Title

Collinear covariates — Treatment of collinear covariates

Description   Remarks and examples   Also see

Description

Lasso, square-root lasso, and elastic net treat collinear covariates differently from traditional estimators. With these models, you specify variables that might be included in the model, and they choose the variables to be included. When you specify those variables, it is important that you present them with all possible alternatives. This means that, when including factor variables, you must include the full collinear set of indicators.

If you use Stata's factor-variable notation, it is handled automatically for you. If you create indicator variables for yourself, you must create and include them all.

Remarks and examples

Remarks are presented under the following headings:

Summary

Consider factor variable `group` that takes on the values 1, 2, and 3. If you type

```
. lasso linear y i.group ...
```

`lasso` will know that separate covariates for `group` 1, 2, and 3 are to be included among the variables to be potentially included in the model.

If you create your own indicator variables, you need to create and specify indicators for all the values of the factor variable:

```
. generate g1 = (group==1)
. generate g2 = (group==2)
. generate g3 = (group==3)
. lasso linear y g1 g2 g3 ...
```

It is important that you do not omit one of them, say, `g1`, and instead type

```
. lasso linear y g2 g3 ...
```

Explanation

With no loss of generality, we will focus on lasso for this explanation. Assume lasso has just found the best model for λ_i with k covariates and is now searching for the best model for λ_{i+1}, where $\lambda_{i+1} < \lambda_i$.

The λ_{i+1} model will not always be the same λ_i model with new covariates added, but this is often the case. (Sometimes, covariates in the λ_i model are removed.) Assume this is a case of adding only new covariates. Also assume that g1, g2, and g3 have not been chosen yet and that lasso chooses g1.

But what if we did not specify g1 among the potential covariates? What if rather than typing

```
. lasso linear y g1 g2 g3 ...
```

we typed

```
. lasso linear y g2 g3 ...
```

In that case, lasso would not choose g1 because it could not. It would choose some other covariate or covariates, perhaps g2, perhaps g3, perhaps g2 and g3, or perhaps other covariates. And lasso is on an inferior path because g1 was not among the potential covariates.

Although selecting both g2 and g3 in place of g1 gives an equivalent model for prediction, it may have wasted an extra penalty on the coefficients for g2 and g3. A model with only g1 may have a smaller penalty and allow other covariates to be included, which a model with g2 and g3 would not. By eliminating g1, we have denied lasso the opportunity to find a more parsimonious model.

Applies to inferential commands

You must also specify full collinear sets of potential covariates with the inferential commands. Specify full sets in the controls() option, such as

```
. dsregress y z1 z2, controls(g1 g2 g3 ...)
```

Likewise for the high-dimensional instruments in poivregress and xpoivregress:

```
. poivregress y ... (z1 z2 = g1 g2 g3 ...), controls(...)
```

Just as with lasso, the issue is handled automatically if you use factor-variable notation:

```
. dsregress y z1 z2, controls(i.group ...)
```

Does not apply to alwaysvars

With any lasso, you can specify covariates that will always appear in the model. You specify them in parentheses. For example, for lasso, type

```
. lasso linear y (x1 x2) x3 x4 ...
```

and for the inference commands, type

```
. dsregress y z1 z2, controls((x1 x2) x3 x4 ...)
```

We call the covariates that always appear in the model *alwaysvars*. The *alwaysvars* do not need to be full collinear sets. Indeed, collinear variables among the *alwaysvars* will be omitted.

Factor-variable notation handles the problem automatically in both cases:

```
. lasso linear (i.region ...) i.group ...
```

A base level will be set for i.region (or you can set it explicitly). For i.group, all levels will be included. If you try to set a base level for i.group, it will be ignored.

Also see

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

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

For suggested citations, see the FAQ on citing Stata documentation.