The Stata Journal (2023)
23, Number 2, pp. 402–417 DOI: 10.1177/1536867X231175272

posw: A command for the stepwise Neyman-orthogonal estimator

David M. Drukker
Sam Houston State University
Huntsville, TX
dxd070@shsu.edu

Di Liu
StataCorp
College Station, TX
dliu@stata.com

Abstract. Inference for structural and treatment parameters while having high-dimensional covariates in the model is increasingly common. The Neyman-orthogonal (NO) estimators in Belloni, Chernozhukov, and Wei (2016, *Journal of Business and Economic Statistics* 34: 606–619) produce valid inferences for the parameters of interest while using generalized linear model lasso methods to select the covariates. Drukker and Liu (2022, *Econometric Reviews* 41: 1047–1076) extended the estimators in Belloni, Chernozhukov, and Wei (2016) by using a Bayesian information criterion stepwise method and a testing-stepwise method as the covariate selector. Drukker and Liu (2022) found a family of data-generating processes for which the NO estimator based on Bayesian information criterion stepwise produces much more reliable inferences than the lasso-based NO estimator. In this article, we describe the implementation of `posw`, a command for the stepwise-based NO estimator for the high-dimensional linear, logit, and Poisson models.

Keywords: st0713, posw, high-dimensional model, covariate selection, partialing out, stepwise, Neyman-orthogonal, generalized linear model, postselection inference

1 Introduction

Many researchers face a situation in which they want to make inferences about a few coefficients on variables of interest while having more control variables than they could include in the model for the sample size at hand. This situation is a high-dimensional model (HDM), and the sparse modeling approach is frequently applied. This approach has three key features. First, it assumes that the model is "sparse". The model is sparse when the number of potential controls that must be included in the model is small relative to the sample size. Second, the approach uses a covariate-selection method to choose which of the many potential controls must be included. Third, the approach uses an estimator for the parameters of interest that is robust to the inevitable mistakes made in the covariate-selection step. As discussed in Chernozhukov, Hansen, and Spindler (2015), these robust estimators are known as Neyman-orthogonal (NO) estimators because they extend a technique derived in Neyman (1959, 1979) and they are robust to the covariate-selection mistakes made by high-quality covariate-selection techniques.[1]

[1] A high-quality covariate-selection technique selects the required covariates at a fast enough rate; see Chernozhukov, Hansen, and Spindler (2015).

Why covariate-selection methods inevitably make mistakes deserves an explanation. Covariate selection has a long and somewhat controversial history in statistics and econometrics. When all the nonzero coefficients are large enough in magnitude and the model is sparse, some covariate-selection methods will include the required covariates with probability approaching one as the sample size increases, under some regularity conditions. These conditions require a minimum ability of the statistical model to approximate the true model and place limits on the possible joint distribution of the covariates. The assumption that the nonzero coefficients are large enough in magnitude is known as a "beta-min" assumption, and it is the crucial assumption. It is now commonly accepted that the beta-min assumption is way too strong to be a part of a realistic approach to estimation and inference.

While Leeb and Pötscher (2006, 2008) and Pötscher and Leeb (2009) contain formal results and intuition based on uniform versus pointwise results, there is a simple thought experiment that captures why the beta-min assumption is too strong. It is frequently the case in applied studies that several coefficients have *p*-values that are just a little too large to reject the null hypothesis that their true values are zero. These covariates are on the border of being included or not included in the model. The beta-min assumption would require that the true values of the coefficients on these not-included covariates be either zero or very close to zero, where the definition of "very close" is a function of the sample size. Dropping the beta-min assumption allows for a more realistic scenario in which these coefficients have nonzero values that are just a little too small in magnitude to warrant including the covariates in the model, according to the covariate-selection method.

Dropping the beta-min assumption implies that even the best possible covariate-selection methods will omit some covariates whose coefficients are small in magnitude. When we accept that covariate-selection methods make mistakes, we must use an estimation technique that is robust to these mistakes in covariate selection. Naive estimators that simply include the selected covariates in a model are not robust and do not provide reliable inferences. In repeated samples, the random inclusion or exclusion of covariates with small coefficients causes the distribution of the naive estimators to be multimodal. Using a normal distribution to approximate this multimodal distribution produces unreliable results in theory and practice. Again, see Leeb and Pötscher (2006, 2008) and Pötscher and Leeb (2009) for details.

Belloni et al. (2012), Belloni, Chernozhukov, and Hansen (2014), Chernozhukov, Hansen, and Spindler (2015), and Belloni, Chernozhukov, and Wei (2016) pioneered NO estimators that are robust to the mistakes in covariate selection made by high-quality covariate-selection methods. These theoretical studies formally demonstrated that NO estimators based on lasso methods with a feasible version of the optimal lasso tuning parameters produce valid inferences for the coefficients of interest. The feasible version of the optimal lasso tuning parameters is known as the plugin method. Drukker and Liu (2022, Forthcoming) extended the feasible version to the generalized linear model (GLM) case.

Drukker and Liu (2022) studied the finite-sample performance of the Lasso-based NO estimator for the high-dimensional GLM using different methods to select the lasso tuning parameters. The simulation results in Drukker and Liu (2022) suggest both advantages and disadvantages of using the lasso as a covariate-selection method in a NO estimator. The main advantage of the lasso-based NO estimator is speed. If the tuning parameters are appropriately chosen, the lasso-based NO estimator can relatively quickly provide valid inferences in the presence of small coefficients and many potential covariates. The main disadvantage of the lasso-based NO estimator is that the simulations in Drukker and Liu (2022) reveal a problematic family of data-generating processes (DGPs) for which the lasso-based NO estimator fails, regardless of the choice of the tuning parameter. The problematic family of DGPs has coefficients that alternate in sign. These coefficients are commonly observed in models that use powers and interaction terms among covariates to approximate nonlinear functional forms.

To accommodate this family of DGPs, Drukker and Liu (2022) extended the lasso-based NO estimator in Belloni, Chernozhukov, and Wei (2016) to the Bayesian information criterion (BIC)-stepwise-based NO estimator for the high-dimensional GLM model. The simulation results in Drukker and Liu (2022) show that the BIC-stepwise-based NO estimator performs well on the designs for which the lasso-based NO estimator failed.

The price of the increased performance was computation time. The BIC-stepwise-based NO estimator is much slower than the lasso-based NO estimators.[2] In practical terms, the BIC-stepwise-based NO estimators become computationally infeasible for huge numbers of potential covariates, which the lasso-based NO estimators can handle.

The good performance of the BIC-stepwise-based NO estimator is not without some theoretical support. Kozbur (2020) presents conditions in which a testing-stepwise-based NO estimator will produce valid inference. Ironically, the testing-stepwise-based NO estimator did not perform well in the problematic DGP in Drukker and Liu (2022). One possible reason is that the significance level is essentially an unoptimized tuning parameter in the covariate selection method. We recommend using the BIC-stepwise-based estimator instead of the testing stepwise-based estimator, but both are available in `posw`.

We organize this article as follows. Section 2 describes the high-dimensional GLM model, the BIC-stepwise-based NO estimators, and the testing-stepwise-based NO estimators. Section 6 documents the syntax and options for the command `posw`. Section 4 shows a numerical example of using `posw`. Section 5 presents the simulation results. Finally, section 6 concludes.

[2] For example, in a dataset with 1,000 observations and 100 controls, the BIC-based `popoisson` command takes 0.5 seconds, while `posw` takes 24 seconds.

2 Stepwise-based NO estimator for parameters in HDMs

2.1 HDMs

A cross-sectional high-dimensional GLM can be written as

$$\mathbf{E}\left(y_i|\mathbf{d}_i, \mathbf{x}_i\right) = G\left(\mathbf{d}_i\boldsymbol{\alpha}_0' + \mathbf{x}_i\boldsymbol{\beta}_0'\right)$$

where y is the outcome, $\mathbf{d}_i$ are the covariates of interest, $\mathbf{x}_i$ are the control covariates that potentially need to be included in the model, $\boldsymbol{\alpha}_0$ are the coefficients on $\mathbf{d}_i$, and $\boldsymbol{\beta}_0$ are the coefficients on $\mathbf{x}_i$. $G(\cdot)$ maps the linear index $\mathbf{d}_i\boldsymbol{\alpha}_0' + \mathbf{x}_i\boldsymbol{\beta}_0'$ to the conditional mean. Although there are many other possibilities, three common models are when $G(\cdot)$ is the identity function for linear models, when $G(\cdot)$ is the standard logistic function for logit models, or when $G(\cdot)$ is the exponential function for a (quasi) Poisson or an exponential conditional mean model.

The number of potential covariates in $\mathbf{x}_i$ ($p_{\mathbf{x}}$) can be larger than the sample size n. We are interested in the case in which $p_{\mathbf{x}}$ is too large for a GLM regression of y on $\mathbf{d}$ and $\mathbf{x}$ to produce reliable results for $\boldsymbol{\alpha}$, but the number of covariates in $\mathbf{x}$ that belong in the model ($s_{\mathbf{x}}$) is not too large. Belloni et al. (2012) and Belloni, Chernozhukov, and Wei (2016) derive rates that must bind $s_{\mathbf{x}}$ as a function of n and $p_{\mathbf{x}}$. The assumption that $s_{\mathbf{x}}$ is not too large is the sparsity assumption mentioned in the introduction.

The goal is to obtain reliable estimation and inference for $\boldsymbol{\alpha}_0$. The covariates in $\mathbf{d}_i$ must be specified a priori, and their number is assumed to be small relative to n. Any or all of the coefficients in $\boldsymbol{\alpha}_0$ can be zero. Specifying a covariate to be of interest does not imply that it has a nonzero effect.

The key features of an HDM are that we are interested only in estimating $\boldsymbol{\alpha}_0$, that there are too many covariates in $\mathbf{x}$ to reliably estimate $\boldsymbol{\alpha}_0$ using a quasi–maximum-likelihood (QML) estimator of y on $\mathbf{d}$ and $\mathbf{x}$, and that the sparsity assumption holds.

The sparsity assumption makes the problem feasible and implies that we have a covariate selection problem. Let $\widetilde{\mathbf{x}}_n$ be the subset of $\mathbf{x}$ that we need to include for a QML estimator of y on $\mathbf{d}$ and $\widetilde{\mathbf{x}}_n$ to produce a root-n consistent and asymptotically normal estimator for $\boldsymbol{\alpha}_0$. Belloni et al. (2012), Chernozhukov, Hansen, and Spindler (2015), and Belloni, Chernozhukov, and Wei (2016) provide formal statements and analyses of how to allow for and how to bind the approximation error.

Algorithm 1 gives the naive estimator for the $\boldsymbol{\alpha}_0$ estimator discussed in the Introduction. Leeb and Pötscher (2006, 2008) and Pötscher and Leeb (2009) show that naive estimators like the one in algorithm 1 do not have an asymptotic normal distribution and that they can perform poorly in finite samples when some of the coefficients are small in magnitude. In repeated samples, which of the covariates with small coefficients are included is random. This random inclusion causes small amounts of omitted-variable bias to be randomly added to the estimator. This random omitted-variable bias makes the distribution of the naive estimator have a nonnormal asymptotic distribution. Using a normal distribution to approximate this nonnormal distribution can produce unacceptably poor results in finite samples.

Algorithm 1 Naive estimator for $\boldsymbol{\alpha}_0$

```
1. Use a feasible covariate-selection technique to select the subset of $\mathbf{x}$ that should
   be included in the model. Call these selected covariates $\check{\mathbf{x}}$.
2. Use a QML Poisson estimator of $y$ on $\mathbf{d}$ and $\check{\mathbf{x}}$ to estimate $\boldsymbol{\alpha}_0$.
```

Instead of a naive estimator, `posw` implements NO estimators that were explicitly designed to provide valid inference for $\boldsymbol{\alpha}$ when some of the model's coefficients are small in magnitude. See Belloni et al. (2012), Chernozhukov, Hansen, and Spindler (2015), and Belloni, Chernozhukov, and Wei (2016) for formal results. Instead of naively using the covariates selected in a model of y on $\mathbf{d}$ and $\mathbf{x}$, NO estimators use moment conditions that are robust to the inevitable mistakes that covariate-selection methods make. The NO estimators use multiple covariate-selection steps to form a moment condition for $\boldsymbol{\alpha}$ that is orthogonal to the first-stage selection. This process is an extension of the technique discussed in Neyman (1959, 1979), hence the NO moniker.

The NO estimators can be implemented using different covariate-selection techniques. One popular choice is to use the lasso. Belloni, Chernozhukov, and Wei (2016) derive the Lasso-based NO estimator for the high-dimensional GLM. It uses a particular version of the lasso that selects the tuning parameters using a plugin method. In practice, the NO estimator's performance critically depends on the choice of tuning parameter selection method.

Drukker and Liu (2022) studied the finite sample behavior of the lasso-based NO estimator for the HDM using different tuning parameter selection methods. The BIC-stepwise-based NO estimators are shown in their simulations to provide reliable inferential results for the family of DGPs where the lasso-based NO estimators provide poor inferential results.

Kozbur (2020) presents formal results for testing stepwise selection for linear models. These results show that testing-stepwise-based NO estimators will perform well in the large sample under the conditions described in the article. Given these formal results, it is a little surprising that these estimators did not perform well for the problematic DGP in Drukker and Liu (2022). We conjecture that the poor performance of the estimators on the problematic DGP was due to the choice of the significance level. In practice, we recommend the BIC-stepwise selection because it avoids choosing the significance level and because of how well it has performed in our simulations.

As discussed by Belloni and Chernozhukov (2011), the lasso can be viewed as a convex approximation to the computationally infeasible problem of finding the subset of covariates that best approximates a conditional expectation function. The family of stepwise methods is another approach to solving this best-subset regression problem. Stepwise methods are computationally feasible for many HDMs, but they take much longer than lasso methods and become infeasible for very high-dimensional problems.

2.2 Algorithms

The BIC-stepwise algorithm used by `posw` is algorithm 3 in Drukker and Liu (2022). The testing-stepwise-based algorithm used by `posw` is algorithm 4 in Drukker and Liu (2022).

Belloni, Chernozhukov, and Wei (2016) derived lasso-based NO estimators for the GLM model. We implement versions of NO estimators that use BIC stepwise or testing stepwise for covariate selection. Algorithm 2 provides the details about these versions of the Belloni, Chernozhukov, and Wei (2016) NO estimator. Algorithm 2 generalizes algorithm 6 in Drukker and Liu (2022) from the Poisson regression case to the GLM regression case.

Algorithm 2 Stepwise-based NO GLM estimation

1. In a GLM model of y on $\mathbf{d}$ and $\mathbf{x}$, use covariate selection to find the subset of the $\mathbf{x}$ covariates that have nonzero coefficients. Denote this subset by $\widetilde{\mathbf{x}}$.

    - For the BIC-stepwise NO estimator, we find the subset of the $\mathbf{x}$ that BIC stepwise includes.
    - For the testing-stepwise NO estimator, we find the subset of the $\mathbf{x}$ that testing stepwise includes.

2. Use the unpenalized QML GLM regression estimator to estimate the coefficients $\widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\beta}}$ in a GLM model of y on $\mathbf{d}$ and $\widetilde{\mathbf{x}}$.

3. Let $\widetilde{s}_i = \widetilde{\mathbf{x}}_i\widetilde{\boldsymbol{\beta}}'$ be the ith observation of the predicted value of the linear index $\mathbf{x}\boldsymbol{\beta}'$.

4. Let $\omega_i = G'(\mathbf{d}_i\widetilde{\boldsymbol{\alpha}}' + \widetilde{s}_i)$ be the ith observation of the predicted value of the derivative of $G(\cdot)$. Let $\sigma_i^2 = \widehat{\mathrm{Var}}(\mathbf{y}_i|\mathbf{d}_i, \mathbf{x}_i)$. Let $f_i = \omega_i/\sigma_i$.

5. For each $j \in \{1, \ldots, J\}$, use a linear stepwise of the jth variable in $\mathbf{d}$ on $\mathbf{x}$ using observation-level weights f_i, and let $\check{\mathbf{x}}_j$ be the selected covariates.

    - The BIC-stepwise NO estimator uses a weighted BIC-stepwise for covariate selection.
    - The testing-stepwise NO estimator uses a weighted testing-stepwise for covariate selection.

6. For each $j \in \{1, \ldots, J\}$, run a linear, ordinary least-squares regression of the jth variable in $\mathbf{d}$ on $\check{\mathbf{x}}_j$ with observation-level weights f_i. Let $\widetilde{d}_j$ be the unweighted residuals from this regression, and let $\widetilde{d}_{j,i}$ be the ith observation on $\widetilde{d}_j$.

7. Create the vector of instrumental variables $\mathbf{z} = (\widetilde{d}_1, \ldots, \widetilde{d}_J)$, and let $\mathbf{z}_i$ be the ith observation on this vector. Note that $\mathbf{z}_i = (z_{1,i}, \ldots, z_{J,i}) = (\widetilde{d}_{1,i}, \ldots, \widetilde{d}_{J,i})$.

8. Compute $\widehat{\boldsymbol{\alpha}}$ by solving the J sample-moment equations

$$\frac{1}{n}\sum_{i=1}^{n} \{y_i - G(\mathbf{d}_i\boldsymbol{\alpha}' + \widetilde{s}_i)\}\,\mathbf{z}_i = \mathbf{0}$$

    We use the standard robust estimator for the asymptotic variance of a method-of-moments estimator.

3 Syntax of posw

`posw` has the following syntax.

`posw` *depvar* *varsofinterest* [*if*] [*in*], `controls(`*varlist*`)`
  `model(linear|logit|poisson)` [`method(bic|test)` `alpha(`*#*`)`]

varsofinterest are variables for which coefficients and their standard errors are estimated.

3.1 Options

`controls(`*varlist*`)` specifies the set of control variables, which control for omitted variables. Control variables are also known as confounding variables. `posw` uses the forward stepwise to select the control variables for each of *depvar* and *varsofinterest*. `controls()` is required.

`model(linear|logit|poisson)` specifies the model for the outcome variable *depvar*. It can be one of `linear`, `logit`, or `poisson`. `model()` is required.

`method(bic|test)` specifies the method used in stepwise covariate selection. It can be one of `bic` or `test`. Specifying `bic` implies using the BIC-based stepwise. Specifying `test` implies using the testing-based stepwise. The default is `method(bic)`.

`alpha(`*#*`)` specifies the level of significance for the testing-based stepwise. The default is `alpha(0.05)`.

3.2 Stored results

`posw` stores the following results in `e()`:

Scalars	
`e(N)`	number of observations
`e(k_controls)`	number of controls
`e(k_controls_sel)`	number of selected controls
`e(k_varsofinterest)`	number of variables of interest
Macros	
`e(cmd)`	`posw`
`e(depvar)`	dependent variable
`e(title)`	title in estimation output
`e(vce)`	`robust`
`e(vcetype)`	`Robust`
`e(properties)`	`b V`
`e(varsofinterest)`	variables of interest
`e(controls_sel)`	selected control variables
`e(controls)`	control variables
`e(model)`	type of model
Matrices	
`e(b)`	coefficient vector
`e(V)`	variance–covariance matrix of the estimators
Functions	
`e(sample)`	marks estimation sample

4 A numerical example

We now illustrate the use of posw through an empirical example. We have an extract of data from Sunyer et al. (2017) to measure the effect of air pollution level on the student's response time. The model is

$$\texttt{react}_i = \texttt{no2\_class}_i \alpha + \mathbf{x}_i \boldsymbol{\beta}' + \epsilon_i$$

where $\texttt{react}_i$ is the response time of child i on a test, $\texttt{no2\_class}_i$ is the pollution level in the school attended by child i, $\mathbf{x}_i$ is a high-dimensional control to be included in the model, and ϵ_i is the disturbance term.

To start, we bring breathe.dta into memory.

```
. use https://www.stata-press.com/data/r17/breathe
(Nitrogen dioxide and attention)
```

Next, we need to define the control variables **x**. It would be tedious to separate the factor variables from the continuous variables manually. Instead, we can use Stata's variable management tools vl. Here we directly call a do-file from the Stata website to save some typing.

```
. quietly do https://www.stata-press.com/data/r17/no2
```

The purpose of this do-file is to define a global macro $fc for the factor control variables and a global macro $cc for the continuous control variables. We can display their content.

```
. display "$cc"
no2_home age age0 sev_home green_home noise_school sev_school precip
> siblings_old siblings_young
. display "$fc"
sex grade overweight lbweight breastfeed msmoke meducation feducation
```

Now, we can define the control variables as raw variables and the full second-order interaction among them. The control variables are stored in global macro $controls for later use.

```
. global controls (c.($cc) i.($fc))##(c.($cc) i.($fc))
```

Finally, we can fit our model using `posw`. We specify option `controls()` for the control variables and the option `model()` for the linear model.

```
. posw react no2_class, controls($controls) model(linear)
select controls for react using stepwise bic
select controls for no2_class using stepwise bic
Partialing-out stepwise bic         Number of obs                =     1,036
                                    Number of controls           =       516
                                    Number of selected controls  =        16
                                    Wald chi2(1)                 =     20.97
Model: linear                       Prob > chi2                  =    0.0000

------------------------------------------------------------------------------
             |               Robust
       react | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
   no2_class |   2.493274     .54448     4.58   0.000     1.426113    3.560435
------------------------------------------------------------------------------
Note: Chi-squared test is a Wald test of the coefficients of the variables of
      interest jointly equal to zero.
```

The results imply that another microgram of NO2 per cubic meter increases the mean reaction time by 2.5 milliseconds. Only the coefficient on the covariate of interest is estimated. The coefficients on the control covariates are not estimated. The cost of using covariate selection methods is that these estimators do not produce estimates for the coefficients on the control covariates. Remarkably, there are 516 control variables, and only 16 of them are selected.

5 Simulations

This section describes some Monte Carlo simulations that illustrate the good performance of the `posw` command in linear, logit, and Poisson and exponential conditional mean models. We note that the naive estimators perform poorly on these designs.[3]

5.1 Designs

The design for the simulations reflects the structure of the high-dimensional GLM model. In each design, there are a few covariates whose coefficients have values that are large in magnitude, a few covariates whose coefficients have values that are small in magnitude, and many covariates whose coefficients are zero. Following Drukker and Liu (2022), we specify the value of each small coefficient to be about two times its standard error in the true model for the sample size used in the simulations. We specify the value of each large coefficient to be about four times its standard error in the true model for the sample size used in the simulations. These values encode a representation of DGPs for which there is no beta-min condition. The small coefficients are close to being statistically significant, while the large coefficients should be statistically significant in the vast majority of the repeated samples.

[3] For a more detailed simulation study on the comparison of `posw` and `popoisson`, see Drukker and Liu (2022).

The DGPs for the three designs are given in (1)–(3).

$$\text{Linear} \qquad y_i = w_i + \epsilon_i \tag{1}$$

$$\text{Logit} \qquad y_i = w_i + \epsilon_i > 0 \tag{2}$$

$$\text{Poisson} \qquad y_i = \exp(w_i)\epsilon_i \tag{3}$$

For each design

$$w_i = \alpha_{\text{big}} d_{\text{big},i} + \alpha_{\text{small}} d_{\text{small},i} + \alpha_{\text{zero}} d_{\text{zero},i} + \mathbf{x}_{\text{big},i}\boldsymbol{\beta}'_{\text{big}} + \mathbf{x}_{\text{small},i}\boldsymbol{\beta}'_{\text{small}}$$

where

- d_{big} is the covariate of interest whose coefficient, α_{big}, is large;
- d_{small} is the covariate of interest whose coefficient, α_{small}, is small;
- d_{zero} is the covariate of interest whose coefficient, α_{zero}, is zero;
- $\mathbf{x}_{\text{big}}$ is the vector of control covariates whose coefficients, $\boldsymbol{\beta}_{\text{big}}$, are large;
- $\mathbf{x}_{\text{small}}$ is the vector of control covariates whose coefficients, $\boldsymbol{\beta}_{\text{small}}$, are small;

A type of Toeplitz structure is frequently used to generate the covariates in HDMs; see Belloni et al. (2012) for an example. In these Toeplitz structures, each covariate has an index $j \in \{1, \dots, p\}$, where p is the total number of covariates in the model. (This includes the covariates of interest and the control covariates.) Covariates with nearby indices are significantly correlated, but the amount of correlation decays as the distance between the indices increases. Each covariate was generated as

$$x_j = 0.3x_{j-1} + 0.2x_{j-4} + 0.2x_{j-8} + \eta$$

where $\eta = (\texttt{rchi2(25)} - 25)/\texttt{sqrt(50)})$ and `rchi2(`*a*`)` is the Stata function that generates draws from a χ^2 distribution with a degrees of freedom. The distribution for η has mean zero and variance one, and its higher moments are distinctly different from those in a normal distribution.

For each design, we generated $p = 100$ covariates. For each draw, we drew 120 covariates and discarded the first 20 to burn in the Toeplitz structure.

- Covariate 1 is d_{big}, the covariate of interest whose coefficient is large.
- Covariates 2–5 are $\mathbf{x}_{\text{big}}$, the control covariates whose coefficients are large.
- Covariate 6 is d_{small}, the covariate of interest whose coefficient is small.
- Covariates 7–10 are $\mathbf{x}_{\text{small}}$, the control covariates whose coefficients are small.
- Covariate 11 is d_{zero}, the covariate of interest whose coefficient is zero.
- Covariates 12–100 are $\mathbf{x}_{\text{zero}}$, the control covariates whose coefficients are zero.

Here are true values for the coefficients.

Table 1. True coefficient values

Model	Coefficient	Value
Linear	α_{big}	0.12
Linear	α_{small}	0.06
Linear	β_{big}	0.12
Linear	β_{small}	0.06
Logit	α_{big}	0.32
Logit	α_{small}	0.16
Logit	β_{big}	0.32
Logit	β_{small}	0.16
Poisson	α_{big}	0.08
Poisson	α_{small}	0.04
Poisson	β_{big}	0.08
Poisson	β_{small}	0.04

Here is how the error term was generated for each design. For the linear design, $\epsilon_i = (\texttt{rchi2(25)} - 25)/\texttt{sqrt(50)})$, where `rchi2(`$a$`)` is the Stata function that generates draws from a χ^2 distribution with a degrees of freedom. This distribution has mean zero and variance one, and its higher moments are distinctly different from those in a normal distribution.

For the logit design, $\epsilon_i = \texttt{rlogistic}()$, where `rlogistic()` generates draws from a standard logistic distribution.

For the Poisson design, $\epsilon_i = \texttt{rweibull}(2, b)$, where `rweibull(`$c$`,`$b$`)` generates draws from a Weibull distribution with shape parameter c and scale parameter b. We set $b = 1/\exp(\texttt{lngamma(1 + 1/2)})$ so that the mean of ϵ is 1. Note that we are drawing from a conditional exponential mean model, not from a Poisson model.

We ran 2,400 repetitions for each design. There are 2,400 repetitions because we used Stata's stream random numbers to parallelize the simulations over 40 cores, with 60 repetitions on each core. (See `help rngstream` for an introduction to stream random numbers.)

5.2 Results

Tables 2, 3, and 4 summarize the simulation results. In short, we see that the NO estimators in the BIC step and test step perform well and that the naive estimators in the BIC naive and test naive do not perform well.

Here is how each table is structured.

- Model specifies the DGP design.
- Estimator specifies the estimator used for that row's results.
- Mean specifies the sample mean of the estimates of that coefficient over the repetitions. The sample mean should be close to the true value given in table 1.
- SD specifies the sample standard deviation of the estimates of that coefficient over the repetitions.
- SE specifies the sample mean of the standard errors of that coefficient over the repetitions. The sample mean of the standard errors should be close to the sample standard deviation of the estimates.
- RR specifies the rejection rate of a test against the true null hypothesis. The significance level of each test was 0.05, so the RR should be close to 0.05.

Table 2. Result for large coefficient α_{big}

Model	Estimator	Mean	SD	SE	RR
Linear	BIC step	0.1172	0.0343	0.0336	0.060
Linear	Test step	0.1157	0.0343	0.0337	0.059
Linear	BIC naive	0.1381	0.0387	0.0316	0.146
Linear	Test naive	0.1546	0.0420	0.0311	0.280
Linear	True	0.1198	0.0338	0.0332	0.055
Logit	BIC step	0.3260	0.0854	0.0824	0.054
Logit	Test step	0.3214	0.0863	0.0797	0.072
Logit	BIC naive	0.3649	0.0921	0.0791	0.128
Logit	Test naive	0.3980	0.0979	0.0769	0.240
Logit	True	0.3255	0.0824	0.0820	0.045
Poisson	BIC step	0.0787	0.0192	0.0198	0.047
Poisson	Test step	0.0796	0.0194	0.0192	0.052
Poisson	BIC naive	0.1088	0.0246	0.0298	0.137
Poisson	Test naive	0.0963	0.0224	0.0176	0.218
Poisson	True	0.0799	0.0185	0.0185	0.051

Table 3. Result for small coefficient α_{small}

Model	Estimator	Mean	SD	SE	RR
Linear	BIC step	0.0586	0.0339	0.0341	0.053
Linear	Test step	0.0583	0.0338	0.0341	0.052
Linear	BIC naive	0.0760	0.0360	0.0317	0.111
Linear	Test naive	0.0868	0.0373	0.0314	0.188
Linear	True	0.0602	0.0327	0.0331	0.050
Logit	BIC step	0.1640	0.0851	0.0837	0.055
Logit	Test step	0.1632	0.0844	0.0813	0.061
Logit	BIC naive	0.2005	0.0876	0.0780	0.112
Logit	Test naive	0.2232	0.0868	0.0761	0.165
Logit	True	0.1643	0.0797	0.0809	0.052
Poisson	BIC step	0.0381	0.0196	0.0200	0.048
Poisson	Test step	0.0387	0.0193	0.0196	0.052
Poisson	BIC naive	0.0612	0.0211	0.0308	0.032
Poisson	Test naive	0.0530	0.0208	0.0176	0.168
Poisson	True	0.0391	0.0186	0.0185	0.055

Table 4. Result for zero coefficient α_{zero}

Model	Estimator	Mean	SD	SE	RR
Linear	BIC step	0.0003	0.0341	0.0341	0.053
Linear	Test step	0.0015	0.0343	0.0341	0.055
Linear	BIC naive	0.0072	0.0338	0.0314	0.078
Linear	Test naive	0.0147	0.0353	0.0312	0.114
Linear	True	0.0005	0.0320	0.0317	0.056
Logit	BIC step	0.0020	0.0841	0.0837	0.050
Logit	Test step	0.0051	0.0826	0.0815	0.054
Logit	BIC naive	0.0136	0.0802	0.0765	0.064
Logit	Test naive	0.0316	0.0813	0.0749	0.092
Logit	True	−0.0002	0.0762	0.0773	0.040
Poisson	BIC step	0.0014	0.0197	0.0200	0.051
Poisson	Test step	0.0018	0.0197	0.0196	0.055
Poisson	BIC naive	0.0131	0.0204	0.0310	0.013
Poisson	Test naive	0.0070	0.0194	0.0175	0.100
Poisson	True	0.0003	0.0177	0.0177	0.050

6 Conclusion

We showed the motivation behind and described posw, a command for the stepwise-based NO estimator in the linear, logit, and Poisson models. This command can be viewed as an alternative to the lasso-based NO estimators, which are implemented in official Stata commands poregress, pologit, and popoisson. Simulations in Drukker and Liu (2022) show that the implemented BIC-stepwise-based NO can perform better than the lasso-based NO estimators for a family of DGPs.

The main cost of using a stepwise-based NO estimator instead of a lasso-based NO estimator is an increase in computation time. Future development could speed up posw by using cluster-parallel computation or the sure-independence-screening version of the stepwise partialing-out estimator outlined in Drukker and Liu (2022).

7 Acknowledgments

We thank the editor and an anonymous referee for their comments and suggestions.

8 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 23-2
. net install st0713      (to install program files, if available)
. net get st0713          (to install ancillary files, if available)
```

9 References

Belloni, A., D. Chen, V. Chernozhukov, and C. 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. 2011. l_1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics* 39: 82–130. https://doi.org/10.1214/10-AOS827.

Belloni, A., V. Chernozhukov, and C. Hansen. 2014. Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies* 81: 608–650. https://doi.org/10.1093/restud/rdt044.

Belloni, A., V. Chernozhukov, and Y. Wei. 2016. Post-selection inference for generalized linear models with many controls. *Journal of Business and Economic Statistics* 34: 606–619. https://doi.org/10.1080/07350015.2016.1166116.

Chernozhukov, V., C. Hansen, and M. Spindler. 2015. Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics* 7: 649–688. https://doi.org/10.1146/annurev-economics-012315-015826.

Drukker, D. M., and D. Liu. 2022. Finite-sample results for lasso and stepwise Neyman-orthogonal Poisson estimators. *Econometric Reviews* 41: 1047–1076. https://doi.org/10.1080/07474938.2022.2091363.

———. Forthcoming. A cluster plugin method for selecting the GLM lasso tuning parameters in models for unbalanced panel data. *Econometrics and Statistics* . https://doi.org/10.1016/j.ecosta.2022.02.006.

Kozbur, D. 2020. Testing-based forward model selection. ArXiv Working Paper No. arXiv:1512.02666. https://doi.org/10.48550/arXiv.1512.02666.

Leeb, H., and B. M. Pötscher. 2006. Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics* 34: 2554–2591. https://doi.org/10.1214/009053606000000821.

———. 2008. Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* 142: 201–211. https://doi.org/10.1016/j.jeconom.2007.05.017.

Neyman, J. 1959. Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics: The Harald Cramér Volume*, ed. U. Grenander, 213–234. New York: Wiley.

———. 1979. C(α) tests and their use. *Sankhyā* 41: 1–21.

Pötscher, B. M., and H. Leeb. 2009. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis* 100: 2065–2082. https://doi.org/10.1016/j.jmva.2009.06.010.

Sunyer, J., E. Suades-González, R. García-Esteban, I. Rivas, J. Pujol, M. Alvarez-Pedrerol, J. Forns, X. Querol, and X. Basagaña. 2017. Traffic-related air pollution and attention in primary school children: Short-term association. *Epidemiology* 28: 181–189. https://doi.org/10.1097/ede.0000000000000603.

About the authors

David M. Drukker is an Associate Professor in the Department of Economics and International Business at Sam Houston State University.

Di Liu is a Principal Econometrician at StataCorp.