churdle — Cragg hurdle regression

**Description**

churdle fits a linear or exponential hurdle model for a bounded dependent variable. The hurdle model combines a selection model that determines the boundary points of the dependent variable with an outcome model that determines its nonbounded values. Separate independent covariates are permitted for each model.

**Quick start**

- Linear hurdle model of \( y_1 \) on \( x_1 \) and \( x_2 \), specifying that \( y_1 \) is truncated at 0 with \( x_1 \) and \( x_3 \) predicting selection
  
  ```stata
  churdle linear y1 x1 x2, select(x1 x3) ll(0)
  ```

- Add an upper truncation limit of 40
  
  ```stata
  churdle linear y1 x1 x2, select(x1 x3) ll(0) ul(40)
  ```

- As above, with the upper truncation limit specified in `trunc`
  
  ```stata
  churdle linear y1 x1 x2, select(x1 x3) ll(0) ul(trunc)
  ```

- As above, and use \( x_3 \) to model the variance of the selection model
  
  ```stata
  churdle linear y1 x1 x2, select(x1 x3, het(x3)) ll(0) ul(trunc)
  ```

- As above, and use \( x_4 \) to model the variance of the outcome model
  
  ```stata
  churdle linear y1 x1 x2, select(x1 x3, het(x3)) ll(0) ul(trunc) het(x4)
  ```

- Exponential hurdle model of \( y_2 \) on \( x_1 \) and \( x_2 \), specifying that \( y_2 \) is truncated at 4 with \( x_1 \) and \( x_3 \) predicting selection
  
  ```stata
  churdle exponential y2 x1 x2, select(x1 x3) ll(4)
  ```

**Menu**

- Statistics > Linear models and related > Hurdle regression
Syntax

Basic syntax

churdle linear depvar, select(varlist) {ll(...) | ul(...)}

churdle exponential depvar, select(varlist) ll(...)

Full syntax for churdle linear

churdle linear depvar [indepvars] [if] [in] [weight],
    select(varlist[, noconstant het(varlist_o)])
    {ll(# | varname) | ul(# | varname)} [options]

Full syntax for churdle exponential

churdle exponential depvar [indepvars] [if] [in] [weight],
    select(varlist[, noconstant het(varlist_o)]) ll(# | varname) [options]

options Description

Model

*select() specify independent variables and options for selection model
‡ll(# | varname) lower truncation limit
‡ul(# | varname) upper truncation limit
noconstant suppress constant term
constraints(constraints) apply specified linear constraints
het(varlist) specify variables to model the variance

SE/Robust

vce(vcetype) vcetype may be oim, robust, cluster clustvar, bootstrap, or jackknife

Reporting

_level(#) set confidence level; default is level(95)
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

Maximization

maximize_options control the maximization process; seldom used
coeflegend display legend instead of statistics
*select() is required.

The full specification is \texttt{select(varlist, noconstant het(varlist)}}.

\texttt{noconstant} specifies that the constant be excluded from the selection model.

\texttt{het(varlist)}} specifies the variables in the error-variance function of the selection model.

†You must specify at least one of \texttt{ul(# | varname)} or \texttt{ll(# | varname)} for the linear model and must specify \texttt{ll(# | varname)} for the exponential model.

\texttt{indepvars}, \texttt{varlist}, and \texttt{varlist}} may contain factor variables; see \texttt{[U] 11.4.3 Factor variables}.

\texttt{bootstrap}, \texttt{by}, \texttt{fp}, \texttt{jackknife}, \texttt{rolling}, \texttt{statsby}, and \texttt{svy} are allowed; see \texttt{[U] 11.1.10 Prefix commands}.

Weights are not allowed with the \texttt{bootstrap} prefix; see \texttt{[R] bootstrap}.

\texttt{vce()} and weights are not allowed with the \texttt{svy} prefix; see \texttt{[SVY] svy}.

\texttt{fweights}, \texttt{iweights}, and \texttt{pweights} are allowed; see \texttt{[U] 11.1.6 weight}.

\texttt{coeflegend} does not appear in the dialog box.

See \texttt{[U] 20 Estimation and postestimation commands} for more capabilities of estimation commands.

Options

\texttt{Model}

\texttt{select(varlist, noconstant het(varlist)}} specifies the variables and options for the selection model. \texttt{select()} is required.

\texttt{ll(# | varname)} and \texttt{ul(# | varname)} indicate the lower and upper limits, respectively, for the dependent variable. You must specify one or both for the linear model and must specify a lower limit for the exponential model. Observations with \texttt{depvar} \leq \texttt{ll()} have a lower bound; observations with \texttt{depvar} \geq \texttt{ul()} have an upper bound; and the remaining observations are in the continuous region.

\texttt{noconstant}, \texttt{constraints(constraints)}; see \texttt{[R] Estimation options}.

\texttt{het(varlist)} specifies the variables in the error-variance function of the outcome model.

\texttt{SE/Robust}

\texttt{vce(vcetype)} specifies the type of standard error reported, which includes types that are derived from asymptotic theory (\texttt{oim}), that are robust to some kinds of misspecification (\texttt{robust}), that allow for intragroup correlation (\texttt{cluster, clustvar}), and that use bootstrap or jackknife methods (\texttt{bootstrap, jackknife}); see \texttt{[R] vce_option}.

\texttt{Reporting}

\texttt{level(#), nocnsreport}; see \texttt{[R] Estimation options}.

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

\texttt{Maximization}

\texttt{maximize_options}: \texttt{difficult, technique(algorithm_spec), iterate(#), [no]log, trace, gradient, showstep, hessian, showtolerance, tolerance(#), ltolerance(#), nrtolerance(#), nonrtolerance, and from(init_specs)}; see \texttt{[R] Maximize}. These options are seldom used.
The following option is available with churdle but is not shown in the dialog box: coeflegend; see [R] Estimation options.

Remarks and examples

churdle fits a linear or an exponential hurdle model. It combines a selection model that determines the boundary points of the dependent variable with an outcome model that determines its nonbounded values. Hurdle models treat these boundary values as observed instead of censored. That is to say, observations where the dependent variable is equal to one of the boundary values are not the result of our inability to observe the distribution above or below a certain point; see Wooldridge (2010) chapter 17 for a thorough discussion of this point.

These models were proposed by Cragg (1971) to explain the demand for durable goods. In the Cragg model, individuals purchase zero or a positive amount of the durable good, with different factors determining each of these choices. This may be generalized to other individual decisions, such as money donated to charity, cigarette consumption, and time spent volunteering.

Hurdle models are characterized by the relationship $y_i = s_i h_i^*$, where $y_i$ is the observed value of the dependent variable.

The selection variable, $s_i$, is 1 if the dependent variable is not bounded and 0 otherwise. In the Cragg model, the lower limit that binds the dependent variable is 0 so the selection model is

$$s_i = \begin{cases} 
1 & \text{if } z_i \gamma + \epsilon_i > 0 \\
0 & \text{otherwise}
\end{cases}$$

where $z_i$ is a vector of explanatory variables, $\gamma$ is a vector of coefficients, and $\epsilon_i$ is a standard normal error term. churdle allows a different lower limit to be specified in ll() and, for the linear model, an upper limit in ul(). Conditional heteroskedasticity of the random error $\epsilon_i$ is allowed if suboption het() is specified in select().

The continuous latent variable $h_i^*$ is observed only if $s_i = 1$. The outcome model can be either the linear model or the exponential model, as proposed in Cragg (1971):

$$h_i^* = x_i \beta + \nu_i \quad \text{(linear)}$$

$$h_i^* = \exp(x_i \beta + \nu_i) \quad \text{(exponential)}$$

where $x_i$ is a vector of explanatory variables, $\beta$ is a vector of coefficients, and $\nu_i$ is an error term.

For the linear model, $\nu_i$ has a truncated normal distribution with lower truncation point $-x_i \beta$. For the exponential model, $\nu_i$ has a normal distribution. churdle extends the Cragg hurdle models to allow for conditional heteroskedasticity of the random error $\nu_i$ if the user specifies the het() option.

The parameters and regressors in the models for $h_i^*$ and for $s_i$ may differ.
Example 1: Linear hurdle model

Consider a dataset that contains the number of hours an individual exercises per day (hours), their age (age), whether they are single (single), hours they work per day (whours), whether they smoke (smoke), their weight in kilograms (weight), their distance from the nearest gym (distance), and their average commute from work (commute).

Figure 1 shows that 43.9% of the individuals in the sample do not exercise and that the hours exercised varies among individuals that decide to exercise.

We model the decision to exercise or not as a function of commute, whours, and age. These variables are written in select(). Once a decision to exercise is made, the time an individual exercises is modeled as a linear function of age, smoke, distance, and single.
. use https://www.stata-press.com/data/r16/fitness
. churdle linear hours age i.smoke distance i.single,
> select(commute whours age) ll(0)
Iteration 0:  log likelihood = -23657.236
Iteration 1:  log likelihood = -23344.182
Iteration 2:  log likelihood = -23340.051
Iteration 3:  log likelihood = -23340.044
Iteration 4:  log likelihood = -23340.044
Cragg hurdle regression
Number of obs = 19,831
LR chi2(4) = 9059.26
Prob > chi2 = 0.0000
Log likelihood = -23340.044
Pseudo R2 = 0.1625

| Coef.  | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|--------|-----------|-----|------|---------------------|
| hours  | .0015116  | .000763 | 1.98  | 0.048 | .0000162 .003007 |
| age    |           |        |       |        |                    |
| smoke  | -1.06646  | .0460578 | -23.15  | 0.000 | -1.156731 -.9761879 |
| smoking|           |        |       |        |                    |
| distance| -.1333868 | .0126344 | -10.56  | 0.000 | -.1581497 -.1086238 |
| single |           |        |       |        |                    |
| single | .9940893  | .0258775 | 38.42  | 0.000 | .9433703 1.044808 |
| _cons  | .9138855  | .0396227 | 23.06  | 0.000 | .8362264 .9915447 |

| Coef.  | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|--------|-----------|-----|------|---------------------|
| selection_ll | -0.2953345 | 0.0624665 | -4.73  | 0.000 | -.4177666 -.1729024 |
| commute |           |        |       |        |                    |
| whours  | 0.0022974  | 0.0069306 | 0.33  | 0.740 | -0.0112864 0.0158811 |
| age     | -0.0485347  | 0.0006501 | -74.65  | 0.000 | -.049809 -.0472604 |
| _cons   | 2.649945   | 0.0499795 | 53.02  | 0.000 | 2.551987 2.747903 |

| Coef.  | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|--------|-----------|-----|------|---------------------|
| lnsigma | 0.0083199  | 0.0099648 | 0.83  | 0.404 | -0.0112170 .0278506 |
| _cons  |           |        |       |        |                    |
| /sigma | 1.008355   | 0.010048 | 98.87  | 0.000 | .9888519 1.028242 |

The coefficients in the outcome model for the latent variable appear under `hours`. Because we only specified a lower limit to bind the dependent variable, the output shows parameter estimates for a single selection model under `selection_ll`. Information about the estimated standard deviation of the error term in the outcome model appears under `lnsigma` and `/sigma`.

The coefficient estimates are not directly interpretable. To obtain the effect of a covariate on the model, we need to use the `margins` command; see `[R] churdle postestimation`. Consider the effect of `age`:
. margins, dydx(age)

Average marginal effects
Model VCE : OIM
Expression : Conditional mean estimates of dependent variable, predict()
dy/dx w.r.t. : age

| Delta-method | dy/dx | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|--------------|-------|-----------|-----|------|----------------------|
| age          | -.0216855 | .000289  | -75.03 | 0.000 | -.022252 -.021119 |

Each additional year of age is associated with about −0.02 fewer hours, or 1.2 minutes, of exercise.

Example 2: Linear hurdle with models for the outcome and selection variances

In this example, we illustrate the possibility of fitting a heteroskedastic probit for the selection and latent model. In both cases, this is done by specifying age and single as the variables that affect the conditional variance. As in example 1, we have separate parameters for the outcome model and lower-limit selection model.

. churdle linear hours age i.smoke distance i.single, > select(commute whours age, het(age single)) ll(0) het(age single) nolog

Cragg hurdle regression
Number of obs = 19,831
LR chi2(4) = 9060.63
Prob > chi2 = 0.0000
Log likelihood = -23339.355 Pseudo R2 = 0.1626

| hours          | Coef. | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|----------------|-------|-----------|-----|------|----------------------|
| hours          |       |           |     |      |                      |
| age            | .0012559 | .0008198 | 1.53 | 0.126 | -.0003508 .0028626 |
| smoke          |       |           |     |      |                      |
| smoking        | -1.065564 | .0457657 | -23.28 | 0.000 | -1.155263 -.9758649 |
| distance       | -.1332939 | .0126102 | -10.57 | 0.000 | -.1580094 -.1085783 |
| single         |       |           |     |      |                      |
| single         | 1.002511 | .032535  | 30.81 | 0.000 | .9387436 1.066278 |
| _cons          | .9166356 | .0388318 | 23.61 | 0.000 | .8405268 .9927445 |
| selection_ll   |       |           |     |      |                      |
| commute        | -.2959986 | .0641594 | -4.61 | 0.000 | -.4217488 -.1702484 |
| whours         | .0024514 | .0069769 | 0.35 | 0.725 | -.0112231 .0161259 |
| age            | -.048886 | .0021405 | -22.84 | 0.000 | -.0530814 -.0446806 |
| _cons          | 2.669613 | .1139478 | 23.43 | 0.000 | 2.446282 2.892947 |
| lnsigma        |       |           |     |      |                      |
| age            | .0003537 | .0004026 | 0.88 | 0.380 | -.0004354 .0011427 |
| single         | -.0080667 | .019253 | -0.42 | 0.675 | -.0458019 .0296685 |
| lnsigma_ll     |       |           |     |      |                      |
| age            | -.0002035 | .0008424 | -0.24 | 0.809 | -.0018546 .0014475 |
| single         | .0268271 | .0270133 | 0.99 | 0.321 | -.0261179 .0797721 |
The coefficients on age and single have no effect on the conditional variance of the outcome model or on the conditional variance of the selection model. Thus, there is no evidence that the variance depends on age and marital status.

Example 3: Exponential hurdle model

Returning to example 1, if we believe that the conditional mean of the latent variable has an exponential form instead of a linear form, we use *churdle exponential*.

```
. churdle exponential hours age i.smoke distance i.single, > select(commute whours age) ll(0) nolog
Cragg hurdle regression
Number of obs = 19,831
LR chi2(4) = 8663.21
Prob > chi2 = 0.0000
Log likelihood = -15666.195 Pseudo R2 = 0.2166

Coef.  Std. Err.  z    P>|z|     [95% Conf. Interval]
---------  ---------  ------  --------  ------------------
hours
  age       .0008368  .0005341  1.57  0.117  -.00021    .0018836
smoke
  smoking  -.6431348  .0258509 -24.88  0.000  -.6938016  -.592468
  distance -.0772879  .0079132  -9.77  0.000  -.0927976  -.0617783
single
  single     .5975111  .0161080  37.09  0.000   .5659401   .6290821
  _cons     -.0770619  .0254833  -3.02  0.002  -.1270082   -.0271157
selection_ll
  commute  -.2953345  .0624665  -4.73  0.000  -.4177666  -.1729024
  whours   .0022974  .0069306   0.33  0.740  -.0112864   .0158811
  age      -.0485347  .0006501 -74.65  0.000  -.049809   -.0472604
  _cons    2.649945   .0499795  53.02  0.000   2.551987   2.747903
lnsigma
  _cons   -.186917    .0067067  -27.87  0.000  -.200062  -.1737721
/sigma  .8295126    .0055633   .81868  .8404884
```

What was said previously regarding the interpretation of the effects of the different regressors also holds true for *churdle exponential*. We again use margins to estimate the effect of age on time spent exercising.

```
. margins, dydx(age)
Average marginal effects       Number of obs  =  19,831
Model VCE    : OIM
Expression : Conditional mean estimates of dependent variable, predict()
dy/dx w.r.t.  : age

  Delta-method
            dy/dx  Std. Err.     z  P>|z|     [95% Conf. Interval]
--- -------------  --------  -------  ------  ------------------
age    -.0245582   .0004805 -51.11  0.000  -.0255   -.0236164
```
With the exponential outcome model of the latent variable, our estimate is that each additional year of age decreases exercise time by about 0.025 hours, or 1.5 minutes.

## Stored results

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

**Scalars**
- `e(N)` number of observations
- `e(k_eq_model)` number of equations in overall model test
- `e(df_m)` model degrees of freedom
- `e(r2_p)` pseudo-$R$-squared
- `e(chi2)` $\chi^2$
- `e(ll)` log likelihood
- `e(ll_0)` log likelihood, constant-only model
- `e(chi2type)` Wald or LR; type of model $\chi^2$ test
- `e(p)` $p$-value for model test
- `e(rank)` rank of `e(V)`
- `e(converged)` 1 if converged, 0 otherwise

**Macros**
- `e(cmd)` `churdle`
- `e(cmdline)` command as typed
- `e(estimator)` linear or exponential
- `e(model)` Linear or Exponential
- `e(wtype)` weight type
- `e(wexp)` weight expression
- `e(vce)` vcetype specified in `vce()`
- `e(vcetype)` title used to label Std. Err.
- `e(opt)` type of optimization
- `e(which)` max or min; whether optimizer is to perform maximization or minimization
- `e(properties)` `b V`
- `e(predict)` program used to implement `predict`
- `e(marginsnotok)` predictions disallowed by `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(i)` iteration log
- `e(V)` variance–covariance matrix of the estimators
- `e(V_modelbased)` model-based variance

**Functions**
- `e(sample)` marks estimation sample
Methods and formulas

Let $\ell \ell$ refer to the lower limit and $u \ell$ to the upper limit. Also let the probabilities of being at these limits be given by

$$
Pr(y_i = \ell \ell | z_i) = \Phi (\ell \ell - z'_i \gamma_{\ell \ell})
$$

$$
Pr(y_i = u \ell | z_i) = \Phi (z'_i \gamma_{u \ell} - u \ell)
$$

where $z_i$ are the covariates of the selection model for individual $i$, which may be distinct from the covariates $x_i$ for the latent model; $\Phi$ corresponds to the standard normal cumulative distribution function; $\gamma_{\ell \ell}$ is the parameter vector for the lower-limit selection model; and $\gamma_{u \ell}$ is the parameter vector for the upper-limit selection model.

Under the assumptions that $\nu_i$ has a truncated normal distribution with lower truncation point $\ell \ell - x'_i \beta$ and upper truncation point $u \ell - x'_i \beta$ and has a homoskedastic variance, the log-likelihood function is given by

$$
\ln L = \sum_{i=1}^{n} (y_i \leq \ell \ell) \log \Phi (\ell \ell - z'_i \gamma_{\ell \ell}) + (y_i > \ell \ell) \log \{1 - \Phi (u \ell - z'_i \gamma_{u \ell})\}
$$

$$
+ (u \ell > y_i > \ell \ell) \left[ \log \left\{ \Phi (u \ell - z'_i \gamma_{u \ell}) - \Phi (\ell \ell - z'_i \gamma_{\ell \ell}) \right\} \right]
$$

$$
- (u \ell > y_i > \ell \ell) \left[ \log \left\{ \Phi \left( \frac{u \ell - x'_i \beta}{\sigma} \right) - \Phi \left( \frac{\ell \ell - x'_i \beta}{\sigma} \right) \right\} \right]
$$

$$
+ (u \ell > y_i > \ell \ell) \left[ \log \left\{ \phi \left( \frac{y_i - x'_i \beta}{\sigma} \right) \right\} - \log(\sigma) \right]
$$

Without the homoskedasticity assumption, the heteroskedasticity can be modeled using the form $\sigma^2 (w_i) = \exp (2 w'_i \theta)$, where $w_i$ are the variables that affect the conditional variance of $\nu_i$. The log-likelihood function is obtained by replacing $\sigma$ with $\exp (w'_i \theta)$.

The log-likelihood function for the exponential model is given by

$$
\ln L = \sum_{i=1}^{n} (y_i \leq \ell \ell) \log \Phi (\ell \ell - z'_i \gamma_{\ell \ell}) + (y_i > \ell \ell) \left[ \log \left\{ \Phi (u \ell - z'_i \gamma_{u \ell}) \right\} \right]
$$

$$
+ (y_i > \ell \ell) \left\{ \log \phi \left( \frac{\log(y_i - \ell \ell) - x'_i \beta}{\sigma} \right) \right\} - \log(\sigma) - \log(y_i - \ell \ell) \right] \right.
$$

Analogous to the linear case, we can model heteroskedasticity by $\sigma^2 (w_i) = \exp (2 w'_i \theta)$.

Estimation of both of the aforementioned likelihood functions is done by maximum likelihood.

References


Also see

[R] churdle postestimation — Postestimation tools for churdle

[R] intreg — Interval regression

[R] tobit — Tobit regression

[SVY] svy estimation — Estimation commands for survey data

[U] 20 Estimation and postestimation commands