A plug-in for Poisson lasso and a comparison of partialing-out Poisson estimators that use different methods for selecting the lasso tuning parameters

David M. Drukker *
Stata
ddrukker@stata.com
Di Liu
Stata
dliu@stata.com

September 18, 2019

Abstract

High-dimensional models that include a few covariates of interest and many control covariates which might potentially affect an outcome are increasingly common. The least absolute shrinkage and selection operator (lasso) is frequently used in this context. Belloni et al. (2016b) derived lasso-based partialing-out (PO) estimators for the coefficients of interest in high-dimensional generalized linear models (GLMs). The tuning parameters of the lasso must be selected to implement these PO estimators, but Belloni et al. (2016b) only presented a method of selecting the lasso tuning parameters for the high-dimensional logit model. This paper extends the Belloni et al. (2012) plug-in algorithm for choosing the lasso tuning parameters to high-dimensional GLMs. This paper presents simulation evidence that a PO Poisson estimator that uses this plug-in algorithm performs well. The simulations also show that a PO Poisson estimator that uses cross-validation (CV) or the adaptive lasso (AL) to select the lasso tuning parameters can require a much larger sample size to perform as well. The simulations also show that a PO Poisson estimator that selects the lasso tuning parameters by minimizing the Bayesian information criterion (BIC) performs almost as well as the plug-in based estimator. Finally, the paper explains these simulation results by discussing the covariate-selection tendencies of the Poisson lasso when the tuning parameters are selected by the plug-in method, by CV, by AL, and by the BIC.

*We thank Enrique Pinzon and Joerg Luedicke for discussions and for their work in running previous simulations. We also thank Fang Wang for speeding up the coordinate-descent algorithm that does the numerical optimization.

1 Introduction

High-dimensional models that include many covariates which might potentially affect an outcome are increasingly common. One approach to high-dimensional models makes a sparsity assumption which requires that the number of potential covariates that must be included in the model is small relative to the sample size.[1] This sparse approach to high-dimensional models uses lasso-based covariate selection and moment conditions that are robust to the covariate selection to produce reliable inference for some of the model parameters. We call this sparse approach the partialing-out (PO) approach, because it extends the classic partialing-out technique for obtaining some regression coefficients after removing the impact of other covariates. The PO approach was derived in Belloni et al. (2012), Belloni et al. (2014), and Belloni et al. (2016b).

Belloni et al. (2012) derived a plug-in method for selecting the lasso tuning parameters for linear-model lassos. Belloni et al. (2012) also presented an algorithm for implementing their plug-in method for linear models. Belloni et al. (2016b) derived PO estimators for generalized linear models (GLMs), but they did not extend their method and algorithm for implementing the plug-in method for selecting the tuning parameters to the GLM model.[2] This paper extends their plug-in method and their algorithm to the GLM model. The paper also presents simulation evidence that this extension produces a version of the Belloni et al. (2016b) PO Poisson that performs well in finite samples.

Cross-validation (CV), the adaptive lasso (AL), and minimizing the Bayesian information criterion (BIC) are also commonly used to select the lasso tuning parameters in GLM lassos.[3] This paper presents simulation evidence that the BIC-based PO Poisson estimator can perform almost as well as the plug-in-based PO estimator. This simulation evidence also shows that the CV-based PO Poisson estimator and the AL-based PO Poisson estimator can require a much larger-sample to perform as well as the plug-in-based PO Poisson estimator. For the designs used in this paper, the problem is that CV-based lasso and the AL-based lasso tend to include many covariates whose coefficients are zero in the true model.

The PO estimators were explicitly designed to provide valid inference when some of the coefficients in the model are small in magnitude. Leeb and Pötscher (2008) show that naive estimators that use the covariates selected by a lasso as if they were the covariates in the best approximating model do not have an asymptotic normal distribution and can

[1] Another approach removes the many-covariate bias and uses many-covariate robust methods to estimate the asymptotic variance of the bias corrected estimator. See Cattaneo et al. (2018b) and Cattaneo et al. (2018a) for this approach.

[2] Belloni et al. (2016b) derived values for the lasso tuning parameters for the logit model. For the logit case, they did not need to use their algorithm, because they were able show that a specific value binds the lasso penalty loadings.

[3] For examples and introductions, see Hastie et al. (2015), Zou (2006), Bühlmann and Van de Geer (2011), and Zhang et al. (2010)

perform poorly in finite samples. The simulation designs used in this paper replicate their results. The naive estimators that use any the discussed methods to select the lasso tuning parameters perform poorly on the designs used in this paper. That all naive estimators fail on these designs is evidence that these designs are, to some extent, a good test for the PO estimators.

Here is an outline for the remainder of the paper. Section 2.1 introduces high-dimensional models. Section 2.2 introduces the lasso. Section 2.3 extends the Belloni et al. (2012) to the GLM case. Section 2.4 discusses other methods for selecting the lasso tuning parameters. Section 3 discusses our simulation results.

2 Lasso for inference in high-dimensional models

2.1 High-dimensional models

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

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

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 $\mathbf{x}_i$. $G()$ 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 $G()$ is the identity function for linear models, when $G()$ is the standard logistic distribution for logit models, or when $G()$ is the exponential function for Poisson or exponential conditional mean models.

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 et al. (2016b) derive rates that must bind $s_{\mathbf{x}}$ as a function of n $p_{\mathbf{x}}$.

The goal is to obtain reliable estimation and inference for $\boldsymbol{\alpha}_0$. The number of covariates in $\mathbf{d}_i$ is assumed to be fixed and small relative to n.

The PO approach uses lasso-based covariate selection to determine which of the covariates in $\mathbf{x}$ should be included in the model.[4] The PO approach does not estimate $\boldsymbol{\beta}_0$. Not estimating $\boldsymbol{\beta}_0$ can be viewed as the cost of using covariate-selection methods to determine which of the potential covariates in $\mathbf{x}_i$ should be included in the model. We use a version of the PO estimator in Belloni et al. (2016b). The details of this estimator are given in appendix B.

[4]Other covariate-selection techniques could be used, but they are outside the realm of this paper. See Chernozhukov et al. (2018) and Kozbur (2019) for results and discussions.

2.2 The lasso

The lasso is a widely used technique for covariate selection. Mechanically, the cross-sectional GLM lasso with given penalty parameter λ and given penalty loadings κ_j $j \in \{1, \ldots, p\}$ solves

$$\widehat{\boldsymbol{\delta}} = \arg\min_{\boldsymbol{\delta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} Q(y_i, \mathbf{w}_i \boldsymbol{\delta}') + \lambda \sum_{j=1}^{p} \kappa_j |\boldsymbol{\delta}_j| \right\} \tag{1}$$

where $\boldsymbol{\delta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, $\mathbf{w}_i = (\mathbf{d}_i, \mathbf{x}_i)$, and $Q(y_i, \mathbf{w}_i \boldsymbol{\delta}')$ is the negative of the contribution of the i(th) observation to the GLM quasi-maximum-likelihood (QML) function. See Hastie et al. (2015), Friedman et al. (2007), and Friedman et al. (2010) for descriptions of the coordinate-descent algorithm used to perform the minimization. The formulas for $Q()$ are given in appendix C.

The first term in the objective function in equation (1) is the usual QML objective function. The second term penalizes the objective function for allowing a coefficient to differ from zero. The kink in the absolute value function in the penalty term causes some elements of $\widehat{\boldsymbol{\delta}}$ to be exactly zero at the minimum, while others are not zero; see Hastie et al. (2015) for details.

That some elements of $\widehat{\boldsymbol{\delta}}$ are exactly zero at the minimum, while others are not zero is the basis of the lasso as a covariate selection technique. The j(th) covariate in $\mathbf{w}$ is included, if the j(th) element in $\widehat{\boldsymbol{\delta}}$ is not zero. Analogously, the j(th) covariate in $\mathbf{w}$ is excluded, if the j(th) element in $\widehat{\boldsymbol{\delta}}$ is zero.

The penalty parameter λ and the penalty loadings κ_j $j \in \{1, \ldots, p\}$ are known as the lasso tuning parameters. One must choose the lasso tuning parameters before using the lasso for covariate selection. The values of λ and the κ_j determine which covariates will have estimated coefficients that are not zero and which covariates will have estimated coefficients that are zero. Thus, the properties of the lasso-covariate-selection method depend on the method used to choose the tuning parameters.

The most widely used methods for selecting the tuning parameters are CV, plug-in methods, AL, and minimizing an information criterion. We want a method that finds the important covariates, but does not include too many extra covariates.

Belloni, Chen, Chernozhukov, and Hansen (2012), Belloni, Chernozhukov, and Hansen (2014), Belloni, Chernozhukov, and Wei (2016b), and Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018) derive PO estimators that are robust to the mistakes that the lasso makes in excluding covariates with small coefficients. Belloni, Chen, Chernozhukov, and Hansen (2012),Belloni, Chernozhukov, and Hansen (2014), and Belloni, Chernozhukov, and Wei (2016b), rigorously show that some of these robust methods will perform well when the tuning parameters are selected using their plug-in method. Belloni, Chen, Chernozhukov, and Hansen (2012) derives a plug-in method for a linear model and provides an algorithm to implement it. Belloni et al. (2016b) derives

PO estimators for GLMs, but it does not provide a plug-in method or an algorithm for Poisson models.[5]

2.3 Plug-in for generalized linear models

The penalty-loadings play an essential role in defining the penalty level λ that provides a good lasso estimator. Following Belloni, Chernozhukov, and Wei (2016b), and Bickel, Ritov, and Tsybakov (2009), the GLM lasso will have good selection properties if

$$P\left(\lambda \geq c \max_{1\leq j\leq p}\left|\frac{1}{n}(1/\kappa_j)\sum_{i=1}^{n}\mathbf{h}_j(y_i,\mathbf{w}_i\boldsymbol{\delta}_0')\right|\right) \to_n 1 \tag{2}$$

where we have the following definitions.

- c is a constant greater than 1.
- $\mathbf{h}_j(y_i,\mathbf{w}_i\boldsymbol{\delta}_0')$ is the contribution of the i(th) observation to the score for the unpenalized QML estimator for the j(th) parameter evaluated at $\boldsymbol{\beta}_0$.
- Each penalty loading has its ideal value

$$\kappa_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{h}_j(y_i,\mathbf{w}_i\boldsymbol{\delta}^{0\prime})\right]^2}$$

To be more specific about the scores, the vector of scores for the unpenalized QML estimator is

$$\frac{\partial Q(y_i,\mathbf{w}_i\boldsymbol{\delta}')}{\partial\boldsymbol{\delta}} = \mathbf{h}(y_i,\mathbf{w}_i\boldsymbol{\delta}')$$

and the contribution of the i(th) observation to the j(th) score is $\mathbf{h}_j(y_i,\mathbf{w}_i\boldsymbol{\delta}')$, which is the j(th) element of $\mathbf{h}$.

Note that equation (2) can be written as

$$P\left(\lambda \geq c \max_{1\leq j\leq p}\left|\frac{\frac{1}{n}\sum_{i=1}^{n}\mathbf{h}_j(y_i,\mathbf{w}_i\boldsymbol{\delta}_0')}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{h}_j(y_i,\mathbf{w}_i\boldsymbol{\delta}_0')\right]^2}}\right|\right) \to_n 1 \tag{3}$$

[5]Belloni et al. (2016b) derives a plug-in method for the tuning parameters of logit models, but this method uses a fixed bound. Belloni et al. (2016b) does not extend the algorithm in Belloni et al. (2012) to the GLM case.

In a series of analogous cases, Belloni, Chen, Chernozhukov, and Hansen (2012), Belloni, Chernozhukov, and Hansen (2014), and Belloni, Chernozhukov, and Wei (2016b) use the self-normalized moderate deviation theory developed by Jing, Shao, and Wang (2003) and Peña, Lai, and Shao (2009) to show that

$$P\left(\sqrt{n}\max\left|\frac{1}{n}(1/\kappa_j)\sum_{i=1}^{n}\mathbf{h}_j(y_i\mathbf{w}_i\boldsymbol{\delta}_0^{'})\right| \leq \Phi^{-1}(1-\gamma/(2p))\right) \geq 1-\gamma+o(1) \tag{4}$$

under reasonable conditions.

Using equations (2) and (4), Belloni, Chen, Chernozhukov, and Hansen (2012), Belloni, Chernozhukov, and Hansen (2014), and Belloni, Chernozhukov, and Wei (2016b) define the ideal value for λ in equation (1) to be

$$\lambda = \frac{c}{\sqrt{n}}\Phi^{-1}\left[1-\gamma/(2p)\right] \tag{5}$$

in a series of analogous cases.

2.3.1 What is the logic behind equation 4

Getting a handle on where equation (4) comes from is essential to understanding the extension in this paper. Belloni, Chen, Chernozhukov, and Hansen (2012) use theorem 7.4 in Peña et al. (2009) to show that

$$P\left(\max_{1\leq j\leq p}|S_j| > \Phi^{-1}(1-\gamma/(2p))\right) \leq \gamma\left(1+A/\ell_n^3\right)$$

where

$$S_j = \frac{\sum_{i=1}^{n}U_{ij}}{\sqrt{\sum_{i=1}^{n}U_{ij}^2}}$$

each U_{ij} is an independent realization of a mean zero random variable.

For small γ, this can be shown to imply that

$$P\left(\max_{1\leq j\leq p}|S_j| \leq \Phi^{-1}(1-\gamma/(2p))\right) \approx 1-\gamma \tag{6}$$

Note that the definition of S_j in equation (6) uses sums over i that are not multiplied by $1/n$ but that the sums over i in equation (3) are multiplied by $1/n$. Extending Belloni, Chen, Chernozhukov, and Hansen (2012) to the GLM case, we let

$$U_{i,j} = \mathbf{h}_j(y_i, \mathbf{w}_i\boldsymbol{\delta}_0')$$

Conceptually, we have

$$\sqrt{n}(1/\kappa_j)1/n\sum_{i=1}^{n}\left[\mathbf{h}_j(y_i,\mathbf{w}_i\boldsymbol{\delta}_0')\right] \tag{7}$$

$$=\sqrt{n}\frac{1/n\sum_{i=1}^{n}\left[\mathbf{h}_j(y_i,\mathbf{w}_i\boldsymbol{\delta}_0')\right]}{\sqrt{1/n\sum_{i=1}^{n}\left[\mathbf{h}_j(y_i,\mathbf{w}_i\boldsymbol{\delta}_0')\right]^2}} \tag{8}$$

$$=\sqrt{n}\frac{1/n\sum_{i=1}^{n}\left[U_{i,j}\right]}{\sqrt{1/n\sum_{i=1}^{n}\left[U_{i,j}\right]^2}} \tag{9}$$

$$=S_j \tag{10}$$

Substituting the expression in (7) in for S_j in equation (6) yields equation (4).

2.3.2 Algorithm for the penalty loadings

We use an extension of the algorithm derived by Belloni, Chen, Chernozhukov, and Hansen (2012) to estimate plug-in values of λ and $\kappa_1,\ \ldots,\ \kappa_p$ that can be used to solve the lasso in equation (1).

Algorithm 1: Plug-in method for Poisson lasso tuning parameters

This algorithm assumes that each x_j has been normalized to have mean 0 and variance 1.
On exit, λ contains the penalty value and the penalty loadings are in $(\tilde{\kappa}_1, \ \ldots, \ \tilde{\kappa}_p)$.

1. Set $\gamma = 0.1/\ln(\max(p, n))$ and set $c = 1.1$.
2. Set $\lambda = \frac{c}{\sqrt{n}}\Phi^{-1}\left[1 - \gamma/(2p)\right]$.
3. Find the five covariates that have the highest correlations with y. Denote the vector of them by $\tilde{\mathbf{x}}_0$ and let $\tilde{\mathbf{x}}_{0,i}$ be the i(th) observation of this vector of variables.
4. Estimate the coefficients $\tilde{\boldsymbol{\beta}}_0$ on $\tilde{\mathbf{x}}_0$ by unpenalized GLM QML.
5. For each $j \in \{1, \ldots, p\}$, set
$$\tilde{\kappa}_{0,j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{h}_j(y_i, \tilde{\mathbf{x}}_{0,i}\tilde{\boldsymbol{\beta}}_0')\right]^2}$$
6. Set $k = 1$ and do the following loop. (It will be executed at most 15 times.)
   (a) Using λ and loadings $\{\tilde{\kappa}_{k-1,1}, \ \ldots, \ \tilde{\kappa}_{k-1,p}\}$ to solve (1) which produces estimates $\tilde{\tilde{\boldsymbol{\beta}}}_k$.
   (b) Let $\tilde{\mathbf{x}}_k$ be the covariates with nonzero coefficients in $\tilde{\tilde{\boldsymbol{\beta}}}_k$.
   (c) Estimate the coefficients $\tilde{\boldsymbol{\beta}}_k$ on $\tilde{\mathbf{x}}_k$ by unpenalized GLM QML.
   (d) For each $j \in \{1, \ldots, p\}$, set
$$\tilde{\kappa}_{k,j} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{h}_j(y_i, \tilde{\mathbf{x}}_{k,i}\tilde{\boldsymbol{\beta}}_k')\right]^2}$$
   where $\tilde{\mathbf{x}}_{k,i}$ is the i(th) observation on $\tilde{\mathbf{x}}_k$.
   (e) Set $k = k + 1$
   (f) If $k > 15$ or the variables in $\tilde{\mathbf{x}}_k$ are the same as those in $\tilde{\mathbf{x}}_{k-1}$ set each $\tilde{\kappa}_j = \tilde{\kappa}_{k,j}$ and exit; else go to step 6a.

2.4 Other methods for selecting the tuning parameters

2.4.1 CV

CV sets the $\kappa_j = 1$ and finds the value of λ that minimizes an estimate of the out-of-sample mean squared error (MSE) of the predictions. This standard method is discussed in many places, including Hastie et al. (2015).

2.4.2 AL

CV is known to include many extra covariates whose coefficients are zero in the model that best approximates the data; see for example Bühlmann and Van de Geer (2011). AL uses multiple steps of CV to reduce this over-selection problem.

The first step is CV. The second step does CV among the covariates selected in the first step. In the second step, the penalty loadings are set to the inverse of the absolute value of the first-step coefficient estimates. Covariates with larger-in-magnitude coefficients are more likely to be included in the second step. Covariates with smaller-in-magnitude coefficients are more likely to be excluded in the second step. See Zou (2006) and Bühlmann and Van de Geer (2011) for details.

2.4.3 Minimizing the BIC

Following Zhang et al. (2010), we define the degrees of freedom in the BIC to be the number of nonzero coefficient estimates in the GLM lasso for a particular value of λ. Our implementation finds the λ_q in the grid of candidate values Λ that produces the smallest value of

$$BIC = -2\sum_{i=1}^{n} Q(y_i, \tilde{\mathbf{w}}_i\tilde{\boldsymbol{\delta}}) + s_{\lambda_q}\ln(n)$$

where the notation in this equation is as follows.

- $\tilde{\mathbf{w}}_i$ is the i(th) observation on the vector of covariates that have nonzero coefficients in the GLM lasso in equation (1), when λ is set to λ_q and $\kappa_j = 1$.
- s_{λ_q} is the number of covariates in $\tilde{\mathbf{w}}_i$.
- $\sum_{i=1}^{n} Q(y_i, \tilde{\mathbf{w}}_i\tilde{\boldsymbol{\delta}})$ is the value of the unpenalized GLM log likelihood function that includes only the covariates in $\tilde{\mathbf{w}}_i$
- $\tilde{\boldsymbol{\delta}}$ are the coefficients on $\tilde{\mathbf{w}}_i$ that maximize the unpenalized log likelihood function.

3 Simulations

For each design, the simulations estimate the coverage rate of the PO Poisson estimator for three coefficients when the lasso tuning parameters are selected by the plug-in method, by CV, by AL, by and minimizing the BIC. The plug-in-based PO estimator and the BIC-based PO estimator produce either nominal 5% coverage or close to nominal coverage for all sample sizes and designs. The plug-in-based PO estimator performs better than the BIC-based PO estimator in the smaller sample size of 1,000. The CV-based PO estimator and the AL-based PO estimator produce coverage rates that far exceeded the nominal 5% level for the smaller sample size. Their coverage rates for the larger sample size of 3,000 were closer to nominal but still higher than 5% for many of the designs. The details are in tables 1-4.

The first coefficient is a large-in-magnitude coefficient, the second is a small-in-magnitude coefficient, and the third is zero coefficient. For each design, the values of the large-in-magnitude and small-in-magnitude coefficients are set relative to their standard errors, when they are estimated by Poisson QML in an oracle model that only includes the covariates with nonzero coefficients. In each design, the large-in-magnitude coefficient has a value of about four times its oracle-model standard error. In each design, the small-in-magnitude coefficient has a value of about its oracle-model standard error.

For each design, the simulations also estimate quantiles that describe the selection tendencies of the lasso when the tuning parameters are selected by the plug-in method, by CV, by AL, and by minimizing the BIC. Knowing these selection tendencies can help researchers better use lasso-based methods. In summary, the plug-in has the highest risk of not including a covariate that has a large coefficient. The BIC has some risk of not including a covariate that has a large coefficient. CV and AL have almost no risk of not including a covariate that has a large coefficient. The plug-in includes the fewest covariates with zero coefficients. The BIC includes a few more covariates with zero coefficients than the plug-in. CV and AL include many covariates with zero coefficients.

These results explain why the CV-based and AL-based estimators have problems in smaller samples. The CV-based estimators and AL-based estimators have problems in smaller samples because of the large number covariates with zero coefficients that they include. Estimating the coefficients on these covariates slows the rate of convergence of the CV-based PO estimator and the AL-based PO estimator.

Tables 5-8 contain the simulation results that describe the selection tendencies.

For each design, the simulations also estimate the coverage rates of the naive two-step Poisson estimator for three coefficients when the lasso tuning parameters are selected by the plug-in method, by CV, by AL, and by minimizing the BIC. The first step in the naive two-step Poisson estimator is to use a lasso to select which covariates should be included in a subsequent QML Poisson model. The covariates of interest of interest are always included in the lasso. In the second step, the coefficients on the covariates

of interest and the selected covariates are estimated by QML. As mentioned, Leeb and Pötscher (2008) show that naive estimators like this do not have an asymptotic normal distribution when there are small coefficients in the model. For each design, the naive estimators using each tuning-parameter selection technique produce coverage rates that far exceed the nominal 5% level. The detailed results are in tables 9-12.

We note two other aspects of the simulation designs. First, they include many covariates with zero coefficients. Second, the number of covariates with large nonzero coefficients varies from an easy-to-handle number to a number that borders on being not sparse. All the details of the how the data for each design were generated are discussed in section A.

4 Tables

The simulation results are in the tables in this section. All the simulation results are based on 1,200 repetitions.

There are three subsections. The first contains the coverage results for the PO estimators. The second describes the selection tendencies of the lasso. The third contains results for the coverage results of the naive estimators. The same designs were used for the coverage results and tendency results. Here we discuss aspects of the tables that are common to the tables in each subsection.

For each design, p is the total number covariates in the model and n is the sample size. The "No. of large" column contains the number of covariates with large-in-magnitude coefficients in the design. The "No. of small" column contains the number of covariates with small-in-magnitude coefficients in the design.

4.1 PO estimators

The tables in this section contain the estimated coverage rates of the plug-in-based PO Poisson estimator, the CV-based PO Poisson estimator, the AL-based PO Poisson estimator, and the BIC-based PO Poisson estimator. Nominal coverage is 5%, which would be 0.05 in the tables.

In each table, the results for "Large RP" are the rejection proportions of a Wald test against the null hypothesis that the first large-in-magnitude coefficient in the model equals its true value. The rejection proportion for each PO estimator is presented along side the rejection proportion of the same test produced by the QML estimator in the oracle model.

In each table, the results for "Small RP" are the rejection proportions of a Wald test against the null hypothesis that the first small-in-magnitude coefficient in the model equals its true value. The rejection proportion for each PO estimator is presented along

side the rejection proportion of the same test produced by the QML estimator in the oracle model.

In each table, the results for "Zero RP" are the rejection proportions of a Wald test against the null hypothesis that the first zero coefficient in the model coefficient equals zero. The oracle model does not estimate coefficients on covariates with coefficients of zero, so there are no oracle results.

Table 1: plug-in PO

		No. of	No. of	Large RP		Small RP		Zero RP
p	n	large	small	oracle	plug-in PO	oracle	plug-in PO	plug-in PO
500	1000	10	3	0.052	0.056	0.048	0.050	0.062
500	3000	10	3	0.058	0.052	0.050	0.052	0.049
1000	1000	10	3	0.040	0.052	0.049	0.048	0.054
1000	3000	10	3	0.052	0.058	0.049	0.053	0.048
2000	1000	10	3	0.056	0.058	0.044	0.048	0.042
2000	3000	10	3	0.052	0.048	0.058	0.062	0.059
500	1000	15	3	0.052	0.061	0.055	0.045	0.066
500	3000	15	3	0.050	0.048	0.049	0.052	0.057
1000	1000	15	3	0.052	0.053	0.052	0.053	0.053
1000	3000	15	3	0.040	0.048	0.054	0.062	0.052
2000	1000	15	3	0.051	0.057	0.048	0.041	0.061
2000	3000	15	3	0.057	0.061	0.058	0.052	0.041
500	1000	20	4	0.046	0.050	0.051	0.063	0.057
500	3000	20	4	0.056	0.052	0.052	0.052	0.052
1000	1000	20	4	0.054	0.067	0.050	0.060	0.060
1000	3000	20	4	0.041	0.043	0.046	0.056	0.052
2000	1000	20	4	0.062	0.065	0.048	0.058	0.048
2000	3000	20	4	0.058	0.044	0.052	0.062	0.049

Table 2: CV PO

p	n	No. of large	No. of small	Large RP oracle	Large RP CV PO	Small RP oracle	Small RP CV PO	Zero RP CV PO
500	1000	10	3	0.052	0.118	0.048	0.088	0.068
500	3000	10	3	0.058	0.072	0.050	0.059	0.055
1000	1000	10	3	0.040	0.113	0.049	0.087	0.066
1000	3000	10	3	0.052	0.077	0.049	0.070	0.056
2000	1000	10	3	0.056	0.128	0.044	0.103	0.066
2000	3000	10	3	0.052	0.077	0.058	0.073	0.062
500	1000	15	3	0.052	0.132	0.055	0.069	0.083
500	3000	15	3	0.050	0.085	0.049	0.057	0.058
1000	1000	15	3	0.052	0.129	0.052	0.107	0.076
1000	3000	15	3	0.040	0.080	0.054	0.076	0.052
2000	1000	15	3	0.051	0.142	0.048	0.120	0.072
2000	3000	15	3	0.057	0.086	0.058	0.079	0.056
500	1000	20	4	0.046	0.147	0.051	0.117	0.093
500	3000	20	4	0.056	0.100	0.052	0.083	0.067
1000	1000	20	4	0.054	0.157	0.050	0.128	0.077
1000	3000	20	4	0.041	0.083	0.046	0.081	0.067
2000	1000	20	4	0.062	0.158	0.048	0.128	0.076
2000	3000	20	4	0.058	0.103	0.052	0.105	0.049

Table 3: AL PO

p	n	No. of large	No. of small	Large RP		Small RP		Zero RP
				oracle	AL PO	oracle	AL PO	AL PO
500	1000	10	3	0.052	0.115	0.048	0.083	0.068
500	3000	10	3	0.058	0.062	0.050	0.058	0.052
1000	1000	10	3	0.040	0.104	0.049	0.085	0.066
1000	3000	10	3	0.052	0.072	0.049	0.071	0.056
2000	1000	10	3	0.056	0.127	0.044	0.101	0.054
2000	3000	10	3	0.052	0.069	0.058	0.077	0.063
500	1000	15	3	0.052	0.119	0.055	0.072	0.083
500	3000	15	3	0.050	0.068	0.049	0.057	0.056
1000	1000	15	3	0.052	0.111	0.052	0.108	0.076
1000	3000	15	3	0.040	0.072	0.054	0.075	0.055
2000	1000	15	3	0.051	0.128	0.048	0.105	0.074
2000	3000	15	3	0.057	0.079	0.058	0.066	0.056
500	1000	20	4	0.046	0.129	0.051	0.117	0.083
500	3000	20	4	0.056	0.087	0.052	0.078	0.062
1000	1000	20	4	0.054	0.143	0.050	0.107	0.085
1000	3000	20	4	0.041	0.074	0.046	0.076	0.062
2000	1000	20	4	0.062	0.142	0.048	0.134	0.070
2000	3000	20	4	0.058	0.092	0.052	0.083	0.052

Table 4: BIC PO

p	n	No. of large	No. of small	Large RP oracle	Large RP BIC PO	Small RP oracle	Small RP BIC PO	Zero RP BIC PO
500	1000	10	3	0.052	0.070	0.048	0.052	0.058
500	3000	10	3	0.058	0.062	0.050	0.052	0.049
1000	1000	10	3	0.040	0.049	0.049	0.052	0.057
1000	3000	10	3	0.052	0.057	0.049	0.049	0.048
2000	1000	10	3	0.056	0.064	0.044	0.049	0.051
2000	3000	10	3	0.052	0.052	0.058	0.064	0.062
500	1000	15	3	0.052	0.076	0.055	0.047	0.065
500	3000	15	3	0.050	0.044	0.049	0.051	0.058
1000	1000	15	3	0.052	0.070	0.052	0.061	0.058
1000	3000	15	3	0.040	0.056	0.054	0.062	0.058
2000	1000	15	3	0.051	0.058	0.048	0.056	0.063
2000	3000	15	3	0.057	0.055	0.058	0.058	0.042
500	1000	20	4	0.046	0.066	0.051	0.077	0.065
500	3000	20	4	0.056	0.062	0.052	0.058	0.055
1000	1000	20	4	0.054	0.075	0.050	0.080	0.065
1000	3000	20	4	0.041	0.047	0.046	0.054	0.051
2000	1000	20	4	0.062	0.084	0.048	0.068	0.057
2000	3000	20	4	0.058	0.051	0.052	0.066	0.048

4.2 Selected covariate counts

The tables in this section summarize the selection tendencies of the lasso when the tuning parameters are selected by the plug-in method, by CV, by AL, and by minimizing the BIC.

Each table presents results for a lasso of the dependent variable on all the potential covariates when the tuning parameters were selected using the specified method.

The "Miss large" columns contain the 50(th), 90(th), 95(th), and 99(th) quantiles of the distribution of the number of covariates with large-in-magnitude coefficients that were not found by the lassos.

The "Miss small" columns contain the 50(th), 90(th), 95(th), and 99(th) quantiles of the distribution of the number of covariates with small-in-magnitude coefficients that were not found by the lassos.

The "Found zeros" columns contain the 50(th), 90(th), 95(th), and 99(th) quantiles of the distribution of the number of covariates with zero coefficients that were included

by the lassos.

Table 5: Selection counts using Plugin

		No. of	No. of	Miss large				Miss small				Found zeros			
p	n	large	small	p50	p90	p95	p99	p50	p90	p95	p99	p50	p90	p95	p99
500	1000	10	3	0	1	2	3	2	2	2	2	0	0	0	1
500	3000	10	3	0	0	0	0	2	2	2	2	0	0	0	0
1000	1000	10	3	0	1	2	3	2	2	2	2	0	0	1	1
1000	3000	10	3	0	0	0	0	2	2	2	2	0	0	0	0
2000	1000	10	3	0	2	2	4	2	2	2	2	0	0	1	2
2000	3000	10	3	0	0	0	1	2	2	2	2	0	0	0	0
500	1000	15	3	1	2	3	6	2	2	2	2	0	0	1	1
500	3000	15	3	0	0	0	1	2	2	2	2	0	0	0	1
1000	1000	15	3	1	2	4	7	2	2	2	2	0	0	1	1
1000	3000	15	3	0	0	0	1	2	2	2	2	0	0	0	1
2000	1000	15	3	1	3	4	9	2	2	2	2	0	0	1	2
2000	3000	15	3	0	0	0	1	2	2	2	2	0	0	0	1
500	1000	20	4	1	4	6	12	3	3	3	3	0	1	1	2
500	3000	20	4	0	0	0	5.5	3	3	3	3	0	0	0	1
1000	1000	20	4	1	5	7.5	13	3	3	3	3	0	1	1	3
1000	3000	20	4	0	0	1	6.5	3	3	3	3	0	0	1	1
2000	1000	20	4	1	6	8	14	3	3	3	3	0	1	1.5	3
2000	3000	20	4	0	0	1	7	3	3	3	3	0	0	1	2

Table 6: Selection counts using CV

		No. of	No. of	Miss large				Miss small				Found zeros			
p	n	large	small	p50	p90	p95	p99	p50	p90	p95	p99	p50	p90	p95	p99
500	1000	10	3	0	0	0	0	1	2	2	2	19	42.5	52	72
500	3000	10	3	0	0	0	0	1	2	2	2	17	39	47	67
1000	1000	10	3	0	0	0	0	1	2	2	2	20	47	63	85.5
1000	3000	10	3	0	0	0	0	1	2	2	2	18	46	57.5	78
2000	1000	10	3	0	0	0	0	1	2	2	2	24	58	71	97.5
2000	3000	10	3	0	0	0	0	1	2	2	2	20	49	64	86.5
500	1000	15	3	0	0	0	0	1	2	2	2	21	45	52	66.5
500	3000	15	3	0	0	0	0	1	2	2	2	21	42	50	64
1000	1000	15	3	0	0	0	1	1	2	2	2	27	56	66	90
1000	3000	15	3	0	0	0	0	1	2	2	2	23	50	60	78
2000	1000	15	3	0	0	0	1	1	2	2	2	31	63	75	109.5
2000	3000	15	3	0	0	0	1	1	2	2	2	25	59	72.5	93
500	1000	20	4	0	0	0	.5	0	1	2	2	29	52	60	78
500	3000	20	4	0	0	0	0	1	2	2	2	26	48	57	73
1000	1000	20	4	0	0	0	1	1	1	2	2	33	62	73	95
1000	3000	20	4	0	0	0	1	1	2	2	2	30	58	66.5	95
2000	1000	20	4	0	0	0	1	1	2	2	3	37	67	82	110
2000	3000	20	4	0	0	0	1	1	2	2	3	33	66	81	110

Table 7: Selection counts using Adapt

		No. of	No. of	Miss large				Miss small				Found zeros			
p	n	large	small	p50	p90	p95	p99	p50	p90	p95	p99	p50	p90	p95	p99
500	1000	10	3	0	0	0	0	1	2	2	2	14	33.5	40.5	54
500	3000	10	3	0	0	0	0	1	2	2	2	12	29	35	50.5
1000	1000	10	3	0	0	0	0	1	2	2	2	16	37	47	66.5
1000	3000	10	3	0	0	0	0	1	2	2	2	14	34	42	57.5
2000	1000	10	3	0	0	0	0	1	2	2	2	18	43	53	72
2000	3000	10	3	0	0	0	0	1	2	2	2	15	37	48	62.5
500	1000	15	3	0	0	0	0	1	2	2	2	17	34	40	50
500	3000	15	3	0	0	0	0	1	2	2	2	13	27	32.5	43
1000	1000	15	3	0	0	0	1	1	2	2	2	21	42	50	67
1000	3000	15	3	0	0	0	0	1	2	2	2	14	32	38.5	51
2000	1000	15	3	0	0	0	1	1	2	2	2	24	47	55	77
2000	3000	15	3	0	0	0	1	1	2	2	2	16	37	46	61.5
500	1000	20	4	0	0	0	1	1	2	2	2	21	38	44	59
500	3000	20	4	0	0	0	0	1	2	2	3	13	25	30	40.5
1000	1000	20	4	0	0	0	1	1	2	2	3	24	44	52	64
1000	3000	20	4	0	0	0	1	1	2	3	3	14	30	35	48
2000	1000	20	4	0	0	0	1	1	2	2	3	26	48	56	78
2000	3000	20	4	0	0	0	1	1	2	3	3	16	34	42	58

Table 8: Selection counts using BIC

		No. of	No. of	Miss large				Miss small				Found zeros			
p	n	large	small	p50	p90	p95	p99	p50	p90	p95	p99	p50	p90	p95	p99
500	1000	10	3	0	0	0	0	1	2	2	2	1	4	4	6
500	3000	10	3	0	0	0	0	2	2	2	2	1	2	3	5
1000	1000	10	3	0	0	0	0	2	2	2	2	1	3	4	6
1000	3000	10	3	0	0	0	0	2	2	2	2	1	2	3	4.5
2000	1000	10	3	0	0	0	1	2	2	2	2	1	4	5	7.5
2000	3000	10	3	0	0	0	0	2	2	2	2	1	2	3	4
500	1000	15	3	0	0	0	1	1	2	2	2	2	5	6	8
500	3000	15	3	0	0	0	0	2	2	2	2	1	3	4	5
1000	1000	15	3	0	0	1	1	2	2	2	2	2	5	6	8
1000	3000	15	3	0	0	0	0	2	2	2	2	1	3	4	5
2000	1000	15	3	0	0	1	1	2	2	2	2	2	5	6	8
2000	3000	15	3	0	0	0	1	2	2	2	2	1	3	4	6
500	1000	20	4	0	0	1	1	1	2	3	3	3	6	8	10
500	3000	20	4	0	0	0	0	1	3	3	3	1	4	5	7
1000	1000	20	4	0	0	1	1	1	2	3	3	3	6	8	11
1000	3000	20	4	0	0	0	1	2	3	3	3	2	4	5	7
2000	1000	20	4	0	1	1	1	2	3	3	3	3	7	8	12
2000	3000	20	4	0	0	0	1	2	3	3	3	2	5	5	7

4.3 Naive estimators

The tables in this section contain the estimated coverage rates of the plug-in-based naive Poisson estimator, the CV-based naive Poisson estimator, the AL-based naive Poisson estimator, and the BIC-based naive Poisson estimator. Nominal coverage is 5%, which would be 0.05 in the tables.

In each table, the results for "Large RP" are the rejection proportions of a Wald test against the null hypothesis that the first large-in-magnitude coefficient in the model equals its true value. The rejection proportion for each naive estimator is presented along side the rejection proportion of the same test produced by the QML estimator in the oracle model.

In each table, the results for "Small RP" are the rejection proportions of a Wald test against the null hypothesis that the first small-in-magnitude coefficient in the model equals its true value. The rejection proportion for each naive estimator is presented

along side the rejection proportion of the same test produced by the QML estimator in the oracle model.

In each table, the results for "Zero RP" are the rejection proportions of a Wald test against the null hypothesis that the first zero coefficient in the model coefficient equals zero. The oracle model does not estimate coefficients on covariates with coefficients of zero, so there are no oracle results.

Table 9: plug-in Naive

		No. of	No. of	Large RP		Small RP		Zero RP
p	n	large	small	oracle	plug-in Naive	oracle	plug-in Naive	plug-in Naive
500	1000	10	3	0.052	0.198	0.048	0.278	0.347
500	3000	10	3	0.058	0.058	0.050	0.183	0.265
1000	1000	10	3	0.040	0.205	0.049	0.316	0.347
1000	3000	10	3	0.052	0.052	0.049	0.173	0.256
2000	1000	10	3	0.056	0.242	0.044	0.346	0.338
2000	3000	10	3	0.052	0.055	0.058	0.189	0.278
500	1000	15	3	0.052	0.268	0.055	0.389	0.344
500	3000	15	3	0.050	0.062	0.049	0.204	0.270
1000	1000	15	3	0.052	0.296	0.052	0.411	0.360
1000	3000	15	3	0.040	0.049	0.054	0.205	0.259
2000	1000	15	3	0.051	0.345	0.048	0.437	0.343
2000	3000	15	3	0.057	0.071	0.058	0.209	0.258
500	1000	20	4	0.046	0.332	0.051	0.470	0.438
500	3000	20	4	0.056	0.092	0.052	0.268	0.351
1000	1000	20	4	0.054	0.392	0.050	0.517	0.491
1000	3000	20	4	0.041	0.082	0.046	0.255	0.378
2000	1000	20	4	0.062	0.412	0.048	0.583	0.546
2000	3000	20	4	0.058	0.102	0.052	0.268	0.388

Table 10: CV Naive

		No. of	No. of	Large RP		Small RP		Zero RP
p	n	large	small	oracle	CV Naive	oracle	CV Naive	CV Naive
500	1000	10	3	0.052	0.095	0.048	0.101	0.122
500	3000	10	3	0.058	0.056	0.050	0.083	0.091
1000	1000	10	3	0.040	0.090	0.049	0.138	0.155
1000	3000	10	3	0.052	0.065	0.049	0.101	0.107
2000	1000	10	3	0.056	0.115	0.044	0.181	0.155
2000	3000	10	3	0.052	0.069	0.058	0.133	0.140
500	1000	15	3	0.052	0.082	0.055	0.102	0.114
500	3000	15	3	0.050	0.072	0.049	0.092	0.093
1000	1000	15	3	0.052	0.103	0.052	0.133	0.129
1000	3000	15	3	0.040	0.072	0.054	0.107	0.106
2000	1000	15	3	0.051	0.128	0.048	0.163	0.158
2000	3000	15	3	0.057	0.094	0.058	0.135	0.132
500	1000	20	4	0.046	0.102	0.051	0.108	0.088
500	3000	20	4	0.056	0.080	0.052	0.089	0.087
1000	1000	20	4	0.054	0.111	0.050	0.120	0.112
1000	3000	20	4	0.041	0.059	0.046	0.096	0.107
2000	1000	20	4	0.062	0.126	0.048	0.192	0.165
2000	3000	20	4	0.058	0.102	0.052	0.134	0.117

Table 11: AL Naive

p	n	No. of large	No. of small	Large RP		Small RP		Zero RP
				oracle	AL Naive	oracle	AL Naive	AL Naive
500	1000	10	3	0.052	0.083	0.048	0.107	0.124
500	3000	10	3	0.058	0.055	0.050	0.095	0.099
1000	1000	10	3	0.040	0.085	0.049	0.139	0.163
1000	3000	10	3	0.052	0.060	0.049	0.107	0.115
2000	1000	10	3	0.056	0.108	0.044	0.168	0.158
2000	3000	10	3	0.052	0.063	0.058	0.134	0.156
500	1000	15	3	0.052	0.077	0.055	0.107	0.120
500	3000	15	3	0.050	0.064	0.049	0.096	0.097
1000	1000	15	3	0.052	0.101	0.052	0.137	0.139
1000	3000	15	3	0.040	0.068	0.054	0.120	0.124
2000	1000	15	3	0.051	0.124	0.048	0.177	0.158
2000	3000	15	3	0.057	0.076	0.058	0.145	0.136
500	1000	20	4	0.046	0.093	0.051	0.117	0.097
500	3000	20	4	0.056	0.072	0.052	0.093	0.102
1000	1000	20	4	0.054	0.108	0.050	0.120	0.130
1000	3000	20	4	0.041	0.057	0.046	0.098	0.147
2000	1000	20	4	0.062	0.127	0.048	0.188	0.168
2000	3000	20	4	0.058	0.086	0.052	0.129	0.133

Table 12: BIC Naive

p	n	No. of large	No. of small	Large RP oracle	Large RP BIC Naive	Small RP oracle	Small RP BIC Naive	Zero RP BIC Naive
500	1000	10	3	0.052	0.061	0.048	0.128	0.212
500	3000	10	3	0.058	0.055	0.050	0.142	0.191
1000	1000	10	3	0.040	0.046	0.049	0.172	0.228
1000	3000	10	3	0.052	0.052	0.049	0.153	0.189
2000	1000	10	3	0.056	0.070	0.044	0.186	0.224
2000	3000	10	3	0.052	0.051	0.058	0.165	0.228
500	1000	15	3	0.052	0.067	0.055	0.151	0.181
500	3000	15	3	0.050	0.055	0.049	0.151	0.173
1000	1000	15	3	0.052	0.077	0.052	0.163	0.225
1000	3000	15	3	0.040	0.048	0.054	0.142	0.187
2000	1000	15	3	0.051	0.079	0.048	0.195	0.212
2000	3000	15	3	0.057	0.063	0.058	0.174	0.184
500	1000	20	4	0.046	0.071	0.051	0.124	0.125
500	3000	20	4	0.056	0.063	0.052	0.115	0.139
1000	1000	20	4	0.054	0.087	0.050	0.133	0.167
1000	3000	20	4	0.041	0.046	0.046	0.116	0.182
2000	1000	20	4	0.062	0.095	0.048	0.174	0.206
2000	3000	20	4	0.058	0.058	0.052	0.147	0.177

5 Conclusion and future research

This paper extended the Belloni et al. (2012) plug-in algorithm for choosing the lasso tuning parameters to high-dimensional GLMs. It presented simulation evidence that that a PO Poisson estimator that uses this plug-in algorithm performs well. The simulations also show that a PO Poisson estimator that uses CV or AL to select the lasso tuning parameters can require a much larger sample size to perform as well. The simulations also show that a PO Poisson estimator that selects the lasso tuning parameters by minimizing the BIC performs almost as well as the plug-in based estimator. Finally, the paper explains these simulation results by discussing the covariate-selection tendencies of the Poisson lasso when the tuning parameters are selected by the plug-in method, by CV, by AL, and by the BIC.

We are currently extending the plug-in method and algorithm discussed in Belloni et al. (2012) and Belloni et al. (2016a) to case of unbalanced panels for GLMs.

A Simulation Designs

The designs vary by sample size, the number of potential covariates, and the specification for the nonzero coefficients. There are 2 sample sizes, 3 values for the number of potential covariates, and 3 specifications for the nonzero covariates. Thus, there are 18 ($= 2*3*3$) designs in total. The 2 sample sizes are 1000, and 3000. The 3 values for the number of potential covariates are 500, 1000, and 2000. The 3 specifications for the nonzero coefficients are

1. 10 covariates with large coefficients and 3 covariates with small coefficients,
2. 15 covariates with large coefficients and 3 covariates with small coefficients, and
3. 20 covariates with large coefficients and 4 covariates with small coefficients.

The i(th) observation of the dependent variable y_i was generated from a Poisson distribution with mean $\exp(\mathbf{x}_i\boldsymbol{\beta}_c')$, where the variables in $\mathbf{x}_i$ come from the skewed, asymmetric distribution described below and $\boldsymbol{\beta}_c$ the vector of coefficients for case $c \in \{1, 2, \ldots, 18\}$. In each case, $\boldsymbol{\beta}_c$ contains some large coefficients, some small coefficients, and many zero coefficients. Table 13 gives the values of the coefficients for each design. In each design, each small coefficient was set to a value close to twice its standard error in the Poisson regression model that includes only the covariates with large and small coefficients. (This model is sometimes called the "oracle" model in the literature.) In each design, each large coefficient was set to be about twice the value of a small coefficient in that design.

In each design, there are p covariates in $\mathbf{x}$. For each design, a Toeplitz covariance matrix $\mathbf{V}$ is constructed from the $p \times 1$ vector $\mathbf{r}$, where the j(th) element of $\mathbf{r}$ is $j^(-1.1)$. And we let the Cholesky factor of this be $\mathbf{L}$.

For each repetition in each design, p variables ($\mathbf{w}$) of sample size n are drawn from χ-squared distribution with 15 degrees of freedom. Each of these variables in $\mathbf{w}$ is then normalized by removing it's mean of 15 and dividing it by it's standard deviation of $\sqrt{30}$. For each draw, $\mathbf{x} = \mathbf{w}\mathbf{L}'$.

Table 13: Coefficients by design

p	n	No of big	No of small	Value of big	Value of small
500	1000	10	3	.16	.06
500	3000	10	3	.13	.035
1000	1000	10	3	.16	.06
1000	3000	10	3	.13	.035
2000	1000	10	3	.16	.06
2000	3000	10	3	.13	.035
500	1000	15	3	.13	.06
500	3000	15	3	.13	.032
1000	1000	15	3	.13	.06
1000	3000	15	3	.13	.032
2000	1000	15	3	.13	.06
2000	3000	15	3	.13	.032
500	1000	20	4	.12	.06
500	3000	20	4	.13	.03
1000	1000	20	4	.12	.06
1000	3000	20	4	.13	.03
2000	1000	20	4	.12	.06
2000	3000	20	4	.13	.03

B Details of PO estimator

Belloni et al. (2016b) derived PO estimators for GLM model. Algorithm 2 specifies the version of the Belloni et al. (2016b) PO estimator we use.

Algorithm 2: PO Poisson estimation

1. Run a Poisson lasso of y on $\mathbf{d}$ and $\mathbf{x}$, and let $\widetilde{\mathbf{x}}$ be the subset of the $\mathbf{x}$ covariates that have nonzero estimated coefficients.

    - Our results for the plug-in PO Poisson estimator use our version of the plug-in method to select the lasso tuning parameters in this lasso. Similarly, our results for the CV PO Poisson estimator, the AL PO Poisson estimator, and BIC PO Poisson estimator respectively use CV, AL, or minimizing the BIC to select the lasso tuning parameters.

2. Use the unpenalized quasi maximum likelihood Poisson regression estimator to estimate the coefficients $\widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\beta}}$ in a Poisson 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 $G(\cdot)$.

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

    - Our results for the plug-in PO Poisson estimator use the heteroskedastic plug-in method of Belloni et al. (2012) to select the lasso tuning parameters in this lasso. Similarly, our results for the CV PO Poisson estimator, the AL PO Poisson estimator, and BIC PO Poisson estimator respectively use CV, AL, or minimizing the BIC to select the lasso tuning parameters.

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 ω_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 instrumental variables $\mathbf{z} = (\widetilde{d}_1, \ldots, \widetilde{d}_J)$ and $\mathbf{z}_i$ be the ith observation on this vector of instrumental variables. 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

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

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

C Formulas for $Q()$

Here are the formulas for $Q()$

- For linear models,
$$Q(y_i, \mathbf{w}_i\boldsymbol{\delta}') = (y_i - \mathbf{w}_i\boldsymbol{\delta}')^2$$
- For Poisson models
$$Q(y_i, \mathbf{w}_i\boldsymbol{\delta}') = -[y_i\mathbf{w}_{i,t}\boldsymbol{\delta}' - \exp(\mathbf{w}_i\boldsymbol{\delta}') - \ln(y_i!)]$$
- For logit models
$$Q(y_i, \mathbf{w}_i\boldsymbol{\delta}') = \ln[1 + \exp(\mathbf{w}_i\boldsymbol{\delta}')] - y_i(\mathbf{w}_i\boldsymbol{\delta})$$

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(6): 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen. 2014. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* 81(2): 608–650.

Belloni, A., V. Chernozhukov, C. Hansen, and D. Kozbur. 2016a. Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics* 34(4): 590–605.

Belloni, A., V. Chernozhukov, and Y. Wei. 2016b. Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics* 34(4): 606–619.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov. 2009. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* 37(4): 1705–1732.

Bühlmann, P., and S. Van de Geer. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Publishing Company, Incorporated.

Cattaneo, M. D., M. Jansson, and X. Ma. 2018a. Two-step estimation and inference with possibly many included covariates. *The Review of Economic Studies* 86(3): 1095–1122.

Cattaneo, M. D., M. Jansson, and W. K. Newey. 2018b. Inference in Linear Regression Models with Many Covariates and Heteroskedasticity. *Journal of the American Statistical Association* 113(523): 1350–1361.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1): C1–C68.

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani. 2007. Pathwise coordinate optimization. *The Annals of Applied Statistics* 1(2): 302–332.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* 33(1): 1–22.

Hastie, T., R. Tibshirani, and M. Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Rotaon: CRC Press.

Jing, B.-Y., Q.-M. Shao, and Q. Wang. 2003. Self-normalized Cramér-type large deviations for independent random variables. *The Annals of probability* 31(4): 2167–2215.

Kozbur, D. 2019. Testing-Based Forward Model Selection . URL `https://www..`

Leeb, H., and B. M. Pötscher. 2008. Sparse estimators and the oracle property, or the return of Hodges estimator. *Journal of Econometrics* 142(1): 201–211.

Peña, V. H., T. L. Lai, and Q.-M. Shao. 2009. *Self-normalized processes: Limit theory and Statistical Applications*. Berlin: Springer-Verlag.

Zhang, Y., R. Li, and C.-L. Tsai. 2010. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* 105(489): 312–323.

Zou, H. 2006. The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* 101(476): 1418–1429.