Prediction Intervals for GLMs, GAMs, and Some Survival Regression Models

David J. Olive, Rasanji C. Rathnayake, and Mulubrhan G. Haile
School of Mathematical & Statistical Sciences
Southern Illinois University
Carbondale, Illinois 62901-4408
dolive@siu.edu rathnaya@siu.edu haile@siu.edu

Keywords Lasso; logistic regression; Poisson regression; variable selection; Weibull Regression.

Mathematics Subject Classification Primary 62F25; Secondary 62F12.

Abstract

Consider a regression model, $Y|\boldsymbol{x} \sim D(\boldsymbol{x})$, where D is a parametric distribution that depends on the $p \times 1$ vector of predictors $\boldsymbol{x}$. Generalized linear models, generalized additive models, and some survival regression models have this form. To obtain a prediction interval for a future value of the response variable Y_f given a vector of predictors $\boldsymbol{x}_f$, apply the nonparametric shorth prediction interval to $Y_1^*, ..., Y_B^*$ where the Y_i^* are independent and identically distributed from the distribution $\hat{D}(\boldsymbol{x}_f)$ which is a consistent estimator of $D(\boldsymbol{x}_f)$. A second prediction interval modifies the shorth prediction interval to work after variable selection and if $p > n$ where n is the sample size. Competing prediction intervals, when they exist, tend to be for one family of D (such as Poisson regression), tend to need $n \geq 10p$, and usually have not been proven to work after variable selection.

1. Introduction

This paper presents simple large sample $100(1-\delta)\%$ prediction intervals (PIs) for a future value of the response variable Y_f given a $p \times 1$ vector of predictors $\boldsymbol{x}_f$ and training data $(\boldsymbol{x}_1, Y_1), ..., (\boldsymbol{x}_n, Y_n)$ for a parametric regression model $Y|\boldsymbol{x} \sim D(\boldsymbol{x})$, where D is a parametric distribution that depends on the $p \times 1$ vector of predictors $\boldsymbol{x}$. Often the conditioning and

subscript i is suppressed: $Y \sim D(\boldsymbol{x})$ means $Y_f|\boldsymbol{x}_f \sim D(\boldsymbol{x}_f)$ and $Y_i|\boldsymbol{x}_i \sim D(\boldsymbol{x}_i)$ for $i = 1, ..., n$.

In a 1D regression model, the response variable Y is conditionally independent of $\boldsymbol{x}$ given the sufficient predictor $SP = h(\boldsymbol{x})$, where the real valued function $h : \mathbb{R}^p \to \mathbb{R}$. The estimated sufficient predictor $ESP = \hat{h}(\boldsymbol{x})$. An important special case is a model with a linear predictor $h(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$ where $ESP = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$. A parametric 1D regression model is $Y|\boldsymbol{x} \sim D(h(\boldsymbol{x}), \boldsymbol{\gamma})$ where the parametric distribution D depends on $\boldsymbol{x}$ only through $h(\boldsymbol{x})$, and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of parameters. Then $\hat{D}(\boldsymbol{x}) = D(\hat{h}(\boldsymbol{x}), \hat{\boldsymbol{\gamma}})$. This class of models includes the generalized linear model (GLM), and the generalized additive model (GAM), where Y is independent of $\boldsymbol{x} = (1, x_2, ..., x_p)^T$ given the additive predictor (AP) where, for example, $AP = SP = \alpha + \sum_{j=2}^{p} S_j(x_j)$ for some (usually unknown) functions S_j. Then the estimated additive predictor $EAP = ESP = \hat{\alpha} + \sum_{j=2}^{p} \hat{S}_j(x_j)$.

A large sample $100(1-\delta)\%$ prediction interval (PI) for Y_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1 - \delta$ as the sample size $n \to \infty$. A large sample $100(1-\delta)\%$ PI is asymptotically optimal if it has the shortest asymptotic length: the length of $[\hat{L}_n, \hat{U}_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the population shorth: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

The shorth(c) estimator of the population shorth is useful for making asymptotically optimal prediction intervals if the data are independent and identically distributed (iid). Let $Z_{(1)}, ..., Z_{(n)}$ be the order statistics of $Z_1, ..., Z_n$. Then let the shortest closed interval containing at least c of the Z_i be

$$\text{shorth(c)} = [\text{Z}_{(\text{s})}, \text{Z}_{(\text{s+c}-1)}]. \tag{1}$$

Let $\lceil x \rceil$ be the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. Let $k_n = \lceil n(1-\delta) \rceil$. Frey (2013) showed that for large $n\delta$ and iid data, the shorth(k_n) PI has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$, and used the shorth($c$) estimator as the large sample $100(1-\delta)\%$ PI where

$$c = \min(n, \lceil n[1 - \delta + 1.12\sqrt{\delta/n}\ ]\ \rceil). \tag{2}$$

The maximum undercoverage occurs for the family of uniform distributions, where the population shorth is not unique. Of course the length of the population shorth is unique.

The large sample $100(1-\delta)\%$ shorth PI (2) may or may not be asymptotically optimal if the $100(1-\delta)\%$ population shorth is $[L_s, U_s]$ and $F(x)$ is not strictly increasing in intervals $(L_s - \delta, L_s + \delta)$ and $(U_s - \delta, U_s + \delta)$ for some $\delta > 0$. To see the issue, suppose Y has probability mass function (pmf) $p(0) = 0.4$, $p(1) = 0.3$, $p(2) = 0.2$, $p(3) = 0.06$, and $p(4) = 0.04$. Then the 90% population shorth is [0,2] and the $100(1-\delta)\%$ population shorth is [0,3] for $(1-\delta) \in (0.9, 0.96]$. Let $I(Y_i \leq x) = 1$ if $Y_i \leq x$ and 0, otherwise. The empirical cumulative distribution function

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq x)$$

is the sample proportion of $Y_i \leq x$. If $Y_1, ..., Y_n$ are iid, then for fixed x, $n\hat{F}_n(x) \sim$ *binomial*$(n, F(x))$. Thus an asymptotic normal approximation is $\hat{F}_n(x) \sim AN(F(x), F(x)(1 - F(x))/n)$. For the Y with the above pmf, $\hat{F}_n(2) \xrightarrow{P} 0.9$ as $n \to \infty$ with $P(\hat{F}_n(2) < 0.9) \to 0.5$ and $P(\hat{F}_n(2) \geq 0.9) \to 0.5$ as $n \to \infty$. Hence the large sample 90% PI (2) will be [0,2] or [0,3] with probabilities $\to 0.5$ as $n \to \infty$ with expected asymptotic length 2.5 and expected asymptotic coverage 0.93. However, the large sample $100(1-\delta)\%$ PI (2) converges to [0,3] and is asymptotically optimal with asymptotic coverage 0.96 for $(1-\delta) \in (0.9, 0.96)$.

We will illustrate the new prediction intervals with the following three types of regression models. The *binomial logistic regression model* is $Y_i \sim \text{binomial}\left(\text{m}_\text{i}, \rho = \frac{\text{e}^{\text{SP}}}{1+\text{e}^{\text{SP}}}\right)$. Then $E(Y_i|SP) = m_i\rho(SP)$ and $V(Y_i|SP) = m_i\rho(SP)(1 - \rho(SP))$, and $\hat{E}(Y_i|\boldsymbol{x}_i) = m_i\hat{\rho} = \frac{m_i e^{ESP}}{1+e^{ESP}}$ is the estimated mean function. The binary logistic regression model has $m_i \equiv 1$ for $i = 1, ..., n$. Note that $\hat{D}(\boldsymbol{x}_f) \sim \text{binomial}(\text{m}_\text{f}, \rho(\text{ESP}))$ where if $SP = h(\boldsymbol{x})$, then $ESP = \hat{h}(\boldsymbol{x}_f)$. We will use the GLM with $ESP = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}$, and we will also use the GAM. Note that Y_f is the number of "successes" in the known number m_f of binomial trials.

A *Poisson regression (PR) model* $Y \sim \text{Poisson}\left(\text{e}^{\text{SP}}\right)$ has $E(Y|SP) = V(Y|SP) = \exp(SP)$. The estimated mean and variance functions are $\hat{E}(Y|\boldsymbol{x}) = e^{ESP}$, and $\hat{D}(\boldsymbol{x}_f) \sim \text{Poisson}\left(\text{e}^{\text{ESP}}\right)$. We will use the GLM and the GAM.

The *Weibull proportional hazards regression model* is

$$Y|SP \sim W(\gamma = 1/\sigma, \lambda_0 \exp(SP)),$$

where $\lambda_0 = \exp(-\alpha/\sigma)$, and Y has a Weibull $W(\gamma, \lambda)$ distribution if the probability density function (pdf) of Y is

$$f(y) = \lambda\gamma y^{\gamma-1} \exp[-\lambda y^{\gamma}]$$

for $y > 0$. The data is $(T_i, \delta_i, \boldsymbol{x}_i)$ where $T_i = Y_i$ is the observed survival time if $\delta_i = 1$, and T_i is the right censored survival time if $\delta_i = 0$. The PIs are for the survival times, not censored survival times. Then $\hat{D}(\boldsymbol{x}_f) \sim W(\hat{\gamma} = 1/\hat{\sigma}, \hat{\lambda}_0 \exp(ESP))$. We will use $ESP = \boldsymbol{x}_f^T \hat{\boldsymbol{\beta}}$. If Y follows the above model, then $\log(Y)$ follows the Weibull accelerated failure time model $\log(Y) = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}_A + \sigma e$ where the variance $V(e) = 1$, and the e_i are iid from a smallest extreme value SEV(0,1) distribution. Then $\boldsymbol{\beta} = \boldsymbol{\beta}_A/\sigma$.

Section 2 describes the new prediction intervals, and Section 3 gives a simulation.

2. The New Prediction Intervals

The first new large sample $100(1-\delta)\%$ prediction interval for Y_f is

$$[Y^*_{(s)}, Y^*_{(s+c-1)}] \text{ with } \mathrm{c} = \min(\mathrm{B}, \lceil \mathrm{B}[1 - \delta + 1.12\sqrt{\delta/\mathrm{B}}\ ]\ \rceil), \tag{3}$$

and applies the shorth(c) prediction interval to the parametric bootstrap sample $Y^*_1, ..., Y^*_B$ where the Y^*_i are iid from the distribution $\hat{D}(\boldsymbol{x}_f)$. If $Y|\boldsymbol{x}_f \sim D(\boldsymbol{x}_f)$ and the regression method produces a consistent estimator $\hat{D}(\boldsymbol{x}_f)$ of $D(\boldsymbol{x}_f)$, then this new prediction interval is a large sample $100(1-\delta)\%$ PI. For a parametric 1D regression model $Y|\boldsymbol{x}_f \sim D(h(\boldsymbol{x}_f), \boldsymbol{\gamma})$, we need $(\hat{h}(\boldsymbol{x}_f), \hat{\boldsymbol{\gamma}})$ to be a consistent estimator of $(h(\boldsymbol{x}_f), \boldsymbol{\gamma})$.

For models with a linear predictor, we will want prediction intervals after variable selection or model selection. Variable selection is the search for a subset of predictor variables that can be deleted with little loss of information if n/p is large, and so that the model with the remaining predictors is useful for prediction. Following Olive and Hawkins (2005), a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S, \tag{4}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given

that the subset S is in the model. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Suppose that S is a subset of I and that model (4) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\mathbf{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I, \tag{5}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \mathbf{0}$ if $S \subseteq I$.

Forward selection or backward elimination with the Akaike (1973) AIC criterion or Schwarz (1978) BIC criterion are often used for GLM variable selection. GLM model selection with lasso and the elastic net is also common. See Hastie et al. (2015, ch. 3) and Friedman et al. (2010). *Lasso variable selection* applies the regression method, such as a GLM, to the active predictors with nonzero coefficients selected by lasso. For $n \geq 10p$, Olive and Hawkins (2005) suggested using multiple linear regression variable selection software with the Mallows (1973) C_p criterion to get a subset I, then fit the GLM using Y and $\boldsymbol{x}_I$. If the regression model contains a $q \times 1$ vector of parameters $\boldsymbol{\gamma}$, then we may need $n \geq 10(p+q)$.

The prediction interval (3) can have undercoverage if n is small compared to the number of estimated parameters. The modified shorth PI (6) inflates PI (3) to compensate for parameter estimation and model selection. Let d be the number of variables $x_1^*, ..., x_d^*$ used by the full model, forward selection, backward elimination, lasso, or lasso variable selection. (We could let $d = j$ if j is the degrees of freedom of the selected model if that model was chosen in advance without model or variable selection. Hence $d = j$ is not the model degrees of freedom if model selection was used. For a GAM full model, suppose the "degrees of freedom" d_i for $S(x_i)$ is bounded by k. We could let $d = 1 + \sum_{i=2}^{p} d_i$ with $p \leq d \leq pk$.) We want $n \geq 10d$, and the prediction interval length will be increased (penalized) if n/d is not large. For the second new prediction interval, let $q_n = \min(1 - \delta + 0.05, 1 - \delta + d/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta d/n), \quad \text{otherwise.}$$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 1-\delta$. Then compute the shorth(c_{mod}) PI

$$[Y^*_{(s)}, Y^*_{(s+c_{mod}-1)}] \text{ with } \mathrm{c_{mod}} = \min(\mathrm{B}, \lceil \mathrm{B}[\mathrm{q_n} + 1.12\sqrt{\delta/\mathrm{B}}\ ]\ \rceil). \quad (6)$$

Olive (2007, 2013a, 2018) and Pelawa Watagoda and Olive (2020) used similar correction factors for additive error regression models $Y = h(\boldsymbol{x}) + e$ since the maximum simulated undercoverage was about 0.05 when $n = 20d$. If a $q \times 1$ vector of parameters $\boldsymbol{\gamma}$ is also estimated, we may need to replace d by $d_q = d + q$.

If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$ is the estimator that minimized the variable selection criterion, then $\hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. This estimator is needed since typically $\hat{\boldsymbol{\beta}}_{I_{min}}$ is not a consistent estimator of any vector $\boldsymbol{\beta}_I$, e.g. $\boldsymbol{\beta}_I = (\beta_1, \beta_3)^T$. We will show that $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ in the following paragraph.

Hong et al. (2018) explain why classical PIs after AIC variable selection may not work. Fix p and let I_{min} correspond to the predictors used after variable selection. To show that (3) and (6) are large sample prediction intervals for a parametric 1D regression model with $SP = \boldsymbol{x}^T\boldsymbol{\beta}$, we need to show that $(\hat{\boldsymbol{\beta}}_{I_{min},0}, \hat{\gamma}_{I_{min}})$ is a consistent estimator of $(\boldsymbol{\beta}, \gamma)$. Suppose $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. This assumption tends to hold for AIC and BIC. See Charkhi and Claeskens (2018), Claeskens and Hjort (2008, pp. 70, 101, 102, 114, 232), and Haughton (1988, 1989) for more information and references about this assumption. For lasso variable selection, the assumption holds if lasso is a consistent estimator. See Pelawa Watagoda and Olive (2020) and Rathnayake and Olive (2021). Suppose model (4) holds with $S \subseteq I_j$. Then under regularity conditions that are often mild, $(\hat{\boldsymbol{\beta}}_{I_j}, \hat{\gamma}_{I_j})$ is a consistent estimator of $(\boldsymbol{\beta}_{I_j}, \gamma)$ with $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$. Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0}), \quad (7)$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the x_i not in I_j. Then $\hat{\gamma}_{I_{min}}$ is a consistent estimator of γ and $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under model (4) if the variable selection criterion AIC or BIC is used with forward selection, backward elimination,

or all subsets. Hence if $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, then (3) and (6) are large sample PIs. Regularity conditions for (3) and (6) to be large sample PIs when $p > n$ are much stronger.

Prediction intervals (3) and (6) often have higher than the nominal coverage if n is large and Y_f can only take on a few values. Consider binary regression where $Y_f \in \{0, 1\}$ and the PIs (3) and (6) are [0,1] with 100% coverage, [0,0], or [1,1]. If [0,0] or [1,1] is the PI, coverage tends to be higher than nominal coverage unless $P(Y_f = 1|\boldsymbol{x}_f)$ is near δ or $1 - \delta$, e.g., if $P(Y_f = 1|\boldsymbol{x}_f) = 0.01$, then [0,0] has coverage near 99% even if $1 - \delta < 0.99$.

3. Examples and Simulations

Example 1. For the Ceriodaphnia data of Myers et al. (2002, pp. 136-139), the response variable Y is the number of Ceriodaphnia organisms counted in a container. The sample size was $n = 70$, and the predictors were a constant (x_1), seven concentrations of jet fuel (x_2), and an indicator for two strains of organism (x_3). The jet fuel was believed to impair reproduction so high concentrations should have smaller counts. Figure 1 shows the response plot of ESP versus Y for this data. In this plot, the lowess curve is represented as a jagged curve to distinguish it from the estimated Poisson regression mean function (the exponential curve). The horizontal line corresponds to the sample mean $\overline{Y}$. We also computed PI (6) using $\boldsymbol{x}_f = \boldsymbol{x}_i$ for $i = 1, ..., n$ corresponding to the observed training data $(\boldsymbol{x}_i, Y_i)$. The circles correspond to the Y_i and the $\times$'s to the PIs (6) with $d = 3$. The $n = 70$ large sample 95% PIs contained 97% of the Y_i. There was no evidence of overdispersion for this example. There were 5 replications for each of the 14 strain–species combinations, which helps show the bootstrap PI variability tracks the data variability when $B = 1000$. Increasing B from 1000 decreases the average PI length slightly, but using $B = 1000000$ gave a plot very similar to Figure 1 with similar coverage. Using $B = 50$ had longer PIs and sometimes had undercoverage. Using $B = 1000$ several times gave coverage between 97% and 100%.

This example illustrates a useful goodness of fit diagnostic: if the model D is a useful approximation for the data and n is large enough, we expect the coverage on the training data to be close to or higher than the nominal coverage $1 - \delta$. For example, there may be undercoverage if a Poisson regression model is used when a negative binomial regression

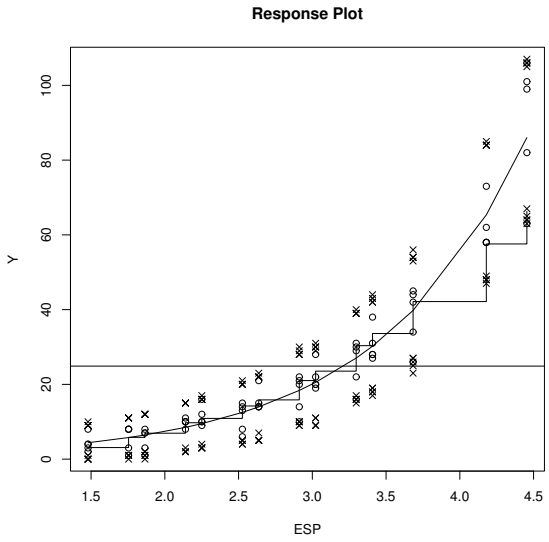

Figure 1: Ceriodaphnia Data Response Plot.

model is needed, as illustrated in the following example.

Example 2. For the species data of Johnson and Raven (1973), the response variable is the total *number of species* recorded on each of $n = 29$ islands in the Galápagos Archipelago. We used a constant and the logarithm of four predictors *endem* = the number of endemic species (those that were not introduced from elsewhere), the *area* of the island, the *distance* to the closest island, the *areanear* = the area of the closest island. The Poisson regression response plot looks good, but Olive (2017b, pp. 438-440) showed that there is overdispersion and that a negative binomial regression model fits the data well. When the incorrect Poisson regression model was used, the n large sample 95% PIs (6) contained 89.7% of the Y_i.

Example 3. The Flury and Riedwyl (1988, pp. 5-6) banknote data consists of 100 counterfeit and 100 genuine Swiss banknotes. The response variable is an indicator for whether the banknote is counterfeit. The six predictors are measurements on the banknote: *bottom, diagonal, left, length, right,* and *top.* We used a constant, right, and bottom as predictors to get a model that did not have perfect classification. The response plot for this model is shown in the left plot of Figure 2 with $Z = Z_i = Y_i/m_i = Y_i$ and the large sample

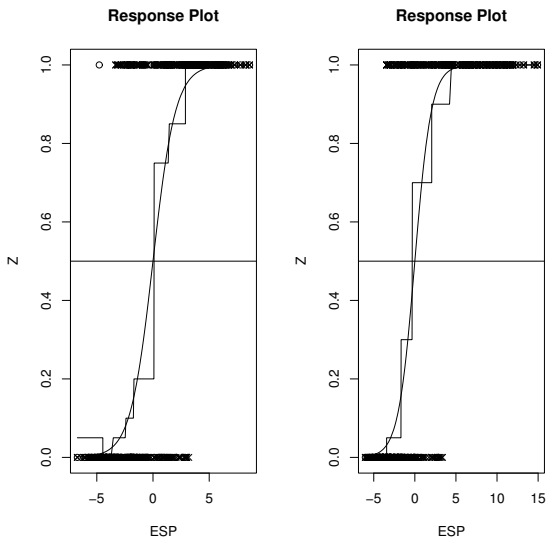

Figure 2: Banknote Data GLM and GAM Response Plots.

95% PIs for $Z_i = Y_i$. The circles correspond to the Y_i and the $\times$'s to the PIs (6) with $d = 3$, and 199 of the 200 PIs contain Y_i. The PI [0,0] that did not contain Y_i corresponds to the circle in the upper left corner. The PIs were [0,0], [0,1], or [1,1] since the data is binary. The mean function is the smooth curve and the step function gives the sample proportion of ones in the interval. The step function approximates the smooth curve closely, hence the binary logistic regression model seems reasonable. The right plot of Figure 2 shows the GAM using right and bottom with $d = 3$. The coverage was 100% for the training data and the GAM had many [1,1] intervals.

For the simulations, generating $\boldsymbol{x}^T\boldsymbol{\beta}$ is important. For example, for binomial logistic regression, typically $-5 \leq \boldsymbol{x}^T\boldsymbol{\beta} \leq 5$ or there can be problems with the MLE. Let $\boldsymbol{x} = (1\ \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1) \times 1$ vector of nontrivial predictors. In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I})$ where the $m = p-1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\mathrm{Cov}(\boldsymbol{z}_i) = \boldsymbol{\Sigma}_{\boldsymbol{z}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi + (m-2)\psi^2]$. Hence

the correlations are $cor(z_i, z_j) = \rho = (2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$ for $i \neq j$. Then $\sum_{j=1}^{k} z_j \sim N(0, k\sigma_{ii} + k(k-1)\sigma_{ij}) = N(0, v^2)$. Let $\boldsymbol{u} = a\boldsymbol{z}/v$. Then $cor(x_i, x_j) = \rho$ for $i \neq j$ where x_i and x_j are nontrivial predictors. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors $\boldsymbol{u}_i$ cluster about the line in the direction of $(1, ..., 1)^T$. Let $SP = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_1 + 1x_{i,2} + \cdots + 1x_{i,k+1} \sim N(\beta_1, a^2)$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (\beta_1, 1, .., 1, 0, ..., 0)^T$ with β_1, k ones and $p - k - 1$ zeros. The default settings for Poisson regression use $\beta_1 = 1 = a$. The default settings for binomial regression use $\beta_1 = 0$ and $a = 5/3$.

The simulation used 5000 runs, so an observed coverage in [0.94, 0.96] gives no reason to doubt that the PI has the nominal coverage of 0.95. The simulation used $B = 1000$; $p = 4, 50, n$, or $2n$; $\psi = 0, 1/\sqrt{p}$, or 0.9; and $k = 1, 19$, or $p - 1$. The simulated data sets are rather small since the R estimators are rather slow. For binomial and Poisson regression, we only computed the GAM for $p = 4$ with $SP = AP = \alpha + S_2(x_2) + S_2(x_3) + S_4(x_4)$ and $d = p = 4$. We only computed the full model GLM if $n \geq 5p$. Lasso and lasso variable selection were computed for all cases for Tables 1 to 5. The regression model was computed from the training data, and a prediction interval was made for the test case Y_f given $\boldsymbol{x}_f$. The "length" and "coverage" were the average length and the proportion of the 5000 prediction intervals that contained Y_f. Two rows per table were used to display these quantities for the first five tables. For a larger simulation, see Rathnayake (2019).

Tables 1 and 2 show some simulation results for Poisson regression. Lasso minimized 10-fold cross validation and lasso variable selection was applied to the selected lasso model. The full GLM, full GAM, and backward elimination (BE in the tables) used PI (3) while lasso, lasso variable selection (LVS in the tables), and forward selection using the Olive and Hawkins (2005) method (OHFS in the tables) used PI (6). For $n \geq 10p$, coverages tended to be near or higher than the nominal value of 0.95, except for lasso and the Olive and Hawkins (2005) method. These two methods sometimes had severe undercoverage. In Table 1, the Poisson counts were not small, so the discreteness of the distribution did not affect the coverage much. For Table 2, $p = 50$, and PI (3) had slight undercoverage for the full GLM

Table 1: Simulated Large Sample 95% PI Coverages and Lengths for Poisson Regression, $p = 4$, $\beta_1 = 5$, $a = 2$

n	ψ	k		GLM	GAM	lasso	LVS	OHFS	BE
100	0	1	cov	0.9500	0.9440	0.7730	0.9664	0.9654	0.9520
			len	77.6072	77.6306	84.1066	81.8374	82.4752	84.1432
400	0	1	cov	0.9580	0.9564	0.7566	0.9622	0.9628	0.9534
			len	82.0126	82.0212	85.5704	83.2692	83.4374	80.9897
100	0.5	1	cov	0.9456	0.9424	0.7646	0.9634	0.9408	0.9512
			len	83.0236	82.9034	90.5822	88.3060	88.6700	79.6887
400	0.5	1	cov	0.9530	0.9500	0.7584	0.9604	0.9566	0.9678
			len	83.8588	83.8292	87.4336	85.1042	85.1434	79.9855
100	0.9	1	cov	0.9492	0.9452	0.7688	0.9646	0.7712	0.9654
			len	78.3554	78.3798	87.0086	84.6072	83.4980	81.5432
400	0.9	1	cov	0.9550	0.9574	0.7606	0.9606	0.7928	0.9513
			len	76.7028	76.7594	80.5070	78.2308	78.2538	80.1298
100	0	3	cov	0.9544	0.9466	0.7798	0.9708	0.9404	0.9487
			len	80.1476	80.1362	92.1372	89.8532	90.3456	79.4565
400	0	3	cov	0.9560	0.9548	0.7514	0.9582	0.9566	0.9567
			len	80.7868	80.8976	85.0642	82.7982	82.7912	79.4522
100	0.5	3	cov	0.9516	0.9478	0.7848	0.9694	0.3324	0.9515
			len	77.1120	77.1130	88.9346	86.4680	85.8634	81.5643
400	0.5	3	cov	0.9568	0.9558	0.7534	0.9636	0.5214	0.9528
			len	80.4226	80.4932	84.7646	82.5590	83.7526	79.9786
100	0.9	3	cov	0.9492	0.9456	0.7882	0.9620	0.7510	0.9554
			len	79.5374	79.6172	91.2052	89.0692	84.5648	81.8544
400	0.9	3	cov	0.9544	0.9546	0.7638	0.9554	0.7384	0.9586
			len	79.7384	79.6906	83.8318	81.6862	81.0882	80.7521

Table 2: Simulated Large Sample 95% PI Coverages and Lengths for Poisson Regression, $p = 50$, $\beta_1 = 5$, $a = 2$

n	ψ	k		GLM	lasso	LVS	OHFS	BE
500	0	1	cov	0.9352	0.7564	0.9598	0.9640	0.9476
			len	81.2668	84.3188	81.8934	85.2922	81.1010
500	0.14	1	cov	0.9370	0.7508	0.9580	0.9628	0.9458
			len	81.1820	84.4530	82.1894	85.2304	81.1146
500	0.9	1	cov	0.9368	0.7630	0.9620	0.8994	0.9456
			len	80.4568	86.3506	84.4942	84.1448	80.4202
500	0	19	cov	0.9388	0.7592	0.9756	0.3778	0.9472
			len	81.6922	96.8546	94.6350	99.7436	81.7218
500	0.14	19	cov	0.9368	0.7556	0.9730	0.2770	0.9438
			len	80.0654	95.2964	93.2748	87.3814	80.1276
500	0.9	19	cov	0.9350	0.7544	0.9536	0.9480	0.9352
			len	79.7324	86.3448	84.0674	83.2958	79.6172
500	0	49	cov	0.9386	0.7104	0.9666	0.1004	0.9364
			len	81.1422	96.4304	94.8818	108.0518	81.2516
500	0.14	49	cov	0.9396	0.7194	0.9558	0.2858	0.9402
			len	79.7874	94.8908	93.2538	86.4234	79.8692
500	0.9	49	cov	0.9380	0.7640	0.9480	0.9512	0.9430
			len	78.8146	85.5786	83.2812	82.4104	78.8316

since $n = 10p$. Table 2 helps illustrate the importance of the correction factor: PI (6) would have higher coverage and longer average length. Lasso was good at choosing subsets that contain S since lasso variable selection had good coverage. The Olive and Hawkins (2005) method is partly graphical, and graphs were not used in the simulation. For the same n, ψ, and k as Table 1, we simulated the Poisson regression model with $p = 4$, and $\beta_1 = 1 = a$. Then the response variable took on few values and the coverages were between 0.959 and 0.984 for the six methods used in Table 1.

Tables 3 and 4 are for binomial regression where only PI (6) was used. For large n, coverage is likely to be higher than the nominal if the binomial probability of success can get close to 0 or 1. For binomial regression, neither lasso nor the Olive and Hawkins (2005) method had undercoverage in any of the simulations with $n \geq 10p$.

For $n \leq p$, good performance needed stronger regularity conditions, and Table 5 shows some results with $n = 100$ and $p = 200$. For $k = 1$, lasso variable selection performed well as did lasso except in the second to last column of Table 5. With $k = 19$ and $\psi = 0$, there was undercoverage since $n < 10(k+1)$. For the dense models with $k = 199$ and $\psi = 0$, there was often severe undercoverage, lasso sometimes picked 100 predictors including the constant, and then lasso variable selection caused the program to fail with 5000 runs. Coverage was usually good for $\psi > 0$ except for the second to last column and sometimes the last column of Table 5.

For the Weibull regression model, there is no constant since the constant appears in the corresponding accelerated failure time model. The data was generated as for the Poisson and Binomial regression, but replace $\boldsymbol{u}$ by $\boldsymbol{x}$ and $p-1$ by p. Let $SP = \boldsymbol{x}_i^T \boldsymbol{\beta} = 1x_{i,1} + \cdots + 1x_{i,k} \sim N(0, a^2)$ for $i = 1, ..., n$. The simulations use $a = 1$ where $\boldsymbol{\beta} = (1, ..., 1, 0, ..., 0)^T$ with k ones and $p-k$ zeros. The right censored Weibull regression data was generated in a manner similar to Zhou (2001) with $\gamma = 4$. Since the Weibull distribution is continuous, the coverage of PI (6) converges to $1 - \delta$. For 5000 runs, we needed $n \geq 100p$ or MLE convergence problems could cause the program to fail often. With $n = 100p$, we occasionally needed to run the program twice to get output. Table 6 shows some results for the full model, and the coverages

Table 3: Simulated Large Sample 95% PI Coverages and Lengths for Binomial Regression, $p = 4$, $m = 40$

n	ψ	k		GLM	GAM	lasso	LVS	OHFS	BE
100	0	1	cov	0.9786	0.9788	0.9774	0.9744	0.9720	0.9726
			len	10.7696	10.7656	10.5332	10.4430	10.1990	10.2016
400	0	1	cov	0.9708	0.9700	0.9696	0.9708	0.9702	0.9688
			len	9.8374	9.8426	9.8292	9.7866	9.7518	9.7548
100	0.5	1	cov	0.9792	0.9720	0.9742	0.9750	0.9724	0.9708
			len	10.6668	10.6426	10.3790	10.3282	10.1060	10.1012
400	0.5	1	cov	0.9678	0.9676	0.9692	0.9670	0.9668	0.9656
			len	9.8352	9.8452	9.8196	9.7890	9.7612	9.7590
100	0.9	1	cov	0.9780	0.9766	0.9762	0.9742	0.9704	0.9714
			len	10.7324	10.7222	10.3774	10.3186	10.1438	10.1602
400	0.9	1	cov	0.9688	0.9672	0.9680	0.9674	0.9684	0.9672
			len	9.7554	9.7646	9.7392	9.7012	9.6778	9.6790
100	0	3	cov	0.9790	0.9750	0.9782	0.9772	0.9780	0.9776
			len	10.6974	10.6960	10.7388	10.7030	10.6956	10.7020
400	0	3	cov	0.9652	0.9652	0.9654	0.9656	0.9650	0.9626
			len	9.7838	9.7878	9.8244	9.7864	9.7800	9.7722
100	0.5	3	cov	0.9780	0.9734	0.9776	0.9766	0.9770	0.9784
			len	10.7224	10.7034	10.7482	10.7042	10.7162	10.7134
400	0.5	3	cov	0.9686	0.9688	0.9726	0.9702	0.9704	0.9706
			len	9.7250	9.7170	9.7460	9.7172	9.7152	9.7290
100	0.9	3	cov	0.9800	0.9798	0.9802	0.9786	0.9698	0.9720
			len	10.6978	10.6994	10.5820	10.5414	10.0660	10.1802
400	0.9	3	cov	0.9682	0.9684	0.9696	0.9674	0.9678	0.9676
			len	9.8146	9.8074	9.8364	9.8190	9.7594	9.7764

Table 4: Simulated Large Sample 95% PI Coverages and Lengths for Binomial Regression, $p = 50$, $m = 7$

n	ψ	k		GLM	lasso	LVS	OHFS	BE
1000	0	1	cov	0.9896	0.9838	0.9802	0.9798	0.9798
			len	4.0008	3.6666	3.5744	3.5838	3.5842
1000	0.14	1	cov	0.9868	0.9818	0.9782	0.9774	0.9770
			len	4.0422	3.6836	3.6158	3.6226	3.6312
1000	0.9	1	cov	0.9894	0.9794	0.9796	0.9800	0.9798
			len	4.0214	3.5994	3.5794	3.6122	3.6114
1000	0	19	cov	0.9888	0.9870	0.9848	0.9814	0.9812
			len	4.0294	3.9730	3.8438	3.7110	3.7030
1000	0.14	19	cov	0.9872	0.9846	0.9852	0.9804	0.9806
			len	4.0376	3.8350	3.7834	3.7170	3.7066
1000	0.9	19	cov	0.9884	0.9804	0.9808	0.9802	0.9772
			len	4.0348	3.6170	3.5948	3.6226	3.6216
1000	0	49	cov	0.990	0.9904	0.9904	0.9900	0.9904
			len	4.0428	4.0726	4.0528	4.0490	4.0460
1000	0.14	49	cov	0.9866	0.9866	0.9856	0.9806	0.9796
			len	4.0396	3.9044	3.8640	3.7046	3.6988
1000	0.9	49	cov	0.9874	0.9808	0.9792	0.9790	0.9772
			len	4.0660	3.6444	3.6230	3.6556	3.6490

Table 5: Simulated Large Sample 95% PI Coverages and Lengths, $n = 100$, $p = 200$

		BR	m=7	BR	m=40	PR,a=1	$\beta_1 = 1$	PR,a=2	$\beta_1 = 5$
ψ,k		lasso	LVS	lasso	LVS	lasso	LVS	lasso	LVS
0	cov	0.9912	0.9654	0.9836	0.9602	0.9816	0.9612	0.7620	0.9662
1	len	4.2774	3.8356	11.3482	11.001	7.8350	7.5660	93.7318	91.4898
0.07	cov	0.9904	0.9698	0.9796	0.9644	0.9790	0.9696	0.7652	0.9706
1	len	4.2570	3.9256	11.4018	11.1318	7.8488	7.6680	92.0774	89.7966
0.9	cov	0.9844	0.9832	0.9820	0.9820	0.9880	0.9858	0.7850	0.9628
1	len	3.8242	3.7844	10.9600	10.8716	7.6380	7.5954	98.2158	95.9954
0	cov	0.9146	0.8216	0.8532	0.7874	0.8678	0.8038	0.1610	0.6754
19	len	4.7868	3.8632	12.0152	11.3966	7.8126	7.5188	88.0896	90.6916
0.07	cov	0.9814	0.9568	0.9424	0.9208	0.9620	0.9444	0.3790	0.5832
19	len	4.1992	3.8266	11.3818	11.0382	7.9010	7.7828	92.3918	92.1424
0.9	cov	0.9858	0.9840	0.9812	0.9802	0.9838	0.9848	0.7884	0.9594
19	len	3.8156	3.7810	10.9194	10.8166	7.6900	7.6454	97.744	95.2898
0.07	cov	0.9820	0.9640	0.9604	0.9390	0.9720	0.9548	0.3076	0.4394
199	len	4.1260	3.7730	11.2488	10.9248	8.0784	7.9956	90.4494	88.0354
0.9	cov	0.9886	0.9870	0.9822	0.9804	0.9834	0.9814	0.7888	0.9586
199	len	3.8558	3.8172	10.9714	10.8778	7.6728	7.6602	97.0954	94.7604

Table 6: Simulated Large Sample 95% PI Coverages and Lengths

n	p	k	ψ	cov	len
400	4	1	0	0.9540	1.0520
400	4	1	0.5	0.9532	1.0506
400	4	1	0.9	0.9504	1.0514
1000	10	4	0	0.9458	1.0529
1000	10	4	0.3162	0.9526	1.0512
1000	10	4	0.9	0.9476	1.0554

were good. The simulated data is such that $Y|SP$ and SP depend on a and γ, but not on p, k, or ψ. If n is large and $\hat{\boldsymbol{\beta}}, \hat{\gamma}$, and $\hat{\lambda}_0$ are good estimators, then we would expect the simulated average lengths to be nearly the same. For Table 6, these lengths are near 1.052. Similar remarks apply to Tables 1 to 4.

4. Conclusions

We recommend using PI (6) if there is a good choice for d. PI (3) can be useful if n is large or if $Y|\boldsymbol{x}_f$ takes on few values with high probability. Since PIs (3) and (6) are for a parametric regression model, it is crucial to check that the parametric model is appropriate. For example, if a negative binomial regression model is appropriate, but a Poisson regression model is fit, then the PI coverage will likely be poor, as in Example 2. The response plot of the ESP on the horizontal axis versus the response on the vertical axis is useful. This plot and the OD plot for detecting overdispersion are described in Olive (2013b, 2017b: ch. 13). Olive (2021: ch. 4) shows that these plots can be useful for methods such as lasso and elastic net if n/p is not large, although estimation becomes more difficult.

For the additive error regression model $Y = h(\boldsymbol{x}) + e = SP + e$ where the e_i are iid, which includes the multiple linear regression model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, we do not recommend using a parametric model. It is often incorrectly assumed that $Y_f \sim N(h(\boldsymbol{x}_f), \sigma^2)$, and then the prediction intervals tend to have undercoverage since the error distribution of e has heavier tails than the normal distribution. Find the shorth $[\hat{L}_{nr}, \hat{U}_{nr}]$ of the residuals and use PI

$[\hat{Y}_f + \hat{L}_{nr}, \hat{Y}_f + \hat{U}_{nr}]$ where $\hat{Y}_f = \hat{h}(\boldsymbol{x}_f)$. See Olive (2013a). For multiple linear regression see Olive (2007) for $n \geq 10p$ and Pelawa Watagoda and Olive (2020) for PIs that can work for variable selection and model selection estimators even if n/p is not large. Also see Lei et al. (2018).

For a parametric multivariate regression model, $\boldsymbol{y} \sim D(\boldsymbol{x})$, there are m response variables $\boldsymbol{y} = (Y_1, ..., Y_m)^T$. To find a prediction region for $\boldsymbol{y}_f|\boldsymbol{x}_f$, generate $\boldsymbol{y}_1^*, ..., \boldsymbol{y}_B^*$ where the iid $\boldsymbol{y}_i^* \sim \hat{D}(\boldsymbol{x}_f)$. A large sample $100(1-\delta)\%$ prediction region for $\boldsymbol{y}_f$ that is analogous to PI (3) applies the Olive (2013a, 2017a: p. 152) prediction region to the $\boldsymbol{y}_i^*$. Olive (2017ab, 2018) gave nonparametric prediction regions for the multivariate linear regression model.

There are not many references for prediction intervals for GLMs and GAMs. The prediction intervals tend to have complicated correction factors, lack software, tend to be only applicable to the GLM full model when $n \geq 10p$, and tend to not be asymptotically optimal. The PIs tend to be constructed using Chebyshev's inequality, percentiles of $\hat{D}$, or Bayesian predictive distributions. See Cai et al. (2008), Hall and Maiti (2006), Hall and Rieck (2001), Lawless and Fredette (2005), Ueki and Fueda (2007), Vidoni (2001, 2003), Wasef Hattab (2016), and Wood (2005). The highest density Bayesian credible interval is the population shorth of the posterior distribution, and Chen and Shao (1999) and Olive (2014, p. 364) used the shorth estimator to estimate Bayesian credible intervals.

Generalized linear models were introduced by Nelder and Wedderburn (1972). Useful references for generalized additive models include Hastie and Tibshirani (1986, 1990) and Wood (2017). Yee (2015) considers many parametric (multivariate) regression models.

The simulations were done in *R*. See R Core Team (2018). We used several *R* functions including lasso with the `cv.glmnet` functions from the Friedman et al. (2015) `glmnet` library. The Wood (2017) library `mgcv` was used for fitting a generalized additive model, and the Venables and Ripley (2010) library `MASS` was used for backward elimination. The Therneau and Grambsch (2000) `survival library` and Lumley (2009) `leaps` library were also used.

The data for examples 1 and 2 are available from (http://parker.ad.siu.edu/Olive/sldata.txt). The data for example 3 is available from (https://cran.r-project.org/web/packages/alr3/

index.html) corresponding to Weisberg (2005). The three data sets are also available from the Cook and Weisberg (1999) *Arc* software. The collection of Olive (2021) *R* functions *slpack*, available from (http://parker.ad.siu.edu/Olive/slpack.txt), has some useful functions for the inference. Table entries for Poisson regression were made with `prpisim2` while entries for binomial regression were made with `brpisim`. Table entries for Weibull regression were made with `wpisim`. The functions `prpiplot2` and `lrpiplot` were used to make Figures 1 and 2. The function `prplot` can be used to check the full Poisson regression model for overdispersion. The function `prplot2` can be used to check other Poisson regression models such as a GAM or lasso. See Olive (2021, ch. 4). Sample *R* code is available from (http://parker.ad.siu.edu/Olive/ppRcodepigam.pdf).

Acknowledgments

The authors thank the referees, editor, and associate editor for their work.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In *Proceedings, 2nd international symposium on information theory*, B. N. Petrov and F. Csakim ed. 267-281. Budapest: Akademiai Kiado.

Cai, T., L. Tian, S. D. Solomon, and L. J. Wei. 2008. Predicting future responses based on possibly misspecified working models. *Biometrika* 95 (1):75-92. doi:10.1093/biomet/asm078.

Charkhi, A., and G. Claeskens. 2018. Asymptotic post-selection inference for the Akaike information criterion. *Biometrika* 105 (3):645-664. doi:10.1093/biomet/asy018.

Chen, M. H., and Q. M. Shao. 1999. Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics* 8 (1):69-92. doi:10.1080/10618600.1999.10474802.

Claeskens, G., and N. L. Hjort. 2008. *Model selection and model averaging*. New York: Cambridge University Press.

Cook, R. D., and S. Weisberg. 1999. *Applied regression including computing and graphics*. New York: Wiley.

Flury, B., and H. Riedwyl. 1988. *Multivariate statistics: A practical approach*. New York: Chapman & Hall.

Frey, J. 2013. Data-driven nonparametric prediction intervals. *Journal of Statistical Planning and Inference* 143 (6):1039-1048. doi:10.1016/j.jspi.2013.01.004.

Friedman, J., T. Hastie, N. Simon, and R. Tibshirani. 2015. *glmnet*: Lasso and elastic-net regularized generalized linear models. *R* package version 2.0. http://cran.r-project.org/package=glmnet.

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.

Hall, P., and T. Maiti. 2006. On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society, B* 68 (2):221-238. doi:10.1111/j.1467-9868.2006.00541.x.

Hall, P., and A. Rieck. 2001. Improving coverage accuracy of nonparametric prediction intervals. *Journal of the Royal Statistical Society, B* 63 (4):717-725. doi:10.1111/1467-9868.00308.

Hastie, T., R. Tibshirani, and M. Wainwright. 2015. *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton, FL: CRC Press Taylor & Francis.

Hastie, T. J., and R.J. Tibshirani. 1986. Generalized additive models (with discussion). *Statistical Science* 1 (3):297-318. doi:10.1214/ss/1177013604.

Hastie, T. J., and R.J. Tibshirani. 1990. *Generalized additive models*. London: Chapman & Hall.

Haughton, D. 1989. Size of the error in the choice of a model to fit data from an exponential

family. *Sankhyā, A* 51 (1):45-58.

Haughton, D. M. A. 1988. On the choice of a model to fit data from an exponential family. *The Annals of Statistics* 16 (1):342-355. doi:10.1214/aos/1176350709

Hong, L., T. A. Kuffner, and R. Martin. 2018. On overfitting and post-selection uncertainty assessments. *Biometrika* 105 (1):221-224. doi:10.1093/biomet/asx083

Johnson, M. P., and P. H. Raven. 1973. Species number and endemism, the Galápagos archipelago revisited. *Science* 179 (4076):893-895. doi:10.1126/science.179.4076.893.

Lawless, J. F., and M. Fredette. 2005. Frequentist prediction intervals and predictive distributions. *Biometrika* 92 (3):529-542. doi:10.1093/biomet/92.3.529.

Lei, J., M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association* 113 (523):1094-1111. doi:10.1080/01621459.2017.1307116.

Lumley, T. 2009. *leaps*: Regression subset selection. *R* package version 2.9. https://cran.r-project.org/package=leaps.

Mallows, C. 1973. Some comments on C_p. *Technometrics* 15 (4):661-676. doi:10.2307/1267380.

Myers, R. H., D. C. Montgomery, and G. G. Vining. 2002. *Generalized linear models with applications in engineering and the sciences*. New York: Wiley.

Nelder, J. A., and R. W. M. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society, A* 135 (3):370-380. doi:10.2307/2344614.

Olive, D. J. 2007. Prediction intervals for regression models. *Computational Statistics & Data Analysis* 51 (6):3115-3122. doi:10.1016/j.csda.2006.02.006.

Olive, D. J. 2013a. Asymptotically optimal regression prediction intervals and prediction regions for multivariate data. *International Journal of Statistics and Probability* 2 (1):90-100. doi:10.5539/ijsp.v2n1p90.

Olive, D. J. 2013b. Plots for generalized additive models. *Communications in Statistics: Theory and Methods* 42 (18):2610-2628. doi:10.1080/03610920902923494

Olive, D. J. 2014. *Statistical theory and inference.* New York: Springer.

Olive, D. J. 2017a. *Robust multivariate analysis.* New York: Springer.

Olive, D. J. 2017b. *Linear regression.* New York: Springer.

Olive, D. J. 2018. Applications of hyperellipsoidal prediction regions. *Statistical Papers* 59 (3):913-931. doi:10.1007/s00362-016-0796-1.

Olive, D. J. 2021. *Prediction and statistical learning.* Online course notes. http://parker.ad.siu.edu/Olive/slearnbk.htm.

Olive, D. J., and D. M. Hawkins. 2005. Variable selection for 1D regression models. *Technometrics* 47 (1):43-50. doi:10.1198/004017004000000590.

Pelawa Watagoda, L. C. R., and Olive, D. J. 2020, Comparing six shrinkage estimators with large sample theory and asymptotically optimal prediction intervals. *Statistical Papers*, to appear. doi:10.1007/s00362-020-01193-1

R Core Team. 2018. *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing, www.R-project.org.

Rathnayake, R. C. 2019. Inference for some GLMs and survival regression models after variable selection. Ph.D. Thesis, Southern Illinois University, USA. http://parker.ad.siu.edu/Olive/srasanjiphd.pdf.

Rathnayake, R. C., and D. J. Olive. 2021. Bootstrapping some GLM and survival regression variable selection estimators. Unpublished manuscript. http://parker.ad.siu.edu/Olive/ppbootglm.pdf.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6 (2):461-464. doi:10.1214/aos/1176344136.

Therneau, T. M. and P. M. Grambsch. 2000. *Modeling survival data: Extending the Cox model*. New York: Springer.

Ueki, M., and K. Fueda. 2007. Adjusting estimative prediction limits. *Biometrika* 94 (2):509-511. doi:10.1093/biomet/asm032.

Venables, W. N., and B. D. Ripley. 2010. *Modern applied statistics with S*. fourth ed., New York: Springer.

Vidoni, P. 2001. Improved prediction for continuous and discrete observations in generalized linear models. *Biometrika* 88 (3):881-887. doi:0.1093/biomet/88.3.881.

Vidoni, P. 2003. Prediction and calibration in generalized linear models. *Annals of the Institute of Statistical Mathematics* 55 (1):169-185. doi:10.1007/BF02530492.

Wasef Hattab, M. 2016. A derivation of prediction intervals for gamma regression. *Journal of Statistical Computation and Simulation* 86 (17):3512-3526. doi:10.1080/00949655.2016.1169421.

Weisberg, S. 2005. *Applied linear regression*. third ed., New York: Wiley.

Wood, G. R. 2005. Confidence and prediction intervals for generalised linear accident models. *Accident Analysis & Prevention* 37 (2):267-273. doi:10.1016/j.aap.2004.10.005.

Wood, S. N. 2017. *Generalized additive models: An introduction with R*. second ed., Boca Rotan, FL: Chapman & Hall/CRC.

Yee, T. W. 2015. *Vector generalized linear and additive models: With an implementation in R*. New York: Springer.

Zhou, M. 2001. Understanding the Cox regression models with time–change covariates. *The American Statistician* 55 (2):153-155. doi:10.1198/000313001750358491.