Title

> **heckman postestimation** — Postestimation tools for heckman

Description

The following postestimation commands are available after `heckman`:

Command	Description
contrast	contrasts and ANOVA-style joint tests of estimates
estat ic[1]	Akaike's and Schwarz's Bayesian information criteria (AIC and BIC)
estat summarize	summary statistics for the estimation sample
estat vce	variance–covariance matrix of the estimators (VCE)
estat (svy)	postestimation statistics for survey data
estimates	cataloging estimation results
lincom	point estimates, standard errors, testing, and inference for linear combinations of coefficients
lrtest[2]	likelihood-ratio test; not available with two-step estimator
margins	marginal means, predictive margins, marginal effects, and average marginal effects
marginsplot	graph the results from margins (profile plots, interaction plots, etc.)
nlcom	point estimates, standard errors, testing, and inference for nonlinear combinations of coefficients
predict	predictions, residuals, influence statistics, and other diagnostic measures
predictnl	point estimates, standard errors, testing, and inference for generalized predictions
pwcompare	pairwise comparisons of estimates
suest[1]	seemingly unrelated estimation
test	Wald tests of simple and composite linear hypotheses
testnl	Wald tests of nonlinear hypotheses

[1] `estat ic` and `suest` are not appropriate after `heckman, twostep`.

[2] `lrtest` is not appropriate with svy estimation results.

Syntax for predict

After ML or twostep

> predict [*type*] *newvar* [*if*] [*in*] [, *statistic* <u>nooff</u>set]

After ML

> predict [*type*] { *stub** | *newvar*_reg *newvar*_sel *newvar*_athrho *newvar*_lnsigma }
>
>   [*if*] [*in*] , <u>sc</u>ores

statistic	Description
Main	
xb	linear prediction; the default
stdp	standard error of the prediction
stdf	standard error of the forecast
xbsel	linear prediction for selection equation
stdpsel	standard error of the linear prediction for selection equation
pr(*a*,*b*)	$\Pr(y_j \mid a < y_j < b)$
e(*a*,*b*)	$E(y_j \mid a < y_j < b)$
ystar(*a*,*b*)	$E(y_j^*)$, $y_j^* = \max\{a, \min(y_j, b)\}$
ycond	$E(y_j \mid y_j \text{ observed})$
yexpected	$E(y_j^*)$, y_j taken to be 0 where unobserved
nshazard or mills	nonselection hazard (also called the inverse of Mills' ratio)
psel	$\Pr(y_j \text{ observed})$

These statistics are available both in and out of sample; type predict ... if e(sample) ... if wanted only for the estimation sample.

stdf is not allowed with svy estimation results.

where *a* and *b* may be numbers or variables; *a* missing ($a \geq .$) means $-\infty$, and *b* missing ($b \geq .$) means $+\infty$; see [U] **12.2.1 Missing values**.

Menu for predict

Statistics > Postestimation > Predictions, residuals, etc.

Options for predict

> ┌─── Main ───

xb, the default, calculates the linear prediction $\mathbf{x}_j \mathbf{b}$.

stdp calculates the standard error of the prediction, which can be thought of as the standard error of the predicted expected value or mean for the observation's covariate pattern. The standard error of the prediction is also referred to as the standard error of the fitted value.

stdf calculates the standard error of the forecast, which is the standard error of the point prediction for 1 observation. It is commonly referred to as the standard error of the future or forecast value. By construction, the standard errors produced by stdf are always larger than those produced by stdp; see *Methods and formulas* in [R] **regress postestimation**.

xbsel calculates the linear prediction for the selection equation.

stdpsel calculates the standard error of the linear prediction for the selection equation.

pr(a,b) calculates $\Pr(a < \mathbf{x}_j\mathbf{b} + u_1 < b)$, the probability that $y_j|\mathbf{x}_j$ would be observed in the interval (a, b).

> a and b may be specified as numbers or variable names; *lb* and *ub* are variable names; pr(20,30) calculates $\Pr(20 < \mathbf{x}_j\mathbf{b} + u_1 < 30)$; pr(*lb*,*ub*) calculates $\Pr(lb < \mathbf{x}_j\mathbf{b} + u_1 < ub)$; and pr(20,*ub*) calculates $\Pr(20 < \mathbf{x}_j\mathbf{b} + u_1 < ub)$.

> a missing ($a \geq .$) means $-\infty$; pr(.,30) calculates $\Pr(-\infty < \mathbf{x}_j\mathbf{b} + u_j < 30)$; pr(*lb*,30) calculates $\Pr(-\infty < \mathbf{x}_j\mathbf{b} + u_j < 30)$ in observations for which *lb* $\geq .$ and calculates $\Pr(lb < \mathbf{x}_j\mathbf{b} + u_j < 30)$ elsewhere.

> b missing ($b \geq .$) means $+\infty$; pr(20,.) calculates $\Pr(+\infty > \mathbf{x}_j\mathbf{b} + u_j > 20)$; pr(20,*ub*) calculates $\Pr(+\infty > \mathbf{x}_j\mathbf{b} + u_j > 20)$ in observations for which *ub* $\geq .$ and calculates $\Pr(20 < \mathbf{x}_j\mathbf{b} + u_j < ub)$ elsewhere.

e(a,b) calculates $E(\mathbf{x}_j\mathbf{b} + u_1 \mid a < \mathbf{x}_j\mathbf{b} + u_1 < b)$, the expected value of $y_j|\mathbf{x}_j$ conditional on $y_j|\mathbf{x}_j$ being in the interval (a, b), meaning that $y_j|\mathbf{x}_j$ is truncated. a and b are specified as they are for pr().

ystar(a,b) calculates $E(y_j^*)$, where $y_j^* = a$ if $\mathbf{x}_j\mathbf{b} + u_j \leq a$, $y_j^* = b$ if $\mathbf{x}_j\mathbf{b} + u_j \geq b$, and $y_j^* = \mathbf{x}_j\mathbf{b} + u_j$ otherwise, meaning that y_j^* is censored. a and b are specified as they are for pr().

ycond calculates the expected value of the dependent variable conditional on the dependent variable being observed, that is, selected; $E(y_j \mid y_j \text{ observed})$.

yexpected calculates the expected value of the dependent variable (y_j^*), where that value is taken to be 0 when it is expected to be unobserved; $y_j^* = \Pr(y_j \text{ observed})E(y_j \mid y_j \text{ observed})$.

> The assumption of 0 is valid for many cases where nonselection implies nonparticipation (for example, unobserved wage levels, insurance claims from those who are uninsured) but may be inappropriate for some problems (for example, unobserved disease incidence).

nshazard and mills are synonyms; both calculate the nonselection hazard—what Heckman (1979) referred to as the inverse of the Mills' ratio—from the selection equation.

psel calculates the probability of selection (or being observed): $\Pr(y_j \text{ observed}) = \Pr(\mathbf{z}_j\boldsymbol{\gamma} + u_{2j} > 0)$.

nooffset is relevant when you specify offset(*varname*) for heckman. It modifies the calculations made by predict so that they ignore the offset variable; the linear prediction is treated as $\mathbf{x}_j\mathbf{b}$ rather than as $\mathbf{x}_j\mathbf{b} + \text{offset}_j$.

scores, not available with twostep, calculates equation-level score variables.

> The first new variable will contain $\partial \ln L / \partial(\mathbf{x}_j\boldsymbol{\beta})$.

> The second new variable will contain $\partial \ln L / \partial(\mathbf{z}_j\boldsymbol{\gamma})$.

> The third new variable will contain $\partial \ln L / \partial(\text{atanh}\,\rho)$.

> The fourth new variable will contain $\partial \ln L / \partial(\ln \sigma)$.

Remarks and examples

▷ Example 1

The default statistic produced by `predict` after `heckman` is the expected value of the dependent variable from the underlying distribution of the regression model. In the wage model of [R] **heckman**, this is the expected wage rate among all women, regardless of whether they were observed to participate in the labor force:

```
. use http://www.stata-press.com/data/r13/womenwk
. heckman wage educ age, select(married children educ age) vce(cluster county)
  (output omitted )
. predict heckwage
(option xb assumed; fitted values)
```

It is instructive to compare these predicted wage values from the Heckman model with an ordinary regression model—a model without the selection adjustment:

```
. regress wage educ age
```

Source	SS	df	MS		Number of obs =	1343
					F(2, 1340) =	227.49
Model	13524.0337	2	6762.01687		Prob > F =	0.0000
Residual	39830.8609	1340	29.7245231		R-squared =	0.2535
					Adj R-squared =	0.2524
Total	53354.8946	1342	39.7577456		Root MSE =	5.452

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
education	.8965829	.0498061	18.00	0.000	.7988765	.9942893
age	.1465739	.0187135	7.83	0.000	.109863	.1832848
_cons	6.084875	.8896182	6.84	0.000	4.339679	7.830071

```
. predict regwage
(option xb assumed; fitted values)
. summarize heckwage regwage
```

Variable	Obs	Mean	Std. Dev.	Min	Max
heckwage	2000	21.15532	3.83965	14.6479	32.85949
regwage	2000	23.12291	3.241911	17.98218	32.66439

Since this dataset was concocted, we know the true coefficients of the wage regression equation to be 1, 0.2, and 1, respectively. We can compute the true mean wage for our sample.

```
. generate truewage = 1 + .2*age + 1*educ
. summarize truewage
```

Variable	Obs	Mean	Std. Dev.	Min	Max
truewage	2000	21.3256	3.797904	15	32.8

Whereas the mean of the predictions from `heckman` is within 18 cents of the true mean wage, ordinary regression yields predictions that are on average about $1.80 per hour too high because of the selection effect. The regression predictions also show somewhat less variation than the true wages.

The coefficients from `heckman` are so close to the true values that they are not worth testing. Conversely, the regression equation is significantly off but seems to give the right sense. Would we be led far astray if we relied on the OLS coefficients? The effect of age is off by more than 5 cents per year of age, and the coefficient on education level is off by about 10%. We can test the OLS coefficient on education level against the true value by using `test`.

```
. test educ = 1

 ( 1)   education = 1

        F(  1,   1340) =     4.31
              Prob > F =    0.0380
```

Not only is the OLS coefficient on education substantially lower than the true parameter, but the difference from the true parameter is also statistically significant beyond the 5% level. We can perform a similar test for the OLS age coefficient:

```
. test age = .2

 ( 1)   age = .2

        F(  1,   1340) =     8.15
              Prob > F =    0.0044
```

We find even stronger evidence that the OLS regression results are biased away from the true parameters.

◁

▷ Example 2

Several other interesting aspects of the Heckman model can be explored with `predict`. Continuing with our wage model, we can obtain the expected wages for women conditional on participating in the labor force with the `ycond` option. Let's get these predictions and compare them with actual wages for women participating in the labor force.

```
. use http://www.stata-press.com/data/r13/womenwk, clear
. heckman wage educ age, select(married children educ age)
  (output omitted)
. predict hcndwage, ycond
. summarize wage hcndwage if wage != .
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	1343	23.69217	6.305374	5.88497	45.80979
hcndwage	1343	23.68239	3.335087	16.18337	33.7567

We see that the average predictions from `heckman` are close to the observed levels but do not have the same mean. These conditional wage predictions are available for all observations in the dataset but can be directly compared only with observed wages, where individuals are participating in the labor force.

What if we were interested in making predictions about mean wages for all women? Here the expected wage is 0 for those who are not expected to participate in the labor force, with expected participation determined by the selection equation. These values can be obtained with the `yexpected` option of `predict`. For comparison, a variable can be generated where the wage is set to 0 for nonparticipants.

```
. predict hexpwage, yexpected
. generate wage0 = wage
(657 missing values generated)
. replace wage0 = 0 if wage == .
(657 real changes made)
```

```
. summarize hexpwage wage0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hexpwage	2000	15.92511	5.979336	2.492469	32.45858
wage0	2000	15.90929	12.27081	0	45.80979

Again we note that the predictions from heckman are close to the observed mean hourly wage rate for all women. Why aren't the predictions using ycond and yexpected equal to their observed sample equivalents? For the Heckman model, unlike linear regression, the sample moments implied by the optimal solution to the model likelihood do not require that these predictions match observed data. Properly accounting for the additional variation from the selection equation requires that the model use more information than just the sample moments of the observed wages.

◁

Reference

Heckman, J. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–161.

Also see

[R] **heckman** — Heckman selection model

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