“Plus/minus” confidence intervals and thresholding

P. Zietkiewicz

April 25, 2024

Table of contents

Contents

Thresholding

- p covariates, then 2^p models. Define $I = \{1, , 2, \ldots, p\}$, $J = \{j \in I : \beta_j \neq 0\}$ and $K = \{j \in I : \beta_j = 0\}$. We define a model selection procedure as the estimator $\hat{J} \subseteq I$, which is the set of selected variables.
- $\hat{J}$ is said to have the oracle property if (1) $\mathbb{P}(\hat{J} = J) \to 1$ and (2) the limiting distribution of its subvector corresponding to the non-zero coefficients is the same as if this subvector was known prior to estimation.

Asymptotic normality

Theorem 2.3.1 (Asymptotic normality). *Let $\hat{J}$ be the selected model, suppose that $\mathbb{P}(\hat{J} = J) \to 1$, then re-estimate the parameters*

$$\hat{\beta}^{\hat{J}} = \underset{b \in \mathbb{R}^p}{\arg \mathit{solve}} \left\{ C_{\hat{J}} X^\top W(\eta_{\hat{J}}) u(\eta_{\hat{J}}) = 0 \right\} \tag{2.3.1}$$

where $\eta_{\hat{J}} = XC_{\hat{J}}b$ and $C_{\hat{J}}$ is a diagonal matrix where $[C_{\hat{J}}]_{jj} = 1$ if $j \in \hat{J}$ or $[C_{\hat{J}}]_{jj} = 0$ if $j \in \hat{K}$. We assume the conditions of Wedderburn (1976, Table 1) and that for the ML estimator we have $\phi^{-1/2}(X^\top W X)^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} N(0_p, I_p)$. Then $\hat{\phi}^{-1/2}(X_J^\top W(\eta_J) X_J)^{1/2}(\hat{\beta}^{\hat{J}}[J] - \beta[J]) \xrightarrow{d} N(0_{|J|}, I_{|J|})$.

Wald statistics

- $\hat{s}_j = O_p(n^{-1/2})$ and $\hat{\beta}_j - \beta_j = O_p(n^{-1/2})$
- Define Wald statistic $z_j = \hat{\beta}_j / \hat{s}_j$

Lemma 2.3.2 (Growth rate of asymptotic Wald statistics). *Under the assumptions of Lemma 2.3.1* $|z_j| = O_p(n^{1/2})$ *if* $j \in J$*, and* $|z_j| = O_p(1)$ *if* $j \in K$.

Thresholding for GLMs with ML

Theorem 2.3.2 (Thresholding for GLMs). *Let $\hat{J} = \{j \in I : |z_j| \geq g(n, \gamma)\}$ and $\hat{K} = \{j \in I : |z_j| < g(n, \gamma)\}$ be estimates of J and K respectively, where $g(n, \gamma) = O_p(n^\gamma)$, where $\gamma \in (0, 1/2)$ is some threshold. Then $\mathbb{P}(\hat{J} = J) \to 1$.*

Corollary 2.3.1 (Oracle property for GLMs by ML). *Combining Theorems 2.3.1 and 2.3.2 we get the oracle property for thresholding.*

Remark 1. In the developments thus far we have used the ML estimator. These results hold for other estimators which are $\sqrt{n}$-consistent and have the same asymptotic distribution as the ML estimator. For example, estimators from bias-reducing estimating equations (Kosmidis et al., 2020) have these properties (Firth, 1993). Additionally in logistic regression, those estimators guarantee finite estimates under the sole requirement that the design matrix is full rank (Kosmidis and Firth, 2021).

Optimised threshold

- Wide range of thresholds that would result in a consistent model selection procedures, for example $g(n, 1/4) = n^{1/4}$ and $g(n, 1/3) = n^{1/3}$ would both suffice.
- Let γ_j and g_j denote the coefficient-specific rate and threshold function respectively. We propose to minimise the quantity

$$\omega_j \mathbb{P}(|z_j| > g_j : j \in K) + (1 - \omega_j)\mathbb{P}(|z_j| \leq g_j : j \in J) \tag{1}$$

  with respect to $\gamma \in (0, 1/2)$.
- Propose statistic from Derryberry et al. (2018)

$$\text{dbic}_j = n \log\left(\frac{z_j^2}{n-p} + 1\right) - \log(n) \tag{2}$$

  and use $\omega_j = I(\text{dbic}_j < 0)$ and $\omega_j = \Phi(-\text{dbic}_j)$ as hard and soft weights respectively.

Binomial sparse

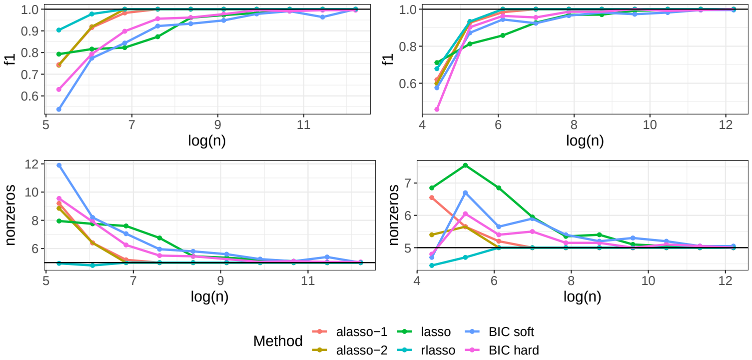


Figure: f1 score and number of non-zero variables in binomial logistic regression with increasing observations with $p = 100$ variables and $s = 5$ non-zero variables (left) $p = 40$ variables and $s = 5$ non-zero variables (right).

Binomial dense

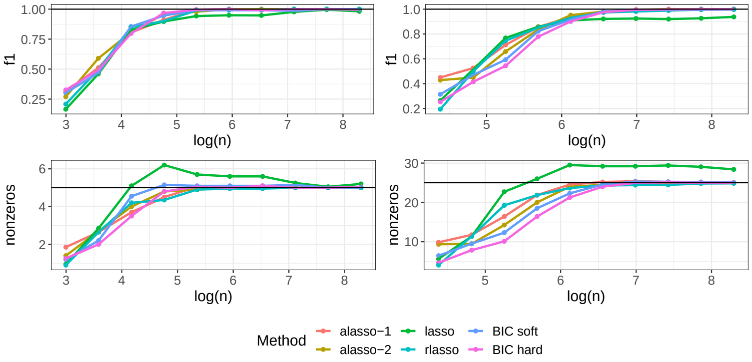


Figure: f1 score and number of non-zero variables in binomial logistic regression with increasing observations with $p = 10$ variables and $s = 5$ non-zero variables (left) $p = 40$ variables and $s = 25$ non-zero variables (right).

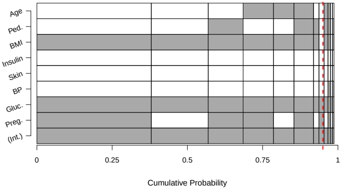
Overall conclusion from this

Overall, we observe that both soft and hard thresholding procedures converge to the true model, although they may not do so as rapidly as the relaxed or adaptive Lasso in the sparse setting. The performance gap between thresholding and relaxed or adaptive Lasso narrows in denser settings when there is a higher proportion of non-zero entries. The broader point is that thresholding is a viable method for consistent model selection whilst being directly implementable as part of the standard maximum likelihood output, making it more accessible and convenient for practitioners.

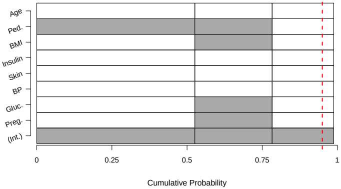
Confidence sets

- Thresholding allows for the direct use of nonparametric bootstrap for the construction of confidence sets of models, in order to quantify the uncertainty associated with the selected model.
- Confidence sets using the diabetes data (Smith et al., 1988) after 1000 bootstrap iterations for relaxed Lasso (top left) and adaptive Lasso with penalty 2 (top right), soft BIC (bottom left) and hard BIC (bottom right) thresholding. A model is seen as a column of tiles where grey and white indicate whether a variable is included or excluded respectively. Models are shown in descending order of the proportion they appeared in the bootstrap samples. The dotted red line is at 0.95.

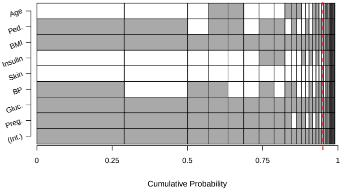
Confidence sets



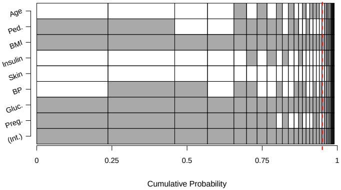
(a) Relaxed Lasso.



(b) Adaptive Lasso with penalty 2.



(c) Soft BIC thresholding.

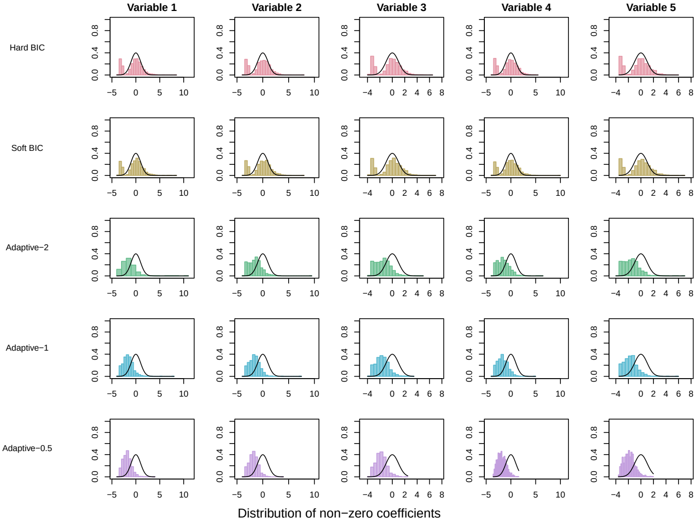


(d) Hard BIC thresholding.

Oracle property

Distribution of standardized non-zero coefficients after model selection using hard and soft BIC thresholding, Hard BIC and Soft BIC respectively, adaptive Lasso with penalties of 2, 1 and 0.5, Adaptive-2, Adaptive-1 and Adaptive-0.5 respectively. Standard normal distribution is superimposed in black. Data is from binomial logistic regression with $\beta_j = 1$ for $j = 1, \ldots, 5$ and $\beta_j = 0$ for $j = 6, \ldots, 40$ with $n = 100$ observations and the simulation is repeated for 1000 iterations.

Oracle property



Summary of Chee et al. (2023)

- Given a dataset (x_i, y_i) for $i \in \{1, \ldots, n\}$ where $x_i \in \mathbb{R}^p$ and y_i is a realization of the random variable Y_i, and a model parameterized by $\theta \in \Theta \subseteq \mathbb{R}^p$, with corresponding log-likelihood function $\ell(\theta)$ and we wish to estimate the parameter θ where the true parameter is $\theta_\star = \text{argmax}_{\theta \in \Theta} \mathbb{E}[\ell(\theta; Y, X)]$. The Fisher information matrix as $F_\star = \mathbb{E}[-\nabla^2 \ell(\theta_\star; Y, X)]$. The basic iteration of SGD is expressed as

$$\theta_{k+1}^{\text{sgd}} = \theta_k^{\text{sgd}} + \varphi_k \nabla \ell(\theta_k^{\text{sgd}}; y_k, x_k)$$

  where $\varphi_k = \varphi_1 k^{-\varphi}$ is a diminishing learning rate with $\varphi_1 > 0$ and $\varphi \in (0.5, 1]$.
- Let θ_n^{sgd} be the one-pass estimator of θ_*.
- Advantages of one-pass over multi-pass: (1) Asymptotic covariance matrix is known in closed form (2) Covariance matrix can be bounded by a factor that depends only on the learning rate φ_1.

Summary of Chee et al. (2023)

- Chee et al. (2023) provide methodology to compute very simple confidence intervals using the one-pass estimate of the form

$$\theta^{\text{sgd}}_{n,j} \pm 2\sqrt{\frac{\varphi_1^*}{n}} \quad j = 1, \ldots, p \tag{3}$$

  where $\theta^{\text{sgd}}_{n,j}$ is the jth element of θ^{sgd}_{n} and φ_1^* is a tuned hyperparameter. Importantly Chee et al. (2023, Theorems 3.1, 3.2) show that the the confidence intervals (3) are asymptotically valid.

- Define $\Sigma_* = \varphi_1^2(2\varphi_1 F_* - I)^{-1} F_*$ where φ_1 is large enough such that $2\varphi_1 F_* - I \succ 0$. And has eigenvalues

$$\text{eigen}(\Sigma_*) = \{\frac{2\varphi_1^2\lambda_j}{2\varphi_1\lambda_j - 1} : j = 1, \ldots, p\}$$

  where λ_j is the jth eigenvalue of F_*.

Results:

Theorem 3.1. *Let $\theta_{N,j}$, denote the j-th component of θ_N in Eq. (4), for $j = 1, \ldots, p$. Suppose that $\gamma_1^* \geq 1/\min_j\{\lambda_j\}$, then $\gamma_1^* I - \Sigma_\star \succ 0$. Define the interval*

$$C_{N,j}(D_N) = \left[\theta_{N,j} - z_{\frac{\alpha}{2}}\sqrt{\frac{\gamma_1^*}{N}},\ \theta_{N,j} + z_{\frac{\alpha}{2}}\sqrt{\frac{\gamma_1^*}{N}}\right], \quad (9)$$

where $z_{\frac{\alpha}{2}} = \Phi^{-1}(1-\alpha/2)$ is the critical value of the standard normal. Then, for every $j = 1, \ldots, p$,

$$\liminf_{N\to\infty} P\big(\theta_{\star,j} \in C_{N,j}(D_N)\big) \geq 1 - \alpha. \quad (10)$$

Theorem 3.2. *Let θ_N be the one-pass SGD in Eq. (4), and suppose that $\gamma_1^* \geq 1/\min_j\{\lambda_j\}$. Define the following confidence region:*

$$\widehat{\Theta} = \left\{\theta \in \Theta : (1/\gamma_1^*)\ ||\theta - \theta_N||^2 < \chi_{\alpha,p}\right\}, \quad (11)$$

where $\chi_{\alpha,p} = \sup\{x \in \mathbb{R} : P(\chi_p^2 \geq x) \leq \alpha\}$ is the α-critical value of a chi-squared random variable with p degrees of freedom. Then,

$$\liminf_{N\to\infty} P(\theta_\star \in \widehat{\Theta}) \geq 1 - \alpha. \quad (12)$$

Selecting γ_1^*:

Linear asymptote in $\Sigma_\star$. At a high level, the variance bound in Theorem 3.1 holds in the regime where the covariance matrix of θ_N is linear with respect to γ_1. One idea is therefore to try and estimate when such regime has been reached. The idea is visualized in Figure 3. Recall from Eq. (8) that the eigenvalues of $\Sigma_\star$ asymptote to $\gamma_1/2$, and so the trace of $\Sigma_\star$ should asymptote to $p\gamma_1/2$, as shown in the figure. The idea is then to slowly increase the learning rate γ_1 and at the same time monitor the trace of $N\mathrm{Var}(\theta_N)$. When γ_1 is large enough for Theorem 3.1 we expect that a linear regression of $\mathrm{trace}(N\mathrm{Var}(\theta_N))$ with respect to γ_1 will give a coefficient around $p/2$ with high confidence. Only a crude estimate of the variance trace is needed, which can be done via bootstrap. See Appendix D.1 for more details, and a practical example.

An eigenvalue bound. In some settings, an estimate $\tilde{F}$ of $F_\star$ exists that may be too crude to be used directly for inference, but may be acceptable for estimating a bound on $\lambda_{\min}$. Then, an alternative way of selecting γ_1^* is to numerically find the maximum eigenvalue of $\tilde{F}^{-1}$, which implies the minimum eigenvalue of $F_\star$ To this end, we propose using inverse power iteration (Trefethen and Bau III, 1997), which is a simple iterative algorithm. More details of this algorithm and its implementation are in Appendix D.2.

Thresholding and SGD

- In the context of thresholding we define the Wald statistic

$$z_{n,j} = \frac{\beta_{n,j}}{\sqrt{\varphi_1^*/n}}. \tag{4}$$

- all the ingredients of our previous Wald statistic. In the numerator if $\beta_j \in J$ we have that $\beta_{n,j} \xrightarrow{p} \beta_j \neq 0$, and if $j \in K$ $\beta_{n,j} \xrightarrow{p} 0$. Then in the denominator we see that $\sqrt{\varphi_1^*/n} = O(n^{-1/2})$. And so,

$$z_{n,j} = \begin{cases} O_p(n^{1/2}) \text{ if } j \in J \\ O_p(1) \text{ if } j \in K \end{cases}$$

Thresholding and SGD

Theorem 3.4.1 (Thresholding for GLMs with SGD). *Let $\hat{J} = \{j \in I : |z_{n,j}| \geq g(n,\gamma)\}$ and $\hat{K} = \{j \in I : |z_{n,j}| < g(n,\gamma)\}$ be estimates of J and K respectively, where $g(n,\gamma) = O_p(n^{\gamma})$, where $\gamma \in (0, 1/2)$ is some threshold. Then $\mathbb{P}(\hat{J} = J) \to 1$.*

Proof. To simplify notation let $g(n,\gamma) = g$. Then $\mathbb{P}(\max_{j\in K}|z_{n,j}| \geq g) \leq \mathbb{P}(\sum_{j\in K}|z_{n,j}$ $g) = \mathbb{P}(g^{-1}\sum_{j\in K}|z_{n,j}| \geq 1)$. Since $\sum_{j\in K}|z_{n,j}| = O_p(1)$ and $g = O_p(n^{\gamma})$ for $\gamma \in (0, 1/2)$, $g^{-1}\sum_{j\in K}|z_{n,j}| = O_p(n^{-\gamma})$ and we have $O_p(n^{-\gamma}) = o_p(n^{-\gamma+1/2})$ (Kosmidis, 2007, Theorem A.4.2) and so $g^{-1}\sum_{j\in K}|z_{n,j}| \xrightarrow{p} 0$. Therefore, $\mathbb{P}(g^{-1}\sum_{j\in K}|z_{n,j}| \geq 1) \to 0$ and $\mathbb{P}(|z_{n,j}| \leq g \text{ for all } j \in K) \to 1$, and $\mathbb{P}(K \subseteq \hat{K}) \to 1$. Similarly,

$$\mathbb{P}\left(\min_{j\in J}|z_{n,j}| \leq g\right) \leq \mathbb{P}\left(\min_{j\in J}\frac{|\beta_j| - |\beta_{n,j} - \beta_j|}{\sqrt{\varphi_1^*/n}} \leq g\right) \tag{3.4.1}$$

$$\leq \mathbb{P}\left(\max_{j\in J}|\beta_{n,j} - \beta_j| \geq \min_{j\in J}|\beta_j| - g\sqrt{\varphi_1^*/n}\right) \tag{3.4.2}$$

$$\leq \mathbb{P}\left(\sum_{j\in J}|\beta_{n,j} - \beta_j| \geq \min_{j\in J}|\beta_j| - g\sqrt{\varphi_1^*/n}\right) \tag{3.4.3}$$

Inequality (3.4.1) comes from using the reverse triangle inequality[1], inequality (3.4. unrestricts the index and takes the largest possible difference between the two terms, and lastly in line (3.4.3) we use the fact that the sum over a set of positive numbers is larger than the maximum of the set. We have $a = \sum_{j\in J}|\beta_{n,j} - \beta_j| = O_p(n^{-1/2})$ following a similar argument used in Lemma 2.3.2, and $b = \min_{j\in J}|\beta_j| - g\sqrt{\varphi_1^*/n} = \min_{j\in J}|\beta_j| - O_p(n^{\gamma-1/2})$ for $\gamma \in (0, 1/2)$. Since $\gamma - 1/2 < 0$ then $a/b \xrightarrow{p} 0/(1+0) = 0$. Therefore

$$\mathbb{P}\left(\sum_{j\in J}|\beta_{n,j} - \beta_j| \geq \min_{j\in J}|\beta_j| - g\sqrt{\varphi_1^*/n}\right) \to 0$$

and $\mathbb{P}(\min_{j\in J}|z_{n,j}| \leq g) \to 0$, so $\mathbb{P}(|z_{n,j}| \geq g \text{ for all } j \in J) \to 1$. Hence $\mathbb{P}(J \cap \hat{K} \neq \emptyset) \to 0$, thus $\mathbb{P}(\hat{J} = J) \to 1$. □

Simulations

- First we see how it is possible to generate a confidence sets on the fly. At each step of SGD we estimate the model set using thresholding.
- Then look at the oracle property of these estimates.

Binomial, $p = 10$, $s = 5$, $n = 200$

Value	Proportion	Cumulative
(1,2,3,4,5)	0.3481	0.3481
(1,2,4)	0.2873	0.6354
()	0.1823	0.8177
(1,2,3,4)	0.1436	0.9613
(1,4)	0.0221	0.9834
(2,4)	0.0110	0.9944
(2)	0.0055	0.9999

Table: $n = 200$, 95% CS: {(1,2,3,4,5), (1,2,4), (), (1,2,3,4)}

Binomial, $p = 10$, $s = 5$, $n = 2000$

Value	Proportion	Cumulative
(1,2,3,4,5)	0.9132	0.9132
(1,3,4,5)	0.0424	0.9556
()	0.0162	0.9718
(1,4,5)	0.0136	0.9854
(4,5)	0.0050	0.9904
(1)	0.0030	0.9934
(3)	0.0030	0.9964
(1,3,5)	0.0010	0.9974
(3,9)	0.0010	0.9984
(1,3)	0.0005	0.9989
(1,9)	0.0005	0.9994
(9)	0.0005	0.9999

Table: $n = 2000$, 95% CS: {(1,2,3,4,5), (1,3,4,5)}

Binomial, $p = 100$, $s = 5$, $n = 2000$

Value	Proportion	Cumulative
(1,2,3,4,5)	0.5786	0.5786
(1,2,3,4,5,29)	0.1571	0.7357
()	0.1016	0.8373
(1,2,3,4,5,28)	0.0861	0.9234
(1,2,3,4,5,29,61)	0.0311	0.9545
(1,3,4,5)	0.0205	0.9750
(1,4)	0.0061	0.9811
(1,2,3,4,5,61)	0.0039	0.9850
(5)	0.0039	0.9889
(1,2,3,4,5,7)	0.0028	0.9917
(1)	0.0028	0.9945
(4)	0.0022	0.9967
(1,3,4)	0.0017	0.9984
(1,3)	0.0011	0.9995
(1,3,5)	0.0006	1.0001

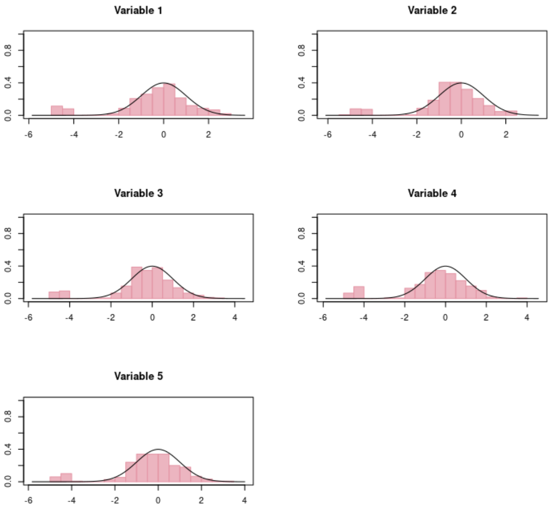
Table: $n = 2000$, 95% CS: {(1,2,3,4,5), (1,2,3,4,5,29), (), (1,2,3,4,5,28), (1,2,3,4,5,29,61)}

Binomial, $p = 40$, $s = 25$, $n = 2000$

Value	Proportion	Cumulative
(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25)	0.2988	0.2988
()	0.1603	0.4591
(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,20,21,22,23,24,25)	0.1046	0.5637
(2,4,14,22)	0.0229	0.5866
(2,4,5,13,14,22)	0.0193	0.6059
(3,14)	0.0187	0.6246
(1,2,4,5,7,8,9,10,11,12,13,14,15,16,20,21,22,23,24,25)	0.0167	0.6413
(3,4,14)	0.0146	0.6559
(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,23,24,25)	0.0141	0.67
(1,2,4,5,6,7,8,9,10,11,12,13,14,15,16,20,21,22,23,24,25)	0.0125	0.6825
⋮	⋮	⋮

Table: $n = 2000$, 95% CS: { (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25), (), (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,20,21,22,23,24,25), (2,4,14,22), (2,4,5,13,14,22), (3,14), (1,2,4,5,7,8,9,10,11,12,13,14,15,16,20,21,22,23,24,25), (3,4,14), (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,23,24,25), (1,2,4,5,6,7,8,9,10,11,12,13,14,15,16,20,21,22,23,24,25), (1,2,4,6,7,8,9,10,11,12,13,14,15,16,17,19,20,21,22,23,24,25), (2,3,4,5,7,9,11,12,13,14,18,22,24,25), (1,2,3,4,5,6,7,8,10,11,12,13,14,15,17,18,20,21,22,23,24,25), (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,20,21,22,23,24,25), (1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,17,18,20,21,22,23,24,25), (2,3,14), (3), (1,2,4,7,10,11,12,13,14,15,16,19,20,21,22,23,24,25), (2,14), (3,12,14), (2,4,14,19,22), (1,2,3,4,5,6,7,8,10,11,12,13,14,15,16,17,18,20,21,22,23,24,25), (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,24,25), (2,4,14,20,22), (3,9,12,14), (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,21,22,23,24,25), (1,2,4,5,7,10,11,12,13,14,15,16,19,20,21,22,23,24,25), (1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,20,21,22,23,24,25)

Oracle property: nsim $= 300$, Binomial, $n = 200$, $p = 10$, $s = 5$, $B = 1$



Questions

- Is there a more efficient way to re-estimate? Using ML is desirable but is that necessary?
- Do you even have to re-estimate? Or could you rerun the SGD algorithm?

References

Chee, J., H. Kim, and P. Toulis (2023, 25–27 Apr). "plus/minus the learning rate": Easy and scalable statistical inference with sgd. In F. Ruiz, J. Dy, and J.-W. van de Meent (Eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, Volume 206 of *Proceedings of Machine Learning Research*, pp. 2285–2309. PMLR.

Derryberry, D., K. Aho, J. Edwards, and T. Peterson (2018, July). Model selection and regression t-statistics. *The American Statistician 72*(4), 379–381.

Smith, J. W., J. E. Everhart, W. Dickson, W. C. Knowler, and R. S. Johannes (1988). Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, pp. 261. American Medical Informatics Association.