	1.66
				Prob > chi2	=	0.4356
Log likelihood = -131.76807				Pseudo R2	=	0.0063

insure	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
Indemnity	(base outcome)					
Uninsure						
age	-.0041055	.0115807	-0.35	0.723	-.0268033	.0185923
male	.4591074	.3595663	1.28	0.202	-.2456296	1.163844
_cons	-1.801774	.5474476	-3.29	0.001	-2.874752	-.7287968

```
. hausman . allcats, alleqs constant
                 ―― Coefficients ――
              │      (b)          (B)          (b-B)     sqrt(diag(V_b-V_B))
              │       .          allcats     Difference          S.E.
──────────────┼────────────────────────────────────────────────────────────
          age │  -.0041055    -.0051925       .001087          .0021355
         male │   .4591074     .4748547      -.0157473             .
        _cons │  -1.801774    -1.756843      -.0449311          .1333421
──────────────┴────────────────────────────────────────────────────────────
                         b = consistent under Ho and Ha; obtained from mlogit
          B = inconsistent under Ha, efficient under Ho; obtained from mlogit
    Test:  Ho:  difference in coefficients not systematic

                  chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =    -0.18    chi2<0 ==> model fitted on these
                                        data fails to meet the asymptotic
                                        assumptions of the Hausman test;
                                        see suest for a generalized test
```

Here the χ^2 statistic is actually negative. We might interpret this result as strong evidence that we cannot reject the null hypothesis. Such a result is not an unusual outcome for the Hausman test, particularly when the sample is relatively small—there are only 45 uninsured individuals in this dataset.

Are we surprised by the results of the Hausman test in this example? Not really. Judging from the z statistics on the original multinomial logit model, we were struggling to identify any structure in the data with the current specification. Even when we were willing to assume IIA and computed the efficient estimator under this assumption, few of the effects could be identified as statistically different from those on the base category. Trying to base a Hausman test on a contrast (difference) between two poor estimates is just asking too much of the existing data.

◁

In example 2, we encountered a case in which the Hausman was not well defined. Unfortunately, in our experience this happens fairly often. Stata provides an alternative to the Hausman test that overcomes this problem through an alternative estimator of the variance of the difference between the two estimators. This other estimator is guaranteed to be positive semidefinite. This alternative estimator also allows a widening of the scope of problems to which Hausman-type tests can be applied by relaxing the assumption that one of the estimators is efficient. For instance, you can perform Hausman-type tests to clustered observations and survey estimators. See [R] suest for details.

Stored results

hausman stores the following in r():

Scalars
  r(chi2) χ^2
  r(df)   degrees of freedom for the statistic
  r(p) p-value for the χ^2
  r(rank)   rank of (V_b-V_B)^(-1)

Methods and formulas

The Hausman statistic is distributed as χ^2 and is computed as

$$H = (\beta_c - \beta_e)'(V_c - V_e)^{-1}(\beta_c - \beta_e)$$

where

β_c	is the coefficient vector from the consistent estimator
β_e	is the coefficient vector from the efficient estimator
V_c	is the covariance matrix of the consistent estimator
V_e	is the covariance matrix of the efficient estimator

When the difference in the variance matrices is not positive definite, a Moore–Penrose generalized inverse is used. As noted in Gourieroux and Monfort (1995, 125–128), the choice of generalized inverse is not important asymptotically.

The number of degrees of freedom for the statistic is the rank of the difference in the variance matrices. When the difference is positive definite, this is the number of common coefficients in the models being compared.

Acknowledgment

Portions of `hausman` are based on an early implementation by Jeroen Weesie of the Department of Sociology at Utrecht University, The Netherlands.

References

Baltagi, B. H. 2011. *Econometrics*. 5th ed. Berlin: Springer.

Gourieroux, C. S., and A. Monfort. 1995. *Statistics and Econometric Models, Vol 2: Testing, Confidence Regions, Model Selection, and Asymptotic Theory*. Trans. Q. Vuong. Cambridge: Cambridge University Press.

Hausman, J. A. 1978. Specification tests in econometrics. *Econometrica* 46: 1251–1271.

Hausman, J. A., and D. L. McFadden. 1984. Specification tests for the multinomial logit model. *Econometrica* 52: 1219–1240.

McFadden, D. L. 1974. Measurement of urban travel demand. *Journal of Public Economics* 3: 303–328.

Also see