Title

> **xtintreg** — Random-effects interval-data regression models

Syntax	Menu	Description	Options
Remarks and examples	Stored results	Methods and formulas	References
Also see			

Syntax

xtintreg *depvar*~lower~ *depvar*~upper~ [*indepvars*] [*if*] [*in*] [*weight*] [, *options*]

options	Description
Model	
noconstant	suppress constant term
offset(*varname*)	include *varname* in model with coefficient constrained to 1
constraints(*constraints*)	apply specified linear constraints
collinear	keep collinear variables
SE	
vce(*vcetype*)	*vcetype* may be oim, bootstrap, or jackknife
Reporting	
level(*#*)	set confidence level; default is level(95)
noskip	perform overall model test as a likelihood-ratio test
intreg	perform likelihood-ratio test against pooled model
nocnsreport	do not display constraints
display_options	control column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
Integration	
intmethod(*intmethod*)	integration method; *intmethod* may be mvaghermite (the default) or ghermite
intpoints(*#*)	use *#* quadrature points; default is intpoints(12)
Maximization	
maximize_options	control the maximization process; see [R] **maximize**
coeflegend	display legend instead of statistics

A panel variable must be specified; use xtset; see [XT] **xtset**.

indepvars may contain factor variables; see [U] **11.4.3 Factor variables**.

depvar~lower~, *depvar*~upper~, and *indepvars* may contain time-series operators; see [U] **11.4.4 Time-series varlists**.

by and statsby are allowed; see [U] **11.1.10 Prefix commands**.

iweights are allowed; see [U] **11.1.6 weight**. Weights must be constant within panel.

coeflegend does not appear in the dialog box.

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

Menu

Description

xtintreg fits a random-effects regression model whose dependent variable may be measured as point data, interval data, left-censored data, or right-censored data. *depvar*$_{\text{lower}}$ and *depvar*$_{\text{upper}}$ represent how the dependent variable was measured.

The values in *depvar*$_{\text{lower}}$ and *depvar*$_{\text{upper}}$ should have the following form:

Type of data		*depvar*$_{\text{lower}}$	*depvar*$_{\text{upper}}$
point data	$a = [a, a]$	a	a
interval data	$[a, b]$	a	b
left-censored data	$(-\infty, b]$	.	b
right-censored data	$[a, +\infty)$	a	.

Options

Model

noconstant, offset(*varname*), constraints(*constraints*), collinear; see [R] **estimation options**.

SE

vce(*vcetype*) specifies the type of standard error reported, which includes types that are derived from asymptotic theory (oim) and that use bootstrap or jackknife methods (bootstrap, jackknife); see [XT] *vce_options*.

Reporting

level(*#*), noskip; see [R] **estimation options**.

intreg specifies that a likelihood-ratio test comparing the random-effects model with the pooled (intreg) model be included in the output.

nocnsreport; see [R] **estimation options**.

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

Integration

intmethod(*intmethod*), intpoints(*#*); see [R] **estimation options**.

Maximization

maximize_options: difficult, technique(*algorithm_spec*), iterate(*#*), [no]log, trace, gradient, showstep, hessian, showtolerance, tolerance(*#*), ltolerance(*#*), nrtolerance(*#*), nonrtolerance, and from(*init_specs*); see [R] **maximize**. These options are seldom used.

The following option is available with xtintreg but is not shown in the dialog box:

coeflegend; see [R] estimation options.

Remarks and examples

Consider the linear regression model with panel-level random effects

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \nu_i + \epsilon_{it}$$

for $i = 1, \ldots, n$ panels, where $t = 1, \ldots, n_i$. The random effects, ν_i, are i.i.d., $N(0, \sigma_\nu^2)$, and ϵ_{it} are i.i.d., $N(0, \sigma_\epsilon^2)$ independently of ν_i. The observed data consist of the couples, (y_{1it}, y_{2it}), such that all that is known is that $y_{1it} \leq y_{it} \leq y_{2it}$, where y_{1it} is possibly $-\infty$ and y_{2it} is possibly $+\infty$.

▷ Example 1

We begin with the nlswork dataset described in [XT] **xt** and create two fictional dependent variables, where the wages are instead reported sometimes as ranges. The wages have been adjusted to 1988 dollars and have further been recoded such that some of the observations are known exactly, some are left-censored, some are right-censored, and some are known only in an interval.

We wish to fit a random-effects interval regression model of adjusted (log) wages:

```
. use http://www.stata-press.com/data/r13/nlswork5
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)

. xtintreg ln_wage1 ln_wage2 union age grade south##c.year occ_code, intreg
  (output omitted)
```

Random-effects interval regression				Number of obs	=	19151
Group variable: idcode				Number of groups	=	4140
Random effects u_i ~ Gaussian				Obs per group: min =		1
				avg =		4.6
				max =		12
Integration method: mvaghermite				Integration points =		12
				Wald chi2(7)	=	2523.84
Log likelihood = -23174.355				Prob > chi2	=	0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf. Interval]	
union	.1441844	.0094245	15.30	0.000	.1257128	.162656
age	.0104083	.0018804	5.54	0.000	.0067228	.0140939
grade	.0794958	.0023469	33.87	0.000	.074896	.0840955
1.south	-.3778103	.0979415	-3.86	0.000	-.5697722	-.1858485
year	.0013528	.0020176	0.67	0.503	-.0026016	.0053071
south#c.year						
1	.0034385	.0012105	2.84	0.005	.0010659	.005811
occ_code	-.0197912	.0014094	-14.04	0.000	-.0225535	-.0170289
_cons	.3791078	.1136641	3.34	0.001	.1563303	.6018853
/sigma_u	.2987074	.0052697	56.68	0.000	.2883789	.309036
/sigma_e	.3528109	.0030935	114.05	0.000	.3467478	.358874
rho	.4175266	.0102529			.3975474	.4377211

Likelihood-ratio test of sigma_u=0: chibar2(01)= 2516.85 Prob>=chibar2 = 0.000

```
    Observation summary:        4757  left-censored observations
                                4792     uncensored observations
                                4830 right-censored observations
                                4772      interval observations
```

The output includes the overall and panel-level variance components (labeled `sigma_e` and `sigma_u`, respectively) together with ρ (labeled `rho`),

$$\rho = \frac{\sigma_\nu^2}{\sigma_\epsilon^2 + \sigma_\nu^2}$$

which is the proportion of the total variance contributed by the panel-level variance component.

When `rho` is zero, the panel-level variance component is unimportant, and the panel estimator is not different from the pooled estimator. A likelihood-ratio test of this is included at the bottom of the output. This test formally compares the pooled estimator (intreg) with the panel estimator.

◁

❑ Technical note

The random-effects model is calculated using quadrature, which is an approximation whose accuracy depends partially on the number of integration points used. We can use the `quadchk` command to see if changing the number of integration points affects the results. If the results change, the quadrature approximation is not accurate given the number of integration points. Try increasing the number of integration points using the `intpoints()` option and run `quadchk` again. Do not attempt to interpret the results of estimates when the coefficients reported by `quadchk` differ substantially. See [XT] **quadchk** for details and [XT] **xtprobit** for an example.

Because the `xtintreg` likelihood function is calculated by Gauss–Hermite quadrature, on large problems the computations can be slow. Computation time is roughly proportional to the number of points used for the quadrature.

❑

Stored results

xtintreg stores the following in e():

Scalars

e(N)	number of observations
e(N_g)	number of groups
e(N_unc)	number of uncensored observations
e(N_lc)	number of left-censored observations
e(N_rc)	number of right-censored observations
e(N_int)	number of interval observations
e(N_cd)	number of completely determined observations
e(k)	number of parameters
e(k_aux)	number of auxiliary parameters
e(k_eq)	number of equations in e(b)
e(k_eq_model)	number of equations in overall model test
e(k_dv)	number of dependent variables
e(df_m)	model degrees of freedom
e(ll)	log likelihood
e(ll_0)	log likelihood, constant-only model
e(chi2)	χ^2
e(chi2_c)	χ^2 for comparison test
e(rho)	ρ
e(sigma_u)	panel-level standard deviation
e(sigma_e)	standard deviation of ϵ_{it}
e(n_quad)	number of quadrature points
e(g_min)	smallest group size
e(g_avg)	average group size
e(g_max)	largest group size
e(p)	significance
e(rank)	rank of e(V)
e(rank0)	rank of e(V) for constant-only model
e(ic)	number of iterations
e(rc)	return code
e(converged)	1 if converged, 0 otherwise

Macros
    e(cmd)   xtintreg
    e(cmdline)   command as typed
    e(depvar)   names of dependent variables
    e(ivar)   variable denoting groups
    e(wtype)   weight type
    e(wexp)   weight expression
    e(title)   title in estimation output
    e(offset1)   offset
    e(chi2type)   Wald or LR; type of model χ^2 test
    e(chi2_ct)   Wald or LR; type of model χ^2 test corresponding to e(chi2_c)
    e(vce)   *vcetype* specified in vce()
    e(vcetype)   title used to label Std. Err.
    e(intmethod)   integration method
    e(distrib)   Gaussian; the distribution of the random effect
    e(opt)   type of optimization
    e(which)   max or min; whether optimizer is to perform maximization or minimization
    e(ml_method)   type of ml method
    e(user)   name of likelihood-evaluator program
    e(technique)   maximization technique
    e(properties)   b V
    e(predict)   program used to implement predict
    e(asbalanced)   factor variables fvset as asbalanced
    e(asobserved)   factor variables fvset as asobserved

Matrices
    e(b)   coefficient vector
    e(Cns)   constraints matrix
    e(ilog)   iteration log
    e(gradient)   gradient vector
    e(V)   variance–covariance matrix of the estimators

Functions
    e(sample)   marks estimation sample

Methods and formulas

Assuming a normal distribution, $N(0, \sigma_\nu^2)$, for the random effects ν_i, we have the joint (unconditional of ν_i) density of the observed data for the ith panel

$$
f\left\{(y_{1i1}, y_{2i1}), \ldots, (y_{1in_i}, y_{2in_i}) | \mathbf{x}_{1i}, \ldots, \mathbf{x}_{in_i}\right\} =
$$
$$
\int_{-\infty}^{\infty} \frac{e^{-\nu_i^2/2\sigma_\nu^2}}{\sqrt{2\pi}\sigma_\nu} \left\{\prod_{t=1}^{n_i} F(y_{1it}, y_{2it}, \mathbf{x}_{it}\boldsymbol{\beta} + \nu_i)\right\} d\nu_i
$$

where

$$
F(y_{1it}, y_{2it}, \Delta_{it}) = \begin{cases}
\left(\sqrt{2\pi}\sigma_\epsilon\right)^{-1} e^{-(y_{1it}-\Delta_{it})^2/(2\sigma_\epsilon^2)} & \text{if } (y_{1it}, y_{2it}) \in C \\
\Phi\left(\frac{y_{2it}-\Delta_{it}}{\sigma_\epsilon}\right) & \text{if } (y_{1it}, y_{2it}) \in L \\
1 - \Phi\left(\frac{y_{1it}-\Delta_{it}}{\sigma_\epsilon}\right) & \text{if } (y_{1it}, y_{2it}) \in R \\
\Phi\left(\frac{y_{2it}-\Delta_{it}}{\sigma_\epsilon}\right) - \Phi\left(\frac{y_{1it}-\Delta_{it}}{\sigma_\epsilon}\right) & \text{if } (y_{1it}, y_{2it}) \in I
\end{cases}
$$

where C is the set of noncensored observations ($y_{1it} = y_{2it}$ and both nonmissing), L is the set of left-censored observations (y_{1it} missing and y_{2it} nonmissing), R is the set of right-censored observations (y_{1it} nonmissing and y_{2it} missing), I is the set of interval observations ($y_{1it} < y_{2it}$ and both nonmissing), and $\Phi()$ is the cumulative normal distribution.

The panel-level likelihood l_i is given by

$$
l_i = \int_{-\infty}^{\infty} \frac{e^{-\nu_i^2/2\sigma_\nu^2}}{\sqrt{2\pi}\sigma_\nu} \left\{ \prod_{t=1}^{n_i} F(y_{1it}, y_{2it}, \mathbf{x}_{it}\boldsymbol{\beta} + \nu_i) \right\} d\nu_i
$$

$$
\equiv \int_{-\infty}^{\infty} g(y_{1it}, y_{2it}, x_{it}, \nu_i) d\nu_i
$$

This integral can be approximated with M-point Gauss–Hermite quadrature

$$
\int_{-\infty}^{\infty} e^{-x^2} h(x) dx \approx \sum_{m=1}^{M} w_m^* h(a_m^*)
$$

This is equivalent to

$$
\int_{-\infty}^{\infty} f(x) dx \approx \sum_{m=1}^{M} w_m^* \exp\left\{(a_m^*)^2\right\} f(a_m^*)
$$

where the w_m^* denote the quadrature weights and the a_m^* denote the quadrature abscissas. The log likelihood, L, is the sum of the logs of the panel-level likelihoods l_i.

The default approximation of the log likelihood is by adaptive Gauss–Hermite quadrature, which approximates the panel-level likelihood with

$$
l_i \approx \sqrt{2}\widehat{\sigma}_i \sum_{m=1}^{M} w_m^* \exp\left\{(a_m^*)^2\right\} g(y_{1it}, y_{2it}, x_{it}, \sqrt{2}\widehat{\sigma}_i a_m^* + \widehat{\mu}_i)
$$

where $\widehat{\sigma}_i$ and $\widehat{\mu}_i$ are the adaptive parameters for panel i. Therefore, using the definition of $g(y_{1it}, y_{2it}, x_{it}, \nu_i)$, the total log likelihood is approximated by

$$L \approx \sum_{i=1}^{n} w_i \log \left[\sqrt{2}\widehat{\sigma}_i \sum_{m=1}^{M} w_m^* \exp\{(a_m^*)^2\} \frac{\exp\{-(\sqrt{2}\widehat{\sigma}_i a_m^* + \widehat{\mu}_i)^2/2\sigma_\nu^2\}}{\sqrt{2\pi}\sigma_\nu} \right.$$

$$\left. \prod_{t=1}^{n_i} F(y_{1it}, y_{2it}, x_{it}\boldsymbol{\beta} + \sqrt{2}\widehat{\sigma}_i a_m^* + \widehat{\mu}_i) \right]$$

where w_i is the user-specified weight for panel i; if no weights are specified, $w_i = 1$.

The default method of adaptive Gauss–Hermite quadrature is to calculate the posterior mean and variance and use those parameters for $\widehat{\mu}_i$ and $\widehat{\sigma}_i$ by following the method of Naylor and Smith (1982), further discussed in Skrondal and Rabe-Hesketh (2004). We start with $\widehat{\sigma}_{i,0} = 1$ and $\widehat{\mu}_{i,0} = 0$, and the posterior means and variances are updated in the kth iteration. That is, at the kth iteration of the optimization for l_i we use

$$l_{i,k} \approx \sum_{m=1}^{M} \sqrt{2}\widehat{\sigma}_{i,k-1} w_m^* \exp\{a_m^*)^2\} g(y_{1it}, y_{2it}, x_{it}, \sqrt{2}\widehat{\sigma}_{i,k-1} a_m^* + \widehat{\mu}_{i,k-1})$$

Letting

$$\tau_{i,m,k-1} = \sqrt{2}\widehat{\sigma}_{i,k-1} a_m^* + \widehat{\mu}_{i,k-1}$$

$$\widehat{\mu}_{i,k} = \sum_{m=1}^{M} (\tau_{i,m,k-1}) \frac{\sqrt{2}\widehat{\sigma}_{i,k-1} w_m^* \exp\{(a_m^*)^2\} g(y_{1it}, y_{2it}, x_{it}, \tau_{i,m,k-1})}{l_{i,k}}$$

and

$$\widehat{\sigma}_{i,k} = \sum_{m=1}^{M} (\tau_{i,m,k-1})^2 \frac{\sqrt{2}\widehat{\sigma}_{i,k-1} w_m^* \exp\{(a_m^*)^2\} g(y_{1it}, y_{2it}, x_{it}, \tau_{i,m,k-1})}{l_{i,k}} - (\widehat{\mu}_{i,k})^2$$

and this is repeated until $\widehat{\mu}_{i,k}$ and $\widehat{\sigma}_{i,k}$ have converged for this iteration of the maximization algorithm. This adaptation is applied on every iteration until the log-likelihood change from the preceding iteration is less than a relative difference of 1e–6; after this, the quadrature parameters are fixed.

The log likelihood can also be calculated by nonadaptive Gauss–Hermite quadrature, the `int-method(ghermite)` option:

$$L = \sum_{i=1}^{n} w_i \log f \{(y_{1i1}, y_{2i1}), \ldots, (y_{1in_i}, y_{2in_i}) | \mathbf{x}_{1i}, \ldots, \mathbf{x}_{in_i}\}$$

$$\approx \sum_{i=1}^{n} w_i \log \left\{ \frac{1}{\sqrt{\pi}} \sum_{m=1}^{M} w_m^* \prod_{t=1}^{n_i} F\left(y_{1it}, y_{2it}, \mathbf{x}_{it}\boldsymbol{\beta} + \sqrt{2}\sigma_\nu a_m^*\right) \right\}$$

Both quadrature formulas require that the integrated function be well approximated by a polynomial of degree equal to the number of quadrature points. The number of periods (panel size) can affect whether

$$\prod_{t=1}^{n_i} F(y_{1it}, y_{2it}, \mathbf{x}_{it}\boldsymbol{\beta} + \nu_i)$$

is well approximated by a polynomial. As panel size and ρ increase, the quadrature approximation can become less accurate. For large ρ, the random-effects model can also become unidentified. Adaptive quadrature gives better results for correlated data and large panels than nonadaptive quadrature; however, we recommend that you use the quadchk command (see [XT] **quadchk**) to verify the quadrature approximation used in this command, whichever approximation you choose.

References

Naylor, J. C., and A. F. M. Smith. 1982. Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society, Series C* 31: 214–225.

Neuhaus, J. M. 1992. Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research* 1: 249–273.

Pendergast, J. F., S. J. Gange, M. A. Newton, M. J. Lindstrom, M. Palta, and M. R. Fisher. 1996. A survey of methods for analyzing clustered binary response data. *International Statistical Review* 64: 89–118.

Skrondal, A., and S. Rabe-Hesketh. 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models.* Boca Raton, FL: Chapman & Hall/CRC.

Also see

[XT] **xtintreg postestimation** — Postestimation tools for xtintreg

[XT] **quadchk** — Check sensitivity of quadrature approximation

[XT] **xtreg** — Fixed-, between-, and random-effects and population-averaged linear models

[XT] **xtset** — Declare data to be panel data

[XT] **xttobit** — Random-effects tobit models

[R] **intreg** — Interval regression

[R] **tobit** — Tobit regression

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