Sankhyā A : The Indian Journal of Statistics
https://doi.org/10.1007/s13171-024-00349-7

Post-Model-Selection Prediction Intervals for Generalized Linear Models

Dean Dustin
Charles Schwab, Lone Tree, USA
Bertrand Clarke
U. Nebraska-Lincoln, Lincoln, USA

Abstract

We give two prediction intervals for Generalized Linear Models that take model selection uncertainty into account. The first is a straightforward extension of asymptotic normality results and the second includes an extra optimization that improves nominal coverage for small-to-moderate samples. Both PI's are wider than would be obtained without incorporating model selection uncertainty. We compare these two PI's with three other PI's. Two are based on bootstrapping procedures and the third is based on a PI from Bayes model averaging. We argue that for general usage the optimized asymptotic normality PI's work best unless sample sizes are large in which case the PI's based only on asymptotic arguments that include model selection will be easier and equivalent. In an Appendix we extend our results to Generalized Linear Mixed Models.

AMS (2000) subject classification. Primary 62J12; Secondary 62M20.
Keywords and phrases. Prediction interval, generalized linear model, post-model selection.

1 Introduction

Linear models and their extensions - generalized linear, linear mixed, and generalized linear mixed models - are the workhorses of statistical analysis. Aside from formulating such models, analysts have to chose amongst competing models usually in the same class. Typically, model selection is from a model list (often chosen pre-experimentally) and done after the data is collected. There are numerous model selection procedures, but regardless of which model is chosen, it will have variability inherited from the data. The main topic of this paper is how to take this variability due to model selection into account properly when making predictions.

Common practice in many predictive contexts is to choose a model and then use it to generate predictions. Such plug-in methods are pragmatic but

Published online: 06 April 2024

neglect the uncertainty dure to model selection or, pehaps more commonly, variable selection. Here we propose prediction intervals (PI's) for generalized linear models (GLM's) that are modified by the model selection principle (MSP) used for variable selection so that their nominal coverage is asymptotically correct at least in the limit of large sample sizes. This is important because Hong et al. (2018a) showed that using model selection procedure procedures can result in predictive intervals with lower than nominal coverage if the PI's do not take the uncertainty of the MSP into account.

The post-model-selection inference problem has gained wide interest in recent years. Notably, the post-selection inference ('PoSI') intervals introduced in Berk et al. (2013) are universally valid for any model selection principle (MSP). However, PoSI intervals are known to be conservative (see Leeb et al. (2015)) partially because they allow for any ad-hoc MSP to be used. The PoSI framework was used to construct universally valid (over all MSP's) confidence intervals for the mean of a predictive distribution in LM's in Bachoc et al. (2019). Universally valid confidence regions for the simultaneous inference problem are constructed (Kuchibhotla et al., 2020).

A different approach was proposed by Efron (2014) who used bootstrap intervals to address the post-model-selection inference problem under a single MSP in somewhat the same spirit as we do here. Earlier, Stine (1985) introduced bootstrapped predictive intervals in linear regression, but these intervals did not consider uncertainty due to model selection. Leeb (2009) introduced a model selection procedure based on cross-validation techniques and proved that using this technique, the resulting prediction interval from the selected model is approximately valid. While this is a seemingly strong and useful result, it holds only in the high sparsity case – low enough parameter dimension in the limit of large n. More recently, predictive intervals based on Shorth – i.e. the shortest interval containing a pre-specified number of values – for GLM's and GAMs are studied in Stine (2021). While these intervals are valid, and account for uncertainty due to the MSP, they are not as intuitive and general as the ones we propose in Sec. 3.

Our methodology is in contrast to the PoSI-based intervals from Berk et al. (2013) that essentially widen PI's until the nominal coverage is achieved. Indeed, the PoSI intervals take the pessimistic (if practical) view that model developers will use MSP's that are not theoretically sound. Our approach is optimistic in that we assume a proper MSP with well known theoretical properties will be used. This allows us to incorporate the variability from a given generic MSP into our PI's.

Here, we present two PI's in the context of GLM's. They account for the uncertainty of an MSP in an intuitive manner. They are easy to understand and, importantly, easy to implement. One is based on an asymptotic argument that includes variability due to model selection. The other also includes a finite sample optimization to improve predictive coverage in small and moderate sample cases. It is this latter PI that we advocate for general use.

The structure of this paper is as follows. In Sec. 2 we define the notation and setting needed for our approach. In Sec. 3, we present the main theorem that gives a PI's dependent on an MSP for the case of GLM's. We then give a finite sample improvement for use with this PI in small-to-moderate sample settings. We also define three other intervals, two based on bootstrapping and one from a Bayes model average. In Sec. 4 we present simulation results to argue that our optimized asymptotic PI is the best of the five we compared. Finally, in Sec. 5 we summarize the implications of our work. In the Appendices, we extend our methodology to GLMM's.

2 Notation and Setting

Throughout this paper we assume model selection and variable selection are synonymous and exclude parameter estimation. Let $\mathcal{D}_n = \{(y_1, x_1), \ldots, (y_n, x_n)\}$ where $Y_i = y_i$ is an outcome of the response variable and x_i is a value of the d-dimensional explanatory variable. We use superscripts to indicate vectors, thus $y^n = (y_1, \ldots, y_n)^T$. Let $m \in \mathcal{M}$ be a candidate model, i.e., selection of variables, in the full collection of models $\mathcal{M}$. We define a variable selection procedure $M = M(\mathcal{D}_n)$ which takes the available data and maps it to a subset of variables based on some objective function we denote Q. We denote a chosen set of variables by $\hat{m} = \arg\min_m Q(m, \mathcal{D}_n)$, i.e., by the model that they form.

We think of Q as an objective function such as the Akaike or Bayes information criterion (AIC, BIC) or as a penalized loss function. For instance, for linear models, if Q is the AIC, we have

$$Q_{AIC}(m, \mathcal{D}_n) = -2\ln(p(y^n | X_m^n, \hat{\beta}_m^{MLE})) + 2d, \tag{2.1}$$

where $\hat{\beta}_m^{MLE}$ is the maximum likelihood estimator (MLE) in model $m \in \mathcal{M} <$ and d is the number of parameters in m that must be estimated. Now, Q defines a function

$$M : \mathcal{D}_n \mapsto \mathcal{M}$$

from the data $\mathcal{D}_n$ into the model space. We think of the variable selection procedure as an estimation problem where the function $M : \mathbb{R}^n \times \mathbb{R}^{n\times d} \mapsto \mathcal{M}$ is the estimator and is conceptually disjoint from parameter estimation.

Other choices for Q include the BIC given by

$$Q_{BIC}(m, \mathcal{D}_n) = 2\ln(p(y^n|X_m^n, \hat{\beta}_m^{MLE})) + d\log(n), \tag{2.2}$$

and the general Bethel-Shumway class of information criteria defined in Bachoc et al. (1988). Unlike AIC, the BIC and the Bethel-Shumway class of information criteria are consistent for model/variable selection.

2.1 Variable Selection In the predictive context, the interpretation of a parameter in a linear model is consistent across models: If the parameter, say β_j, appears in multiple models it always means the expected change in Y for a unit change in x_j holding other explanatory variables constant. With this in mind, we re-express an MSP M based on Q as follows. Let $X = (x_1, \ldots, x_d)$, and write $M = \{\hat{\delta}_1, \ldots, \hat{\delta}_d\}$ where for $j = 1, \ldots, d$

$$\hat{\delta}_j = \hat{\delta}_j(\mathcal{D}_n) = \begin{cases} 1 \text{ if } X_j \text{ is selected under Q} \\ 0 \text{ otherwise.} \end{cases}$$

For the true model we have $m_T = \{\delta_{1,T}, \ldots, \delta_{d,T}\}$ where

$$\delta_{j,T} = \begin{cases} 1 \text{ if } X_j \in m_T \\ 0 \text{ otherwise.} \end{cases}$$

Define the set $\mathcal{M}_S$ to be the set containing all 2^d possible vectors of zeros and ones M can take so we can regard $\mathcal{M} \subseteq \mathcal{M}_S$. Assuming $m_T \in \mathcal{M}_S$, we want to the estimate the true value $\beta_T = (\beta_1^{\delta_{1,T}}, \ldots, \beta_d^{\delta_{d,T}})$ (where a superscript of zero means that parameter (and its variable) falls out of the model) by

$$\hat{\beta}_M = (\hat{\beta}_1^{\hat{\delta}_1}, \ldots, \hat{\beta}_d^{\hat{\delta}_d}),$$

where we have ensured $\dim(\hat{\beta}_M) = \dim(\beta_{m_T})$. If $\hat{\delta}_j = 0$, then by default we set $\hat{\beta}_{\hat{\delta}_j} = \beta_j = 0$ and we have that $\delta_{j,T} = 0$ is equivalent to $\beta_j = 0$. Thus, we assume we can estimate all parameters in any chosen model.

For clarity, we define the consistency of variable selection under M and henceforth assume it. We say that M is a consistent MSP if and only if, with probability under the true model m_T going to one, M selects the variables

in m_T asymptotically correctly. More formally, we say that $M(\mathcal{D}_n)$ selects the model with variables $x_1^{\delta_1}, \ldots, x_d^{\delta_d}$. So, M is consistent if and only if all

$$\hat{\delta}_j(M(\mathcal{D}_n)) \to \delta_j(m_T) \tag{2.3}$$

in the probability defined by m_T; Eq. 2.3 assumes m_T is not mis-specified.

We specify our target of inference as $\beta_T = \beta_{m_T}$ and note

$$\beta_{m_T} = \left(X'_{m_T} X_{m_T}\right)^{-1} X'_{m_T} E(Y)$$

in the linear models context where the subscript m_T indicates which explanatory variables are in the design matrix. This is in contrast to the random target of inference

$$\beta_M = \left(X'_M X_M\right)^{-1} X'_M E(Y)$$

defined in Berk et al. (2013). Thus, for linear models, we define the estimate $\hat{M}$ for β_j under M using $\mathcal{D}_n$ by

$$\hat{\beta}_{\hat{M},j} = \begin{cases} \left[\left(X'_{\hat{M}} X_{\hat{M}}\right)^{-1} X'_{\hat{M}} y\right]_j & \text{if } \hat{\delta}_j = 1 \\ 0 & \text{if } \hat{\delta}_j = 0. \end{cases}$$

Now there are two steps in the process of obtaining the true model. The first step is to estimate the δ_j's. In this step we would like M to give $\hat{\delta}_j = 1$ if $\delta_{j,T} = 1$. However, M may also give $\hat{\delta}_j = 1$ even if $\delta_{j,T} = 0$. In this case, our definition allows the estimate $\hat{\beta}_{\delta_j}$ to be zero. Thus, even if M includes variables that are not in M_T we can still estimate their coefficients to be zero which allows $M \to m_T$ asymptotically (as seen in Theorem 3.1).

2.2 Prediction in Generalized Linear Models As noted, we restrict attention to GLM's here and discuss GLMM's in the appendices. To be more precise, suppose $Y \sim \mathcal{G}(\mu, R)$ where $\mathcal{G}$ is an exponential family with mean μ and variance R. Then the pdf of Y given the canonical parameter θ is

$$f(y|\theta) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)} \tag{2.4}$$

where ϕ is a scale parameter. In one parameter exponential families such as Poisson or Binomial distributions, $a(\phi) = 1$. From Eq. 2.4 we have the following:

- $E(Y|X) = \frac{\partial b(\theta)}{\partial \theta} = \mu$,

- $Var(Y|X) = a(\phi)\frac{\partial^2 b(\theta)}{\partial \theta^2} = a(\phi)V(\mu)$,
- $I(\theta) = Var(\ell(\theta|y,\phi))$, where ℓ is the log-likelihood, see Eq. 2.7.

Following standard GLM practice, we model the mean of Y by transforming it to a linear function of the explanatory variables. The function we use to transform $E(Y) = \mu$ is called the link function and we denote it by $g(\cdot)$. Note that $g(\cdot)$ is a continuous invertible function. This gives us the linear predictor

$$\eta = g(E(Y|X)) = g(\mu) = X\beta \tag{2.5}$$

and we define the inverse link function to be the inverse of $g(\cdot)$ which is

$$\mu = E(Y|X) = g^{-1}(X\beta). \tag{2.6}$$

Here, we assume that X is of full rank, to avoid problems with estimability.

Note that the canonical parameter θ is a function of μ so we write $\theta = \theta(\mu) = \theta(g^{-1}(X\beta))$. Now we can write the log-likelihood of Eq. 2.4 as

$$\ell(\beta|y,\phi) = \frac{y\left(\theta(g^{-1}(X\beta))\right) - b(\theta(g^{-1}(X\beta)))}{a(\phi)} + c(y,\phi). \tag{2.7}$$

For ease of exposition, we start by assuming ϕ is known and use the maximum likelihood estimators (MLE's) for β, denoted generically by $\hat{\beta}$. Although $\hat{\beta}$ does not have a closed form expression outside the normal case, the Newton-Raphson algorithm or Fisher scoring can be used to find it. Conditions for the consistency and asymptotic normality for the MLE are well-known (see Theorems 2 and 3 in the classic paper Fahrmeir and Kaufmann (1985)) and henceforth assumed. We comment on estimating ϕ after proving Theorem 3.1.

Suppose the inferential goal is predicting the next outcome Y^{n+1}. The usual point predictor under a consistent MSP M is

$$\hat{Y}_M^{n+1} = \hat{\mu}_M = g^{-1}(X_M'^{n+1}\hat{\beta}_M), \tag{2.8}$$

where $\hat{\beta}_M$ is the MLE for β whether found in closed form or by some iterative algorithm that generates consistent and asymptotically normal estimates. Henceforth, our focus in on constructing valid PI's for Eq. 2.8. Note that $n+1$ just denotes a generic outcome we don't yet have. It need not be the first outcome after the n data point.

3 Candidate PI's

In this section we define four PI's. The first is derived in Theorem 3.1 below. The second is an improvement on this PI by incorporating an extra optimization to ensure more rapid convergence to the nominal coverage for small sample sizes. Both of these are in Sec. 3.1. In Sec. 3.2 we give our third and fourth intervals are based on bootstrapping approach. (We define a fifth in Sec. 4.) We will argue that our optimized interval provides the best performance.

3.1 Main Result and Two PI's One choice for a PI uses asymptotic normality of the point predictor Eq. (2.8). Define the statistic

$$Z_{pred} = Z_{pred}(M) = \frac{\hat{Y}_M^{n+1} - Y^{n+1}}{\sqrt{Var(\hat{Y}_M^{n+1} - Y^{n+1})}}. \tag{3.1}$$

It is the predictor for a future outcome that includes variable selection in its numerator and hence has a larger denominator to account for the extra variability. We have the following result giving our first PI. When we need to indicate this interval in contrast to others we use a superscript AN.

Theorem 3.1 *Suppose* Y^n, Y^{n+1} *come from an exponential family distribution as defined in* Sec. 2.2.[1] *and let* M *be a consistent MSP. Suppose also that the MLE is consistent and asymptotically normal with variance decreasing as* $\mathcal{O}(1/n)$. *An asymptotically* $(1-\alpha)$ *coverage PI for a new outcome is* $PI(M) = PI_\alpha(M)$ *given by*

$$g^{-1}(X_M'^{n+1}\hat{\beta}_M) \pm z_{1-\alpha/2}\sqrt{\left[\frac{d}{d\eta}g^{-1}(\hat{\eta}_M^{n+1})\right]^2 X_M'^{n+1}Var(\hat{\beta}_M)X_M^{n+1} + a(\phi)V(\hat{\mu})_M}. \tag{3.2}$$

Remark: Any consistent and asymptotically normal estimator can be used in place of the MLE if the corresponding modifications are made to the proof.

Proof The key quantity we want to control asymptotically is

$$\sqrt{n}(X_M'^{n+1}\hat{\beta}_M - X_{m_T}'^{n+1}\beta_{m_T}). \tag{3.3}$$

To do so, define the 'good' set

$$S_n = \{\omega | \forall j, \hat{\delta}_j(\omega) = \delta_{j,T}\}$$

and let $\mathbf{1}_{S_n}$ be the indicator that $\omega \in S_n$. Note that for ease of exposition we have expressed the random variables $\hat{\delta}$ in terms of the underlying measure

[1] See Eq. 2.4 and the conditions following it.

space. Letting $\mathbf{1}_{S_n^c}$ be the indicator that ω is in the complement of S_n we can write

$$\sqrt{n}(X_M'^{n+1}\hat{\beta}_M - X_{m_T}'^{n+1}\beta_{m_T}) = \sqrt{n}(X_M'^{n+1}\hat{\beta}_M - X_{m_T}'^{n+1}\beta_{m_T})\mathbf{1}_{S_n} + \sqrt{n}(X_M'^{n+1}\hat{\beta}_M - X_{m_T}'^{n+1}\beta_{m_T})\mathbf{1}_{S_n^c}. \quad (3.4)$$

There are now three technical steps. First, we show that $\sqrt{n}\mathbf{1}_{S_n^c} = o_p(1)$. The union of events bound gives

$$P(S_n^c) \leq \sum_{j=1}^{p} P(|\hat{\delta}_j - \delta_{j,T}| > \eta),$$

for some $\eta > 0$, so using symmetry in the MSP it is enough to show

$$\lim_{n\to\infty} P(|\hat{\delta}_j - \delta_{j,T}| > 1/\sqrt{n}) = 0$$

for any j. It is easy to see that

$$\lim_{n\to\infty} P(|\hat{\delta}_j - \delta_{j,T}| > 1/\sqrt{n}) = \lim_{n\to\infty} P(|\hat{\delta}_j - \delta_{j,T}| = 1) \quad (3.5)$$

because $\delta_{j,T}$ and $\hat{\delta}_j$ are either 1 or 0. Now because we have chosen a consistent MSP we have

$$\lim_{n\to\infty} P(|\hat{\delta}_j - \delta_{j,T}| = 1) = 0.$$

Hence, the left hand side of Eq. 3.5 is also equal to zero and we get $\sqrt{n}\mathbf{1}_{S_n^c} = o_p(1)$.

Second, dealing with the 'bad' set first, observe the second term on the RHS of Eq. 3.4 is

$$\begin{aligned}
&\sqrt{n}(X_M'^{n+1}\hat{\beta}_M - X_M'^{n+1}\beta_{m_T} + X_M'^{n+1}\beta_{m_T} - X_{m_T}'^{n+1}\beta_{m_T})\mathbf{1}_{S_n^c} \\
&= \sqrt{n}(X_M'^{n+1}\hat{\beta}_M - X_M'^{n+1}\beta_{m_T})\mathbf{1}_{S_n^c} + \sqrt{n}(X_M'^{n+1}\beta_{m_T} - X_{m_T}'^{n+1}\beta_{m_T})\mathbf{1}_{S_n^c} \\
&= X_M'^{n+1}\sqrt{n}(\hat{\beta}_M - \beta_{m_T})\mathbf{1}_{S_n^c} + \sqrt{n}(X_M'^{n+1} - X_{m_T}'^{n+1})\beta_{m_T}\mathbf{1}_{S_n^c}. \quad (3.6)
\end{aligned}$$

In both terms of Eq. 3.6 we use $\sqrt{n}\mathbf{1}_{S_n^c} = o_p(1)$ so continuing the equality gives

$$X_M'^{n+1}(\hat{\beta}_M - \beta_{m_T})o_P(1) + (X_M'^{n+1} - X_{m_T}'^{n+1})\beta_{m_T}o_P(1). \quad (3.7)$$

It is easy to see that $X_M'^{n+1}$ is a real vector and that consistency of $\hat{\beta}_M$ gives that $(\hat{\beta}_M - \beta_{m_T}) = \mathcal{O}_P(1)$ so Slutsky's theorem gives the first term in Eq. 3.7 is $o_P(1)$. Similarly, β_{m_T} is a constant and $(X_M'^{n+1} - X_{m_T}'^{n+1}) = \mathcal{O}(1)$ so Slutsky's theorem gives the second term in Eq. 3.7 also goes to zero in probability. Thus, the second term on the RHS of Eq. 3.4 goes to zero in probability.

Third, it remains to deal with the first term on the right of Eq. 3.4 involving the good set. Note that on the good set $M = m_T$. So,

$$\begin{aligned}\sqrt{n}(X_M'^{n+1}\hat{\beta}_M - X_{m_T}'^{n+1}\beta_{m_T})\mathbf{1}_{S_n} &= \sqrt{n}(X_{m_T}'^{n+1}\hat{\beta}_{m_T} - X_{m_T}'^{n+1}\beta_{m_T})\mathbf{1}_{S_n} \\ &= X_{m_T}'^{n+1}\left[\sqrt{n}(\hat{\beta}_{m_T} - \beta_{m_T})\right]\mathbf{1}_{S_n}. \quad (3.8)\end{aligned}$$

Again, we use a Slutsky's theorem argument. The first factor $X_{m_T}'^{n+1}$ is bounded. The factor in square brackets is asymptotically normal because we are using the MLE. And the last factor $\mathbf{1}_{S_n} \to 1$ in probability.

Thus, putting all thesse pieces together we get

$$\sqrt{n}(X_M'^{n+1}\hat{\beta}_M - X_{m_T}'^{n+1}\beta_{m_T}) \xrightarrow{D} N\left(0, X_{m_T}'^{n+1}V_{m_T}^* X_{m_T}^{n+1}\right).$$

To get a predictive distribution, we observe that using the delta method on the link function gives

$$\sqrt{n}\left(g^{-1}(X_M'^{n+1}\hat{\beta}_M) - g^{-1}(X_{m_T}'^{n+1}\beta_{m_T})\right) \xrightarrow{D} N\left(0, \left[\frac{d}{d\eta}g^{-1}(\eta_{m_T}^{n+1})\right]^2 X_{m_T}'^{n+1}V_{m_T}^* X_{m_T}^{n+1}\right) \quad (3.9)$$

where $\eta_{m_T}^{n+1} = X_{m_T}'^{n+1}\beta_{m_T}$.

Since $\hat{Y}_M^{n+1} = \hat{Y}_M(X^{n+1}) = g^{-1}(X_M'^{n+1}\hat{\beta}_M)$ and $Y^{n+1} = Y(X^{n+1}) = g^{-1}(X_{m_T}'^{n+1}\beta_{m_T})$, the variance of $\hat{Y}_M^{n+1} - Y^{n+1}$ is

$$\begin{aligned}Var(\hat{Y}_M^{n+1} - Y^{n+1}) &= Var(\hat{Y}_M^{n+1}) + Var(Y^{n+1}) \\ &= Var(g^{-1}(X_M'^{n+1}\hat{\beta}_M)) + a(\phi)V(\mu) \\ &\approx \frac{1}{n}\left[\frac{d}{d\eta}g^{-1}(\eta_{m_T}^{n+1})\right]^2 X_{m_T}'^{n+1}V_{m_T}^* X_{m_T}^{n+1} + a(\phi)V(\hat{\mu}) \quad (3.10)\end{aligned}$$

due to Eq. 3.9. Again, because Y^{n+1} is a random variable, and not a parameter, we must consider the variance of it as well, which we get assuming it

will come from the exponential family distribution as $Y_1, \ldots, Y_n$. This quantity, however, is impossible to compute because we do not know m_T. Hence, we must replace m_T with M, making the variance a random quantity that depends on model selection. Thus, we must replace μ with $\hat{\mu}$ in Eq. 3.10.

Now we use Eq. 3.1 as a pivotal quantity to get

$$\begin{aligned} 1-\alpha &\leq P\left(|Z_{pred}| < z_{1-\alpha/2}\right) \\ &= P\left(\left|\hat{Y}^{n+1} - Y^{n+1}\right| < z_{1-\alpha/2}\sqrt{Var(\hat{Y}^{n+1} - Y^{n+1})}\right) \\ &= P\left(\hat{Y}^{n+1} - z_{1-\alpha/2}\sqrt{Var(\hat{Y}^{n+1} - Y^{n+1})} < Y_{n+1} < \hat{Y}^{n+1} + z_{1-\alpha/2}\sqrt{Var(\hat{Y}^{n+1} - Y^{n+1})}\right). \end{aligned} \tag{3.11}$$

Hence using Eq. 3.10

$$\left[g^{-1}(X_M'^{n+1}\hat{\beta}_M) \pm z_{1-\alpha/2}\sqrt{\frac{1}{n}\left[\frac{d}{d\eta}g^{-1}(\hat{\eta}_M^{n+1})\right]^2 X_M'^{n+1}V_M^* X_M^{n+1} + a(\phi)V(\hat{\mu}_M)}\right]$$

is the $100(1-\alpha)\%$ prediction interval we call PI^{AN} for Y^{n+1} □

In the general case, we may not know ϕ and hence we need to estimate it. There are three main approaches to obtain the estimate $\hat{\phi}$. The approaches are the deviance approach, the Pearson approach and the Fletcher approach. The deviance approach is to estimate ϕ using

$$\hat{\phi}_D = \frac{\hat{D}}{n-d}$$

where $\hat{D} = \frac{2\phi}{a(\phi)}\sum_{i=1}^n (Y_i(g(Y_i) - \hat{\theta}_i) - b(g(Y_i)) + b(\hat{\theta}_i))$. Note here the $a(\phi)$ is proportional to ϕ so $\hat{D}$ does not depend on ϕ.

The Pearson approach uses

$$\hat{\phi}_P = \frac{\sum_{i=1}^n (y_i - \hat{\mu}_i)^2}{(n-d)V(\hat{\mu}_i)}.$$

We use the Pearson approach in our simulated example in Sec. 4.1 because in the Gaussian setting, it is an unbiased estimator of ϕ. If there is overdispersion in the data, the approach in Fletcher (2012) approach may be better than both the deviance and Pearson approach.

In the non-Gaussian case the estimates of the dispersion parameter are often biased and can lead to the width of our PI being larger or smaller than required depending on the direction of the bias. This issue may be alleviated

by using the PI we describe next because it offers a data driven adjustment to the width of the interval to optimize the coverage.

We end this subsection with a small sample improvement on the PI from Theorem 3.1. Note that Eq. 3.2 uses the standard normal quantile to define the predictive interval. However, we can adjust the width of the interval to correct for poor coverage. To do this, let $PI(M, C)$ denote the interval in Eq. 3.2 with the normal quantile replaced by C and consider the interval

$$PI(C_{\alpha,M}) = g^{-1}(X_M'^{n+1}\hat{\beta}_M) \pm C_{\alpha,M} \times \sqrt{\left[\frac{d}{d\eta}g^{-1}(\hat{\eta}_M^{n+1})\right]^2 X_M'^{n+1} Var(\hat{\beta}_M) X_M^{n+1} + a(\phi)V(\hat{\mu})} \tag{3.12}$$

where $C_{\alpha,M}$ is chosen to satisfy

$$C_{\alpha,M} = \arg\min_C P\left(Y^{n+1} \in PI(M, C)\right) \tag{3.13}$$

for all C such that $P\left(Y^{n+1} \in PI(M, C)\right) \geq 1 - \alpha$. Importantly, this probability also sees the random variable M and hence inherits the uncertainty associated with M as well as Y^{n+1}. This is in the same spirit of the PoSI constant in Berk et al. (2013). That is, we enlarge $C_{\alpha,M}$ to account for the uncertainty in M. We can approximate $C_{\alpha,M}$ using Monte Carlo cross validation.

We begin by choosing an interval on $\mathbb{R}^+$ denoted $\mathcal{C} = [C_{low}, C_{high}]$ that we will perform the line search on to estimate $C_{\alpha,M}$. Next, we randomly split $\mathcal{D}_n$ into $L \in \mathbb{N}$ test and train sets, $\mathcal{D}_{train,\ell}$ and $\mathcal{D}_{test,\ell}$ for $\ell = 1, \ldots, L$. Then, in the Gaussian case, for each ℓ we estimate β by

$$\hat{\beta}_\ell = (X'_{train,M,\ell} X_{train,M,\ell})^{-1} X'_{train,M,\ell} y_{train,\ell}$$

using $\mathcal{D}_{train,m}$, and form the predictor $\hat{Y}_{M,\ell}^{test} = g^{-1}(X_{test,M,\ell}\hat{\beta}_\ell)$. In the non-Gaussian case we estimate $\hat{\beta}_\ell$ with the MLE using the Newton-Rapshon or Fisher scoring algorithms. Now form the prediction interval $PI_{M,\ell}(C)$

$$\hat{Y}_{M,\ell}^{n+1} \pm C\sqrt{\left[\frac{d}{d\eta}g^{-1}(\hat{\eta}_{test,M,\ell}^{n+1})\right]^2 X_{test,M,\ell}'^{n+1} Var(\hat{\beta})_{test,M,\ell} X_{test,M,\ell}^{n+1} + a(\phi)V(\hat{\mu})_{M,\ell}} \tag{3.14}$$

and for each $C \in \mathcal{C}$, check if $y_{test,\ell} \in PI_{M,\ell}(C)$. Then we choose the value C

that gives us $1-\alpha$ coverage for the Monte Carlo samples. More formally, we can approximate $C_{\alpha,M}$ by

$$\hat{C}^{MC} = \arg\min_{C} \frac{1}{L} \sum_{\ell=1}^{L} \left| \frac{1}{\#(\mathcal{D}_{test,\ell})} \sum_{i=1}^{\#(\mathcal{D}_{test,\ell})} I_{y_{test,\ell} \in PI_{M,\ell}(C)} - (1-\alpha) \right| \quad (3.15)$$

where $I_{y_{test,\ell} \in PI_{M,\ell}(C)}$ is the indicator that the test values are in the constructed intervals.

Using $\hat{C}^{MC}$, we define the follow PI

$$PI(M)^{\hat{C}^{MC}} = \hat{Y}_M^{n+1} \pm \hat{C}^{MC} \sqrt{\left[\frac{d}{d\eta} g^{-1}(\hat{\eta}_M^{n+1})\right]^2 X_M'^{n+1} Var(\hat{\beta}_M) X_M^{n+1} + a(\phi) V(\hat{\mu})}. \quad (3.16)$$

The intuition behind using this interval in place of the PI in Theorem 3.1 is that, in finite samples, the difference between $z_{1-\alpha/2}$ and $\hat{C}^{MC}$ can be interpreted as the added variability due to model uncertainty.

3.2 Two Bootstrap Based PI's In the frequentist setting, perhaps the most natural way to obtain a PI that takes into account both the uncertainty of model selection and the uncertainty associated with the distribution of the new outcome is to make use of the bootstrap. Accordingly, to form our first bootstrapped PI, we use the bootstrap to estimate the distribution of

$$\hat{\mu}_M = E(Y^{n+1}|X_M^{n+1}) = g^{-1}(X_M^{n+1}\hat{\beta}_M) \quad (3.17)$$

and we bootstrap to estimate the distribution of $\hat{\phi}_M$ when necessary. This gives bootstrap samples $\hat{\mu}_{M,b}$ and $\hat{\phi}_{M,b}$ for $b = 1, \ldots, B$. Then for each of the bootstrapped mean $\hat{\mu}_{M,b}$ and dispersion parameter $\hat{\phi}_{M,b}$, we generate a new observation from the distribution of $Y^{n+1}|X^{n+1}, \mu, \phi$, i.e. $\mathcal{G}$.

Let $\hat{p}(\hat{\mu})$ be the bootstrapped density of Eq. 3.17, $\hat{p}(\phi)$ be the bootstrapped density of ϕ_M, and $\hat{p}(Y^{n+1})$ be the resulting estimated density of Y^{n+1}.

The procedure is as follows. For $b = 1, \ldots, B$,

- generate $\hat{\mu}_{M,b} = g^{-1}(X_{M,b}^{n+1}\hat{\beta}_{M,b})$ using the known link function and the bootstrap sample, denoted $\mu_1^*, \ldots, \mu_B^*$
- if there is a dispersion parameter to estimate, generate B bootstrap replications of ϕ_M, denoted $\phi_1^*, \ldots, \phi_B^*$
- generate observations $y_1^* \mid \mu_1^*, \phi_1^*, \ldots, y_B^* \mid \mu_B^*, \phi_B^*$, by randomly generating a new observation from $\mathcal{G}$.

The sample $y_1^*, \ldots, y_B^*$ can be used to estimate an approximate marginal predictive distribution for Y^{n+1}. To obtain the PI, we use the appropriate percentile interval from this distribution. That is, to obtain a $100(1-\alpha)\%$ PI we use the interval

$$PI^{boot} = [q^*_{1-\alpha/2}, q^*_{\alpha/2}] \tag{3.18}$$

where q^*_α is the α quantile from $\hat{p}(Y^{n+1})$. The use of $\hat{p}(\hat{\mu})$ and, if needed, $\hat{p}(\phi)$ to obtain the estimated predictive distribution $\hat{p}(Y^{n+1})$ allows $\hat{p}(Y^{n+1})$ to inherit the variability from $\hat{p}(\hat{\mu})$, $\hat{p}(\phi)$, and the variability that is already associated with the known parametric distribution $\mathcal{G}$. Hence, the interval Eq. (3.18) is typically widened due to the uncertainty of the model selection procedure as well as the uncertainty of the distribution of Y^{n+1}.

In the GLM setting, coverage for the PI in Eq. 3.18 should be closer to the $1-\alpha$ nominal coverage than the PI resulting from ignoring the model uncertainty. Note that as $n \to \infty$ the variability due to model uncertainty will go to 0 and this interval will converge to the standard PI. Bootstrap PI's for the Gaussian case are studied in a fairly narrow (small d and moderate n) setting in Hong et al. (2018b). These authors suggest that in this setting the bootstrap distribution fails to assess the uncertainty of model selection accurately. We explore different simulation settings to evaluate the performance of bootstrap intervals in Sec. 4.

Our second bootstrapped PI is formed as follows. Recall the interval in Theorem 3.1 is a random because it depends on M. It is directly usable for predictions, but we must use $\hat{M}$ in place of M to get a confidence statement. Nevertheless, we provide an approximate interval by "smoothing" over M, which accounts for the uncertainty of M in both the center and width of the interval. This is similar to the approach used in Efron (2014) for estimation. The method we propose is to use $\hat{p}(\hat{\mu})$, the bootstrap the distribution of $\hat{\mu}_M = g^{-1}(\hat{\eta}_M)$ as described earlier in this Subsection, to obtain an approximation for the predictor and its variance that accounts for model selection uncertainty.

Specifically, we use

$$\tilde{\mu} = \frac{1}{B}\sum_{b=1}^{B} \mu_b^*$$

for the point predictor. We approximate the variance of $\hat{\mu}_M$ with

$$\begin{aligned} Var(\mu^*) &= \widehat{Var}(\hat{\mu}_M) \\ &= \frac{1}{B-1}\sum_{b=1}^{B}(\mu_b^* - \tilde{\mu})^2 \\ &\approx \left[\frac{d}{d\eta}g^{-1}(\hat{\eta}_M^{n+1})\right]^2 X_M'^{n+1} Var(\hat{\beta}_M) X_M^{n+1}, \end{aligned}$$

and the estimated variance of the predictive distribution is given by

$$\begin{aligned} Var(Y^*) &= \widehat{Var}(Y^{n+1}) \\ &= \frac{1}{B-1}\sum_{b=1}^{B}(y^*(\mu_b^*) - \bar{y}^*)^2 \\ &\approx a(\phi)V(\hat{\mu}_M) \end{aligned}$$

where $\bar{y}^* = \frac{1}{B-1}\sum_{b=1}^{B} y^*(\mu_b^*)$. We treat Y^* as a random variable approximating Y^{n+1}. Note also that we are required to estimate $V(\hat{\mu}_M)$ in Eq. 3.2, but this again is a random variable so we use the bootstrap to account for the uncertainty in M for this term also. Since we have assumed ϕ is known we do not need to estimate $a(\phi)$. Now as an ad-hoc fix, we rewrite Eq. 3.2 to give our second bootstrapped PI

$$PI(M)^{S\text{-}boot} = \tilde{\mu} \pm z_{1-\alpha/2}\sqrt{Var(\mu^*) + Var(Y^*))}. \tag{3.19}$$

4 Simulation Results for GLM's

We give two contexts in which the PI's we have defined in Eqs. 3.2, 3.12, 3.18, and 3.19 can be readily found. Respectively, these intervals are labeled the asymptotic normal PI (AN), the optimized AN $\hat{C}^{MC}$, the bootstrapped (boot) PI, and the 'smoothed' asymptotic normal (S-boot) PI. In addition to the intervals we have derived, we give the BMA PI's as well as the 'Naive' PI's obtained by applying the inverse link to the endpoints of a confidence interval for the mean on the linear predictor scale; practitioners often regard this as an acceptable and 'pragmatic' solution.

We include the BMA PI for comparison because it is often used to account for model uncertainty. For the full details on BMA, see Hoeting et al. (1999). For our implementation of BMA here, we used the BAS package in R, and we used Zellner-Siow priors on the model coefficients and a

uniform across model prior. In this case the "models" are referring to a collection of explanatory variables used in the GLM. The PI's we present in our results for BMA are the highest posterior density intervals from the BMA posterior predictive distribution.

The naive PI is given by

$$PI^{naive}(M) = [g^{-1}(\hat{\eta}_{M,low}), g^{-1}(\hat{\eta}_{M,high})]$$

where

$$\hat{\eta}_{M,low} = X_M'^{n+1}\hat{\beta}_M - z_{1-\alpha/2}\sqrt{X_M'^{n+1}Var(\hat{\beta}_M)X_M^{n+1}},$$

and

$$\hat{\eta}_{M,high} = X_M'^{n+1}\hat{\beta}_M + z_{1-\alpha/2}\sqrt{X_M'^{n+1}Var(\hat{\beta}_M)X_M^{n+1}}.$$

In Sec. 4.1 we present these intervals for the standard Gaussian case and in Sec. 4.2, we present the prediction intervals for binomial regression, i.e., a more general case of logistic regression. For both cases we use 500 new observations from their respective distribution and calculate the estimated predictive coverage using

$$\widehat{coverage} = \frac{1}{500}\sum_{i=1}^{500} I_{y_i^{new} \in PI_i(X_i^{new}, X^n, y^n)}, \tag{4.1}$$

where each $PI_i(X_i^{new}, X^n, y^n)$ depends on the data and the new observed explanatory variables. For the PIs that require bootstrapping we resample the data 500 times to obtain the bootstrapped distributions.

4.1 Gaussian Linear Models In the standard case, we assume $Y \sim N(\mu, \sigma^2)$, and the log likelihood is

$$L(\mu_i, \sigma^2|y_i) = \frac{y_i\mu_i - (\mu_i^2/2)}{\sigma^2} - \left(\frac{y_i^2}{2\sigma^2} + \log(\sigma\sqrt{2\pi}\right),$$

the canonical parameter is $\theta_i = \mu_i$, $b(\theta) = \mu + i^2/2$, $a(\phi) = \sigma^2$, and $V(\mu_i) = 1$. In this case we need to estimate the dispersion parameter to estimate σ^2.

The linear predictor uses the identity link function and the point predictor is

$$\hat{Y}_M^{n+1} = X_M'^{n+1}\hat{\beta}_M.$$

The asymptotic normal PI from Eq. 3.2 for Y^{n+1} is

$$PI(M)^{AN} = \left[\hat{Y}_M^{n+1} \pm z_{1-\alpha/2}\hat{\sigma}_M\sqrt{X_M'^{n+1}(X_M'X_M)^{-1}X_M^{n+1} + 1}\right].$$

Our simulation results for Gaussian data includes coverage and width estimates for the normal PI in Eq. 3.2, the PI Eq. (3.14) using $\hat{C}^{MC}$, the bootstrap PI in Eq. 3.18, and the 'smoothed' normal interval Eq. (3.19). We do not include the Naive interval because in the Gaussian case it is equivalent to AN. For the interval using $\hat{C}^{MC}$, we do a grid search for the value of $C_{\alpha,M}$ on the interval from 1.95 to 5 in increments of 0.05. Here we have used 1000 Monte Carlo samples, and used a 70/30 train/test split for evaluating coverage on the test set.

The simulation setup is as follows. First, we consider two model selection procedures, BIC and AIC. Both methods are implemented in R using the step() function by setting the respective penalties for BIC and AIC. We also use BMA implemented with the BAS package in R with prior specification discussed in Sec. 4. We consider various choices for n (30,50,100,200) and choose $p = 25$. We randomly generate values for σ and β once, and fix those values throughout the simulations. Accordingly, let

$$\beta = (\beta_1, \ldots, \beta_{25})' = (6.43, 4.39, 4.26, 4.11, 0, \ldots, 0)'$$

and $\sigma = 0.93$. We simulate n observations for the design matrix X according to

$$X \sim MVN_p(0, I_p),$$

and then draw and $n \times 1$ vector of observations from $Y \sim N(X\beta, \sigma^2 I_n)$. We then calculate estimated coverage using Eq. 4.1. Ideally, we want coverage close to 0.95. When choosing between competing PI's with good coverage, we prefer the one with the narrowest width. The results are seen in Table 1.

For the bootstrap based intervals we use the procedure outlined in Sec. 3.2. In this setting, the steps are as follows, for $b = 1, \ldots, B$,

- generate $\hat{\mu}_{M,b} = g^{-1}(X^{n+1}_{M,b}\hat{\beta}_{M,b}) = X^{n+1}_{M,b}\hat{\beta}_{M,b}$ using the identity link function and the bootstrap sample, denoted $\mu^*_1, \ldots, \mu^*_B$
- generate $\hat{\phi}_{M,b} = \hat{\sigma}^2_{M,b} = \frac{1}{n-d}\sum_{i=1}^{n}(y_i - \hat{\mu}_{M,b})^2$ denoted $\phi^*_1, \ldots, \phi^*_B$
- generate observations $y^*_1 \mid \mu^*_1, \phi^*_1, \ldots, y^*_B \mid \mu^*_B, \phi^*_B$, by randomly generating a new observation from $\mathcal{N}(\hat{\mu}_{M,b}, \hat{\sigma}^2_{M,b})$.

Note that the differences between using AIC and BIC are negligible, so we describe the performance of each PI only once (rather than once for each MSP). It is seen in Table 1 that AN has low coverage for $n = 50$, but gets close to the nominal coverage for the larger sample sizes. For $n = 50$, both

Table 1: Simulation results for Gaussian data with $p = 25$ and $p_0 = 4$

n	MSP	Interval	Coverage	Avg.Width (SE)
50	AIC	AN	0.88	3.63 (0.17)
		$\hat{C}^{MC}$	0.98	5.5 (0.35)
		boot	0.99	8.8 (1.87)
		S-boot	1	14.2 (3.81)
	BIC	AN	0.88	3.63 (0.18)
		$\hat{C}^{MC}$	0.98	5.37 (0.26)
		boot	0.99	8.4 (1.78)
		S-boot	1	13.79 (3.80)
	BMA		0.89	3.98 (0.11)
100	AIC	AN	0.91	3.63 (0.08)
		$\hat{C}^{MC}$	0.93	4.08 (0.09)
		boot	0.92	3.81 (0.26)
		S-boot	0.95	4.43 (0.41)
	BIC	AN	0.92	3.70 (0.06)
		$\hat{C}^{MC}$	0.94	3.97 (0.05)
		boot	0.92	3.74 (0.22)
		S-boot	0.94	4.25 (0.33)
	BMA		0.93	3.73 (0.05)
200	AIC	AN	0.92	3.70 (0.05)
		$\hat{C}^{MC}$	0.93	3.96 (0.06)
		boot	0.91	3.64 (0.14)
		S-boot	0.93	3.95 (0.16)
	BIC	AN	0.94	3.75 (0.03)
		$\hat{C}^{MC}$	0.94	3.92 (0.03)
		boot	0.92	3.65 (0.12)
		S-boot	0.94	3.91 (0.14)
	BMA		0.92	3.74 (0.03)

S-boot and boot give at least the nominal coverage and arguably reasonable width of PI's to be useful. Here, $\hat{C}^{MC}$ gives close to the stated 95% coverage and is noticeably narrower than both S-boot and boot, so it is the preferred PI.

When $n = 100$ and 200, we observe all of the 5 PIs are roughly equal in terms of coverage and width. Since AN is the easiest to implement as it does not require any bootstrapping or cross validation, we recommend using

it with relatively large n. For intermediate n we recommend using $\hat{C}^{MC}$ as it gives appropriate coverage and is narrower than the other PIs.

We give the optimal choices for $\hat{C}^{MC}$ for each sample size in Table 2. We observe that as sample size increases, $\hat{C}^{MC}$ decreases as expected. This reflects the fact that as we gather more data, the uncertainty in model selection also decreases. Overall, the PI using $\hat{C}^{MC}$ is able to capture the uncertainty due to model selection for smaller sample size such that it gives appropriate coverage. In larger sample sizes $\hat{C}^{MC}$ decreases because we have chosen a consistent MSP and there is less uncertainty in model selection when with a larger sample. Thus, we recommend using this PI to ensure the interval is wide enough to have proper coverage regardless of the sample size.

4.2 Binomial Regression Suppose we have n independent but not identically distributed random variables following $Y_i \sim Bin(r_i, p_i)$ so $E(Y_i) = r_i p_i$. We write $W = \frac{Y_i}{r_i}$ as our response to model the proportion of success, and then we convert back to number of successes to form our predictive interval. Now we have $E(W) = p_i$ and the log likelihood for a given i is given by

$$L(p_i|w_i) = \frac{w_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)}{\frac{1}{r_i}} + \log\binom{r_i}{nw_i},$$

which reveals the canonical parameter

$$\theta_i = logit(p_i) = \log\left(\frac{p_i}{1-p_i}\right).$$

We also see that $a(\phi) = \frac{1}{r_i}$, $b(\theta_i) = -\log(1+e^{\theta_i}) = -\log(1-p_i)$, and thus

$$V(p_i) = \frac{\partial^2 b(\theta_i)}{\partial p_i^2} = \frac{p_i(1-p_i)}{r_i}.$$

Table 2: Gaussian cross validation results for the optimal choice for $\hat{C}^{MC}$

n	MSP	$\hat{C}^{MC}$
50	AIC	2.95
	BIC	2.90
100	AIC	2.20
	BIC	2.10
200	AIC	2.10
	BIC	2.05

Thus the linear predictor is defined by the logit link as

$$E\left(\frac{Y_i}{r_i}\right) = g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = X_i'\beta$$

and the inverse link function, which gives the probability of success, is given by

$$p_i = g^{-1}(X_i'\beta) = \frac{1}{1+e^{-X_i'\beta}}.$$

Of course, we do not know p_i, so we estimate p_i by

$$\hat{p}_i = g^{-1}(X_i'\hat{\beta}) = \frac{1}{1+e^{-X_i'\hat{\beta}}}.$$

Given n observations $Y_1, \ldots, Y_n$, our goal is to predict the total number of successes Y_{n+1} in r_{n+1} trials while accounting for model selection. We denote the predicted probability of success $\hat{p}_M^{n+1}$ and its value is given by

$$\hat{p}_M^{n+1} = \frac{1}{1+e^{-X_M'^{n+1}\hat{\beta}_M}}.$$

Recalling that

$$E(Y_{n+1}) = r_{n+1} \cdot g^{-1}(X'^{n+1}\beta) = r_{n+1} \cdot p_{n+1},$$

the form of the post-model selection AN PI for a binomial random variable is

$$PI(M)^{AN} = r_{n+1} \cdot \hat{p}_M^{n+1} \pm z_{1-\alpha/2} \cdot r_{n+1} \sqrt{\frac{e^{-2\hat{\eta}_M^{n+1}}}{\left(1+e^{-\hat{\eta}_M^{n+1}}\right)^4} X_M'^{n+1} Var(\hat{\beta}_M) X_M^{n+1} + \frac{1}{r^{n+1}} \hat{p}_M\,(1-\hat{p}_M)} \tag{4.2}$$

where the factor r_{n+1} in the width of the intervals comes from the the distribution in Eq. 3.9 being multiplied by this factor. The interval in Eq. 4.2 gives a prediction interval for total number of successes in r_{n+1} trials.

In the setting described above, our simulations are as follows. Let $X \sim MVN_p(0, I_p)$ and

$$\beta = (\beta_1, \ldots, \beta_{25})' = (0.252, 0.171, -0.268, 0.09, 0, \ldots 0)'.$$

Now we calculate the estimated coverage using Eq. 4.1. Again, we want coverage close to 0.95 and narrow width.

For the interval using $\hat{C}^{MC}$, we again use 100 Monte Carlo samples and a 70/30 train/test split.

To implement the bootstrap based intervals, we use the same bootstrap procedure to generate the bootstrap intervals as what is outlined in Sec. 3.2. In this scenario we do not need to estimate a dispersion parameter because $\phi = \frac{1}{r}$, is known and hence the same for each bootstrap sample. The steps are as follows. For $b = 1, \ldots, B$:

- generate $\hat{\mu}_{M,b} = g^{-1}(X_{M,b}^{n+1}\hat{\beta}_{M,b}) = \frac{1}{1+e^{-X_{M,b}^{n+1}\hat{\beta}_{M,b}}}$ using the identity link function and the bootstrap sample, denoted $\mu_1^*, \ldots, \mu_B^*$
- generate observations $y_1^* \mid \mu_1^*, r, \ldots, y_B^* \mid \mu_B^*, r$, by randomly generating a new observation from $Bin(r, \hat{p}_{M,b})$.

The simulated results are given in Table 3.

For $n = 50$, the Naive interval has very poor coverage for both AIC and BIC. Using AIC and the AN interval results in undercoverage, but it is much better than the Naive PI. This is also true using BIC as the MSP. Both S-boot and boot are conservative, give coverage larger than the stated coverage. The width of both S-boot and boot make the intervals fairly uninformative despite having better coverage than Naive and AN. Finally, we observe $\hat{C}^{MC}$ performs noticeably better than the other PI's. This suggests that the cross validation step to widen the asymptotic normal PI is useful.

Looking at the $n = 100$ and $n = 200$ cases, we see the Naive interval is worse than in the smaller sample case. AN gives very good coverage and the smallest width among all of the PIs for both AIC and BIC. The other 3 PIs, S-boot, boot, and $\hat{C}^{MC}$ give close stated coverage but they are slightly wider than the AN interval. When $n = 200$ we see Naive is by far the worst among the 5 PIs, but the other 4 are roughly the same with AN and $\hat{C}^{MC}$ having perhaps slightly better coverage and narrower PIs than S-boot and boot.

These results confirm two main points. First, the AN PI achieves the stated 95% coverage as given in Theorem 3.1 when the sample size is large enough. Second $\hat{C}^{MC}$ always gives appropriate coverage, and appears to reduce to AN as n increases. This leads us to recommend using $\hat{C}^{MC}$, especially for small and intermediate sample sizes, and use AN for large n.

Table 3: Simulation results for binomial data with $r = 30$

n	MSP	Interval	Coverage	Avg Width (SE)
50	AIC	Naive	.51	4.9 (1.35)
		AN	0.83	9.51 (0.93)
		$\hat{C}^{MC}$	0.97	15.55 (0.78)
		boot	1	20.82 (3.35)
		S-boot	1	28.31 (2.22)
	BIC	Naive	.48	4.03 (1.08)
		AN	0.90	9.56 (0.76)
		$\hat{C}^{MC}$	0.98	14.35 (0.74)
		boot	1	17.81 (3.06)
		S-boot	1	23.36 (3.93)
100	AIC	Naive	.46	3.29 (0.86)
		AN	0.94	9.42 (0.87)
		$\hat{C}^{MC}$	0.99	13.01 (0.61)
		boot	0.99	11.99(1.66)
		S-boot	0.99	14.11(2.22)
	BIC	Naive	.43	2.77 (0.69)
		AN	0.95	9.38 (0.79)
		$\hat{C}^{MC}$	0.99	12.67 (0.54)
		boot	0.99	11.57 (1.47)
		S-boot	1	13.38 (2.00)
200	AIC	Naive	.37	2.43 (0.90)
		AN	0.94	8.91 (1.60)
		$\hat{C}^{MC}$	0.94	9.15 (0.87)
		boot	0.97	9.37 (1.96)
		S-boot	0.97	10.08(2.10)
	BIC	Naive	.37	2.18 (0.79)
		AN	0.94	8.93 (1.56)
		$\hat{C}^{MC}$	0.95	9.01 (0.83)
		boot	0.98	9.30 (1.82)
		S-boot	0.97	9.75 (1.92)

Again, we list the optimal cross validation constants in Table 4. As in the Gaussian case, we see that $\hat{C}^{MC}$ decreases as the sample size increases. However, in the Binomial case, $\hat{C}^{MC}$ is noticeably larger than the Gaussian case. This may be due to the fact that we are using a normal PI for data that is not normal.

Table 4: Binomial cross validation results for the optimal choice for $\hat{C}^{MC}$

n	MSP	$\hat{C}^{MC}$
50	AIC	5.00
	BIC	4.40
100	AIC	4.10
	BIC	3.90
200	AIC	3.00
	BIC	2.85

5 Discussion

Our main contribution is the PI in Theorem 3.1, and the small sample correction using $\hat{C}^{MC}$ given in Eq. 3.14. Much of the literature on GLM prediction has focused on *confidence* intervals around predictors, which we refer to as the 'Naive' PI, rather than true *prediction* intervals. That is, it is common for analysts to apply the inverse link function to the endpoints of a confidence interval in the linear predictor scale. This approach does not account for uncertainty appropriately because it uses the variability on the linear predictor scale rather than the data scale. Here we have presented PI's that are derived on the model-scale rather than the linear-predictor scale.

We have presented several PI's that take model uncertainty (as well as parameter uncertainty) into account. The PI derived in Theorem 3.1 accounts for model uncertainty via the consistency of the MSP. This PI severely underperforms in terms of predictive coverage in small sample size, e.g. $n \approx p$, cases but as $n \to \infty$ the predictive coverage is roughy the nominal $1-\alpha$ coverage. The boot and S-boot PIs tend to be too wide, suggesting far too much model uncertainty for these PIs to be useful. That is, these two PI's overcorrect the width of the intervals for the uncertainty in the MSP's used here. Again as n increases, both boot and S-boot become usable (due to the MSP choosing the correct model). At this point, the bootstrapping is not necessary because PI^{AN} performs well with large samples.

Taken together, our results provide valid post-model-selection PIs for GLM's for moderate and large samples with the small sample corrected PI being the preferred option overall.

Acknowledgements. The first author acknowledges funding from the University of Nebraska Program of Excellence in Computational Science. No other funding was received and the authors have no conflicts of interest.

References

Bachoc, F., Leeb H. and Pötscher B. (1988). Asymptotic properties of information theoretic methods of model selection. University of California at Davis, Division of Statistics Technical Report.

Bachoc, F., Leeb H. and Pötscher B. (2019). Valid confidence intervals for post-modelselection predictors. *Ann. Stat.,* **47**, 1475–1504.

Berk, R., Brown L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *Ann. Stat.,* **41**, 802–837.

Efron, B. (2014). Estimation and accuracy after model selection. *J. Am. Stat. Assoc.,* **109**, 991–1022.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likeliohood estimator in generalized linear models. *Ann. Stat.,* **13**, 342–368.

Fletcher, D. (2012). Estimating overdispersion when fitting a generalized linear model to sparse data. *Bioketrika* **99**, 230–237.

Hoeting, J., Madigan D. Raftery A. and Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statist. Sci.,* **14**, 382–417.

Hong, L., Kuffner T. and Martin R. (2018a). On prediction of future insurance claims when the model is uncertain. Preprint.

Hong, L., Kuffner T. and Martin R. (2018b). On overfitting and post-selection uncertainty assessments. *Biometrika* **105**, 221–224.

Kuchibhotla, A.K., Brown L.D., Buja A., Cai, J., George, E. and Zhao, L. (2020). A model free perspective for linear regression: Uniform-in-model bounds for post selection inference. *Ann. Stat.,* **49**, 2953–2981.

Leeb, H. (2009). Conditional predictive inference post model selection. *Ann. Stat.,* **37**, 2838–2876.

Leeb, H., Pötscher B. and Ewald K. (2015). On various confidence intervals postmodel-selection. *Stat. Sci.,* **30**, 216–227.

Stine, R. (1985). Bootstrap prediction intervals for regression. *J. Am. Stat. Assoc.,* **80**, 1026–1031.

Stine, R. (2021). Prediction intervals for glms, gams, and some survival regression models. *Commun. Stat. - Theory Methods*

Appendices

Appendix 5.A Extension to GLMM's

The approach described in Sec. 2.2 to obtain valid prediction intervals after model selection extends naturally to the class of generalized linear mixed

models. Here we assume the random variable $Y|X, \beta, Z, U \sim \mathcal{G}$ where $\mathcal{G}$ is a distribution in the exponential family. We write the linear predictor as

$$\eta = g(E(Y|U)) = g(\mu|U) = X\beta + ZU \tag{5.1}$$

where β is the vector of fixed effects, U is a random effect such that $U \sim N(0, \Sigma_U)$. X and Z are their respective design matrices. The mean function is

$$\mu = E(Y|U) = g^{-1}(\eta) = g^{-1}(X\beta + ZU)$$

and the variance is

$$Var(Y|U) = V_\mu^{1/2} A V_\mu^{1/2}$$

where $V_\mu^{1/2} = \text{diag}\left[\sqrt{V(\mu)}\right]$ and $A = \text{diag}\,[1/a(\phi)]$.

As with GLM's, model selection is often performed when forming predictors. In the GLMM setting model selection can be done on both X and Z, however here we focus on model selection on the design matrix X. Analogous to the GLM case, we state an asymptotic normal predictive interval For GLMM's that is derived in the same way as the GLMM. The only difference is the random effects part of the linear predictor. However, recall the random effects have expectation 0, so the location of the asymptotic distribution does not change. The variance, on the other hand, does increase. This is seen in Eq. 5.2 as the width has an extra term for the variance of the random effects: We get that $PI(M, C_\alpha) = g^{-1}(\hat{\eta}_M^{n+1}) \pm$

$$C_\alpha \sqrt{\left[\frac{d}{d\eta} g^{-1}(\hat{\eta}_M^{n+1})\right]^2 \left(X_M'^{n+1} Var(\hat{\beta}_M) X_M^{n+1} + Z'^{n+1} Var(\hat{u}) Z^{n+1}\right) + a(\phi) V(\hat{\mu})}. \tag{5.2}$$

Appendix 5.B GLMM Bootstrap Intervals

As with the GLM AN interval, we can approximate the variance of $g^{-1}(\hat{\eta}_M^{n+1})$ using bootstrapping and replace Eq. 5.2 with

$$PI(M, C_\alpha) = g^{-1}(\hat{\eta}_M^{n+1}) \pm z_{1-\alpha/2} \sqrt{\hat{Var}(g^{-1}(\hat{\eta}_M^{n+1}))^{boot} + a(\phi) V(\hat{\mu})} \tag{5.3}$$

where $\hat{Var}(g^{-1}(\hat{\eta}_M^{n+1}))^{boot}$ is simply the variance of the bootstrapped distribution of $g^{-1}(\hat{\eta}_M^{n+1})$.

We can also use the bootstrap approach, in same way as with the GLM, to obtain a bootstrap distribution for a new outcome. In the GLMM

setting, we bootstrap the expected value of the distribution for a new outcome,

$$\hat{\mu}_M = g^{-1}(X_M'^{n+1}\hat{\beta} + Z'^{n+1}\hat{u}).$$

and we bootstrap to estimate the distribution of $\hat{\phi}_M$ when necessary. This gives bootstrap samples $\hat{\mu}_{M,b}$ and $\hat{\phi}_{M,b}$ for $b = 1, \ldots, B$. Then for each of the bootstrapped mean $\hat{\mu}_{M,b}$ and dispersion parameter $\hat{\phi}_{M,b}$, we generate a new observation from the distribution of $Y^{n+1}|X^{n+1}, \mu, \phi$, i.e. $\mathcal{G}$.

Let $\hat{p}(\hat{\mu})$ be the bootstrapped density of Eq. 3.17, $\hat{p}(\phi)$ be the bootstrapped density of ϕ_M, and $\hat{p}(Y^{n+1})$ be the resulting estimated density of Y^{n+1}.

The procedure is as follows. For $b = 1, \ldots, B$,

- generate $\hat{\mu}_{M,b} = g^{-1}(X_{M,b}^{n+1}\hat{\beta}_{M,b} + Z'^{n+1}\hat{u})$ using the known link function and the bootstrap sample, denoted $\mu_1^*, \ldots, \mu_B^*$
- if there is a dispersion parameter to estimate, generate B bootstrap replications of ϕ_M, denoted $\phi_1^*, \ldots, \phi_B^*$
- generate observations $y_1^* \mid \mu_1^*, \phi_1^*, \ldots, y_B^* \mid \mu_B^*, \phi_B^*$, by randomly generating a new observation from $\mathcal{G}$.

The sample $y_1^*, \ldots, y_B^*$ can used to obtain the predictive interval by extracting the appropriate percentile interval from this distribution. Thus the PI is

$$[q_{1-\alpha/2}^*, q_{\alpha/2}^*], \tag{5.4}$$

the $1-\alpha/2$ and $\alpha/2$ quantiles from $y_1^*, \ldots, y_B^*$ which inherits the uncertainty of M, $\hat{\beta}$ and $\hat{u}$. These intervals are implementable assuming we already have estimates $\hat{\beta}$ and $\hat{u}$. Regardless of which method is used to form predictors, we theoretically can use both intervals Eq. 5.3 or Eq. 5.4 because predictors and mean and variance functions, as well as the uncertainty associated with the MSP can be obtained through the bootstrap procedure. Thus, a closed form solution of parameter estimates is not necessary to obtain valid PI's. Also, we can, at least theoretically, still use both of the intervals presented to account for the uncertainty of model selection in the random effects design matrix.

Appendix 5.C Computational Issues for GLMM's

The theoretical and bootstrap based intervals we have proposed to capture the uncertainty of model selection are not implementable, at least yet.

This is due to convergence issues with implementing GLMM's. Estimation in GLMM's requires integrating out the random effects, and these integrals do not have closed form solutions. Thus, numerical integration is necessary, making the integrals computationally hard.

In practice, now, there is no a single best approach so one tries many approaches until the algorithm converges. Once convergence is achieved, classical approaches to assess model fit are used. In the bootstrapping approach, we require estimation over many repeated samples of the data and this would require convergence of the estimates in the GLMM over each resample. The estimates require numerical integration for each resample of the data which requires a person trying several algorithms until one works. We attempted this, but we were unsuccessful because convergence in each resample using a fixed numerical integration method is not feasible.

Dean Dustin
Charles Schwab, 9899 Schwab Way,
Lone Tree 80124 CO, USA
E-mail: ddustin8@huskers.unl.edu

Bertrand Clarke
Department of Statistics, U.
Nebraska-Lincoln, 340 Hardin Hall
North, Lincoln 68583-0963 NE, USA
E-mail: bclarke3@unl.edu

Paper received: 1 March 2024