lasso postestimation — Postestimation tools for lasso for prediction

Postestimation commands | predict | stcurve | Remarks and examples
Methods and formulas | References | Also see

Postestimation commands

The following postestimation commands are of special interest after `lasso`, `sqrtlasso`, and `elasticnet`:

Command	Description
`bicplot`	plot Bayesian information criterion function
`coefpath`	plot path of coefficients
`cvplot`	plot cross-validation function
`lassocoef`	display selected coefficients
`lassogof`	goodness of fit after lasso for prediction
`lassoinfo`	information about lasso estimation results
`lassoknots`	knot table of coefficient selection and measures of fit
`lassoselect`	select alternative λ^* (and α^* for `elasticnet`)
*`stcurve`	plot the survivor, failure, hazard, or cumulative hazard function

*`stcurve` is appropriate only after `lasso cox` or `elasticnet cox`.

The following standard postestimation commands are also available:

Command	Description
`estat summarize`	summary statistics for the estimation sample
`estimates`	cataloging estimation results
`etable`	table of estimation results
`predict`	linear predictions

predict

Description for predict

`predict` creates a new variable containing predictions such as linear predictions; probabilities when the model is logit or probit; number of events when the model is Poisson; or hazard ratios and baseline survivor, cumulative hazard, and hazard functions when the model is Cox.

Menu for predict

Statistics > Postestimation

Syntax for predict

`predict` [*type*] *newvar* [*if*] [*in*] [, *statistic options*]

statistic	Description
Main	
`xb`	linear predictions; the default for the `linear` model
`pr`	probability of a positive outcome; the default for the `logit` and `probit` models
`n`	number of events; the default for the `poisson` model
`ir`	incidence rate; optional for the `poisson` model
`hr`	predicted hazard ratio, also known as the relative hazard; the default for the `cox` model
`basesurv`	baseline survivor function
`basechazard`	baseline cumulative hazard function
`basehc`	baseline hazard contributions

`pr` is allowed only when the model is `logit` or `probit`.

`n` and `ir` are allowed only when the model is `poisson`.

`hr`, `basesurv`, `basechazard`, and `basehc` are allowed only when the model is `cox`.

options	Description
Main	
`penalized`	use penalized coefficients; the default
`postselection`	use postselection (unpenalized) coefficients
`nooffset`	ignore the offset or exposure variable (if any)

Unstarred statistics are available both in and out of sample; type `predict ... if e(sample) ...` if wanted only for the estimation sample. Starred statistics are calculated only for the estimation sample, even when `e(sample)` is not specified. `nooffset` is allowed only with unstarred statistics.

Options for predict

Main

`xb`, the default for the `linear` model, calculates linear predictions.

`pr`, the default for and only allowed with the `logit` and `probit` models, calculates the probability of a positive event.

`n`, the default for and only allowed with the `poisson` model, calculates the number of events, which is $\exp(\mathbf{x}_j\boldsymbol{\beta})$ if neither `offset()` nor `exposure()` was specified when the model was fit; $\exp(\mathbf{x}_j\boldsymbol{\beta} + \text{offset}_j)$ if `offset()` was specified; or $\exp(\mathbf{x}_j\boldsymbol{\beta}) \times \text{exposure}_j$ if `exposure()` was specified.

`ir` applies to the `poisson` model only. It calculates the incidence rate $\exp(\mathbf{x}\boldsymbol{\beta}')$, which is the predicted number of events when exposure is 1. Specifying `ir` is equivalent to specifying `n` when neither `offset()` nor `exposure()` was specified when the model was fit.

`hr`, the default for the `cox` model, calculates the relative hazard (hazard ratio), that is, the exponentiated linear prediction $\exp(\mathbf{x}\boldsymbol{\beta}')$.

`basesurv` applies to the `cox` model only. It calculates the baseline survivor function. In the null model, this is equivalent to the Kaplan–Meier product-limit estimate.

`basechazard` applies to the `cox` model only. It calculates the cumulative baseline hazard.

`basehc` applies to the `cox` model only. It calculates the baseline hazard contributions. These are used to construct the product-limit type estimator for the baseline survivor function generated by `basesurv`.

`penalized` specifies that penalized coefficients be used to calculate predictions. This is the default. Penalized coefficients are those estimated by lasso in the calculation of the lasso penalty. See *Methods and formulas* in [LASSO] **lasso**.

`postselection` specifies that postselection coefficients be used to calculate predictions. Postselection coefficients are calculated by taking the variables selected by lasso and refitting the model with the appropriate ordinary estimator: linear regression for `linear` models, logistic regression for `logit` models, probit regression for `probit` models, Poisson regression for `poisson` models, and Cox regression for `cox` models.

`nooffset` is relevant only if you specified `offset()` or `exposure()` when you fit the model. It modifies the calculations made by `predict` so that they ignore the offset or exposure variable; the linear prediction is treated as $\mathbf{x}\boldsymbol{\beta}'$ rather than $\mathbf{x}\boldsymbol{\beta}' + \text{offset}$ or $\mathbf{x}\boldsymbol{\beta}' + \ln(\text{exposure})$. For the `poisson` model, specifying `predict ..., nooffset` is equivalent to specifying `predict ..., ir`. This option is not allowed when `basesurv`, `basechazard`, or `basehc` is specified.

stcurve

Description for stcurve

`stcurve` plots the survivor, failure, hazard, or cumulative hazard function after `lasso cox` or `elasticnet cox`.

Menu for stcurve

Statistics > Survival analysis > Regression models > Plot survivor or related function

Syntax for stcurve

`stcurve` [, `penalized` `postselection` *stcurve_options*]

Options for stcurve

`penalized`, the default, specifies that penalized coefficients be used to calculate predictions. Penalized coefficients are those estimated by lasso in the calculation of the lasso penalty. See *Methods and formulas* in [LASSO] **lasso**.

`postselection` specifies that postselection coefficients be used to calculate predictions. Postselection coefficients are calculated by taking the variables selected by lasso and refitting the model with `stcox`.

stcurve_options are options available for `stcurve`; see *Options* in [ST] **stcurve**.

Remarks and examples

stata.com

By default, `predict` after `lasso` uses the penalized coefficient estimates to predict the outcome. Specifying the `postselection` option causes `predict` to use the postselection coefficients to calculate predictions. Postselection coefficients are calculated by taking the variables selected by lasso and refitting the model with the unpenalized estimator.

`stcurve` after `lasso cox` or `elasticnet cox` also uses the penalized coefficients by default. Specifying the `postselection` option causes `stcurve` to use the postselection coefficients.

Belloni and Chernozhukov (2013) and Belloni et al. (2012) provide results under which predictions using postselection coefficients perform at least as well as predictions using penalized coefficients. Their results are only for linear models. Their conditions essentially limit the cases to ones in which the covariates selected by the lasso are close to the set of covariates that best approximates the outcome. Said plainly, this means that under the conditions for which lasso provides valid predictions, the postselection coefficients should do slightly better than the penalized coefficients in most cases; in other cases, they should be about the same.

Rather than relying on theorems, standard practice in prediction applications uses split-sample techniques to find which of several models produces the best predictions. One standard practice in prediction applications is to randomly split the sample into training and testing samples. When you use the training data, the coefficients for several competing predictors are computed. When you use the testing data, an out-of-sample prediction error is computed for each of the predictors whose coefficients were estimated on the training data. The predictor with the smallest out-of-sample prediction error is preferred. This practice is illustrated in [LASSO] **lassogof**.

Methods and formulas

Below, we discuss the methods and formulas for the predictions of baseline survivor function, baseline cumulative hazard function, and baseline hazard contributions after `lasso cox` or `elasticnet cox`.

Define $z_i = \mathbf{x}_i\widehat{\boldsymbol{\beta}}' + \text{offset}_i$, where $\widehat{\boldsymbol{\beta}}$ is either the penalized or the postselection coefficients. The estimated baseline hazard contribution is obtained at each failure time as $h_j = 1 - \widehat{\alpha}_j$, where $\widehat{\alpha}_j$ is the solution to

$$\sum_{k\in D_j} \frac{\exp(z_k)}{1-\widehat{\alpha}_j^{\exp(z_k)}} = \sum_{\ell\in R_j} \exp(z_\ell)$$

(Kalbfleisch and Prentice 2002, eq. 4.34, 115), where j indexes the ordered failure times t_j ($j = 1,\ldots,D$); D_j is the set of d_j observations that fail at t_j; d_j is the number of failures at t_j; and R_j is the set of observations k that are at risk at time t_j (that is, all k such that $t_{0k} < t_j \le t_k$, and t_{0k} is the entry time for the kth observation).

The estimated baseline survivor function is

$$\widehat{S}_0(t) = \prod_{j:t_j\le t} \widehat{\alpha}_j$$

The estimated baseline cumulative hazard function, if requested, is related to the baseline survivor function calculation; yet the values of $\widehat{\alpha}_j$ are set at their starting values and are not iterated. Equivalently,

$$\widehat{H}_0(t) = \sum_{j:t_j\le t} \frac{d_j}{\sum_{\ell\in R_j}\exp(z_\ell)}$$

For an application of this formula in the context of `lasso cox`, see Ternès, Rotolo, and Michiels (2017).

References

Belloni, A., D. Chen, V. Chernozhukov, and C. B. Hansen. 2012. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80: 2369–2429. https://doi.org/10.3982/ECTA9626.

Belloni, A., and V. Chernozhukov. 2013. Least squares after model selection in high-dimensional sparse models. *Bernoulli* 19: 521–547. https://doi.org/10.3150/11-BEJ410.

Kalbfleisch, J. D., and R. L. Prentice. 2002. *The Statistical Analysis of Failure Time Data*. 2nd ed. New York: Wiley.

Ternès, N., F. Rotolo, and S. Michiels. 2017. Robust estimation of the expected survival probabilities from high-dimensional Cox models with biomarker-by-treatment interactions in randomized clinical trials. *BMC Medical Research Methodology* 17(83). https://doi.org/10.1186/s12874-017-0354-0.

Also see

[LASSO] **lasso examples** — Examples of lasso for prediction

[LASSO] **elasticnet** — Elastic net for prediction and model selection

[LASSO] **lasso** — Lasso for prediction and model selection

[LASSO] **sqrtlasso** — Square-root lasso for prediction and model selection

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