On prediction of future insurance claims when the model is uncertain

by Liang Hong, Todd Kuffner, and Ryan Martin

Abstract

Predictive modeling is arguably one of the most important tasks actuaries face in their day-to-day work. In practice, actuaries may have a number of reasonable models to consider, all of which will provide different predictions. The most common strategy is to first use some kind of model selection tool to select a "best model," and then use that model to make predictions. However, there is reason to be concerned about the use of the classical distribution theory to develop predictions because these ignore the selection effect. Since accuracy of predictions is crucial to the insurer's pricing and solvency, care is needed to develop valid prediction methods. In this paper, we undertake an investigation of the effects of model selection on the validity of classical prediction tools and make some recommendations for practitioners.

Keywords and phrases: Bootstrap; post-selection inference; predictive distribution; regression; variable selection.

1 Introduction

In the current property and casualty insurance practice, actuaries often need to predict the value of a future claim based on data from previous claims. This is done by first specifying a statistical model for claims depending on some unknown parameters, learning about the parameters in some specified way based on the observed claim data, and then converting this fitted model into a predictive distribution for the future claim. There are a variety of ways this process of predicting a future claim can be carried out, and we discuss some of these below, but there is an issue of practical importance lurking behind the scenes. In most applications, there will be many candidate models that could be fit to the observed claim data, *but the actuary will not be sure of which one to use.* Standard practice is to pick one of the candidate models, maybe by using one of the many model selection tools available in the statistical literature, treat the selected model as if it were certain, and proceed with model fitting and prediction as usual. Since the distribution theory used to derive properties of the predictive distribution assume a fixed model, there is reason to be concerned that these properties may fail if the data is used to select a model. Prediction

errors can adversely affect the insurer's pricing, potentially hurting its profitability; they also lead to insufficient reserves and hence jeopardize the insurer's solvency. Therefore, the prediction risk is a serious concern to both the insurer and regulator. Motivated by this fact, the goal of this paper is to assess the effects of model uncertainty and selection on the quality of the predictive distribution. While we choose regression model as a main vehicle for our presentation, the issue under consideration and our conclusion carry over to actuarial model selection in general.

In the extant actuarial science literature, parameter and model uncertainty has received some attention; see, e.g., Cairns (2000), Peters et al (2008), and Hartman and Groendyke (2013), Bignozzi and Tsanakas (2016), Huang et al (2016) and Venter and Sahasrabuddhe (2016). Our paper is different in that its focus is on the effect of model selection. Only recently has the potentially devastating effects of selection on inference been noticed in the statistical literature, so bringing these issues to the attention of actuaries is important and of general interest. We will conduct our investigation in the context of prediction because accurate prediction is a crucial task for all actuaries, and, even in the statistics literature, there has virtually no work (except Kabaila 1995 and Leeb 2009) in the statistics literature on the effect of model selection on the validity of predictive distributions for future loss variables and their corresponding prediction intervals.

There are a variety of model selection criteria available for actuaries such as AIC, BIC, S_p and Mallows' C_p, to name a few. Essentially, each model selection criterion can be classified either as "consistent" or "conservative". For a *conservative* model selection criterion, such as AIC, C_p and S_p, the probability of selecting an incorrect model is asymptotically zero, while for a *consistent* model selection critia, such as BIC and description length criterion, the probability of selecting the most parsimonious correct model is asymptotically one. Kabaila (1995) shows that if a consistent model selection criterion is used, then the resulting predictive intervals will not get the correct coverage even asymptotically. No similar work has been done to investigate the effect of selection on predictive interval. One exception is Leeb (2009) where he develops a selection tool based on a version of cross validation, and rigorously proves that the corresponding prediction intervals are approximately valid. However, his approximate validity result holds only when either the dimension p is large compared to n, or if p is not large but n is unrealistically large; see his Proposition 4.3. Therefore, we conclude that Leeb's method is not satisfactory for the typical case where actuaries face relatively small p and moderate n. In view of the above-mentioned result in Kabaila (1995), investigation along this line should be made using conservative model selectors. Our investigation here focus primarily on selection based on the AIC criterion. But limited investigations using lasso and stepwise selection procedures reveal similar conclusions.

The remainder of the paper is organized as follows. Section 2 gives a brief review of the classical theory of prediction in regression. Section 3 is devoted to reviewing some of the investigations in the statistical literature on the effect of model selection on inference. Next, in Section 4, we shift our focus to the effect of model selection on prediction. There, based on our numerical investigations of the available statistical tools for the cases of practical relevance to actuaries and on the real life dangers of prediction errors, we conclude that the

best strategy for making valid predictions is to use the full model, i.e., not to carry out a variable selection step using a model selector. Finally, we conclude the paper in Section 5 with several remarks and open questions. R code for implementing the simulation study in this paper can be found at `http://www4.stat.ncsu.edu/~rmartin`.

2 Prediction in regression

Property and casualty actuaries often need to model the relationship between the loss (response) variable, Y, and a set of rating (predictor) variables, $X_1, \ldots, X_p$. For example, in personal auto insurance, Y might be the claim amount (or a transformation thereof) and $X_1, \ldots, X_p$ might include driver age, educational level, gender, income, marital status, vehicle model, territory, etc (e.g., Werner and Modlin 2010). Once the model is fully specified and the relationship between the X and Y variables is known, then actuaries can use this model to predict the value of a new loss, $\tilde{Y}$, corresponding to a new set of values $\tilde{x}_1, \ldots, \tilde{x}_p$ of the rating variables. For an introduction to these regression models, see Frees (2010), Frees et al (2014), and the references therein. For concreteness, and because it is the most widely used, we will focus our attention on the standard linear regression model

$$Y = X\beta + \sigma\varepsilon, \tag{1}$$

where Y is the n-vector of loss variables, X is the $n \times p$ matrix of rating variables, ε is a n-vector of iid standard normal errors, β is the p-vector of regression coefficients, and $\sigma > 0$ is the scale parameter; if the model includes an intercept term, then the first column of X consists of a n-vector of 1s. The points we make in this paper, however, are not unique to this simple linear model. Indeed, the same conclusions would apply to, say, generalized linear models (McCullagh and Nelder 1989; de Jong and Heller 2008) among others, but the arguments and calculations would be less transparent for the more complex models.

In the remainder of this section, we review the classical theory of prediction in the linear regression model, but with a slight twist. According to Equation (4.6) in Frees (2010), if the goal is to predict a future claim $\tilde{Y}$, corresponding to a vector of rating variables $\tilde{x}$, possibly different from those in (1), a $100(1-\alpha)\%$ prediction interval is

$$\tilde{x}^\top \hat{\beta} \pm t_{n-p}(\alpha)\, \hat{\sigma}\, \{1 + \tilde{x}^\top (X^\top X)^{-1} \tilde{x}\}^{1/2}, \tag{2}$$

where $t_\nu(\alpha)$ denotes the upper α^{th} quantile of the Student-t distribution with ν degrees of freedom. If we denote this prediction interval as $C_\alpha(Y)$, omitting the dependence on X and $\tilde{x}$, then we say that the prediction interval is *valid* in the sense that

$$\mathsf{P}\{C_\alpha(Y) \ni \tilde{Y}\} = 1 - \alpha,$$

where the probability is with respect to the joint distribution of $(Y, \tilde{Y})$ under the model (1). In other words, validity means that the actual prediction coverage of $C_\alpha(Y)$ equals the

nominal level $1-\alpha$. To summarize this result over all values of α, simultaneously, we can construct a *predictive distribution* for $\tilde{Y}$, given Y, which has a density function given by

$$\pi_Y(\tilde{y}) = \{1 + \tilde{x}^\top (X^\top X)^{-1} \tilde{x}\}^{-1/2} f_{n-p}\Big(\frac{\tilde{y} - \tilde{x}^\top \hat{\beta}}{\{1 + \tilde{x}^\top (X^\top X)^{-1} \tilde{x}\}^{1/2}} \Big), \tag{3}$$

where f_ν is the density function corresponding to a Student-t distribution with ν degrees of freedom and, again, we suppress the dependence on X and $\tilde{x}$ in the notation. Then that $100(1-\alpha)\%$ prediction interval described above is exactly the $1-\alpha$ highest predictive density interval corresponding to $\pi_Y(\tilde{y})$. We will use this predictive density primarily for visualization purposes in what follows.

3 Effects of model selection on inference

Inference and prediction based on the model (1) and the least-squares distribution theory is standard, but the actuary often will not know which of the rating variables $X_1, \ldots, X_p$ are relevant to explaining the variation in the loss Y; in other words, the actuary may want to consider which of the coefficients β_j, $j = 1, \ldots, p$, are zero. To facilitate this discussion, it will help to expand a bit on the usual notation. Rewrite the parameter β as a pair (S, β_S), where $S \subseteq \mathcal{S} := \{1, 2, \ldots, p\}$ is the model, i.e., the set of indexes, j, corresponding to non-zero β_j, and β_S is the corresponding $|S|$-vector of non-zero values. This expanded notation helps to make clear that S might also be uncertain, which is not easily reflected in (1).

Established approaches for dealing with an uncertain model, such as the Akaike information criterion (AIC, Akaike 1973) and the Bayesian information criterion (BIC, Schwartz 1978), first use the data to select a suitable model, $\hat{S}$, say, and then estimate the corresponding parameter $\beta_{\hat{S}}$ via least-squares as usual (e.g., Frees 2010, Ch. 5). An alternative to AIC and BIC is the least absolute shrinkage and selection operator (lasso, Tibshirani 1996), which attempts to simultaneously estimate the pair (S, β_S); a recent reference on lasso in the actuarial science literature is Duncan et al (2016), and a detailed summary of its many variants is given in Hastie et al (2009). While AIC, BIC, lasso, forward stepwise, etc, are simple and widely used, there are some often-overlooked concerns.

Our focus here is on the so-called *selection effect*, i.e., using the data first to select a model $\hat{S}$ and will affect the distribution theory for the least-squares estimators, which may invalidate inference and/or prediction. Here we give an example to illustrate the potentially serious problems that may arise; a different example, with a similar message, can be found in Lockhart et al (2014). As a first step in a regression analysis, one often will carry out a full F-test to determine if any of the coefficients β_j, $j \in S$, for a fixed $S \subseteq \mathcal{S}$, are statistically significant (e.g., Frees 2010, Ch. 4 and De Jong and Heller, Ch. 4). Under the null hypothesis that all the coefficients are zero, the p-value for the F-test will have a $\mathsf{Unif}(0,1)$ distribution. But, as discussed above, standard practice is to use data to help select a candidate model, say $\hat{S}$. Here we consider a choice of $\hat{S}$ based on the AIC criterion; this is easy to implement using the R function `regsubsets` provided in the `leaps` package (Lumley 2009). What

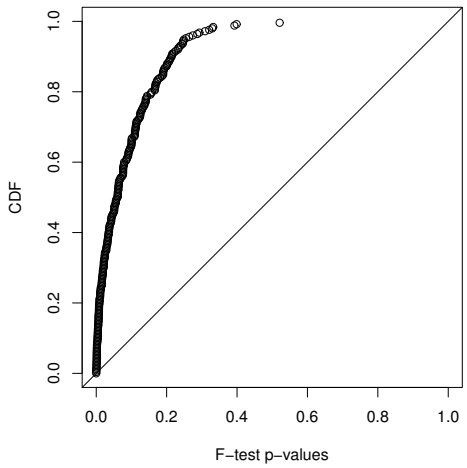


Figure 1: Plot of the distribution function of the F-test p-values, after model selection via AIC, compared to that of $\mathsf{Unif}(0,1)$, diagonal line.

happens when the F-test is applied after model selection via AIC? Is the distribution of the p-value still $\mathsf{Unif}(0,1)$?

To investigate this, we carry out a simulation study. Let the rating variables $X_1, \ldots, X_p$ be iid $\mathsf{N}(0,1)$, and $\beta_1 = \cdots = \beta_p = 0$, so that the null hypothesis is true no matter what model is selected; here we take $p = 10$ and $n = 50$. For each data set, we select a model based on the AIC criterion, carry out the F-test as usual on the selected model, and evaluate the p-value. This process is repeated for 250 data sets, and Figure 1 plots the empirical distribution function of the F-test p-value. While different repetitions may not yield selected models with the same number of predictors, and this means that the reference F distribution under the null hypothesis may be different for different repetitions, the p-values are uniformly distributed under any of the null distributions, and therefore transforming the test statistic at each repetition into a p-value allows for comparing the distribution of the classical F-test null p-values with their actual distribution post-selection. The classical theory suggests that this empirical distribution function should match that of $\mathsf{Unif}(0,1)$, the diagonal line in the plot, but clearly it does not. The p-values after selection are stochastically considerably larger than $\mathsf{Unif}(0,1)$, so the the classical F-test, applied post selection, is not valid.

The effects of model selection on inference is a serious concern, and it has become something of a "hot topic" in the statistics literature in recent years; important references include Benjamini et al (2005), Leeb and Pötscher (2005, 2006, 2008), Berk et al (2013), Efron (2014), Fithian (2015), and Taylor and Tibshirani (2015, 2017). Aside from identifying issues that arise as a result of model selection, there are important questions about what even is the inferential target post-selection. The aforementioned papers address some of these issues, and propose various corrections for the selection effect. It is beyond the scope of this paper to review the various proposals; besides, this is still a very active area of research so new developments are to be expected. However, we should mention briefly one of the general

strategies that can be used to correct for the selection effect, one that we will consider in the next section. As was made clear in Figure 1, the act of selecting a model first changes the classical distribution theory. To correct for the selection effect, one only needs to understand *how* the classical distribution theory changes. Only in rare cases can this selection-adjusted distribution theory be worked out analytically, but numerical approximations may be possible. In Section 4 we will consider an approach to adjust for the selection effect based on the *bootstrap* (Efron 1979; Klugman et al 2012).

4 Effects of model selection on prediction?

4.1 Setup and first observations

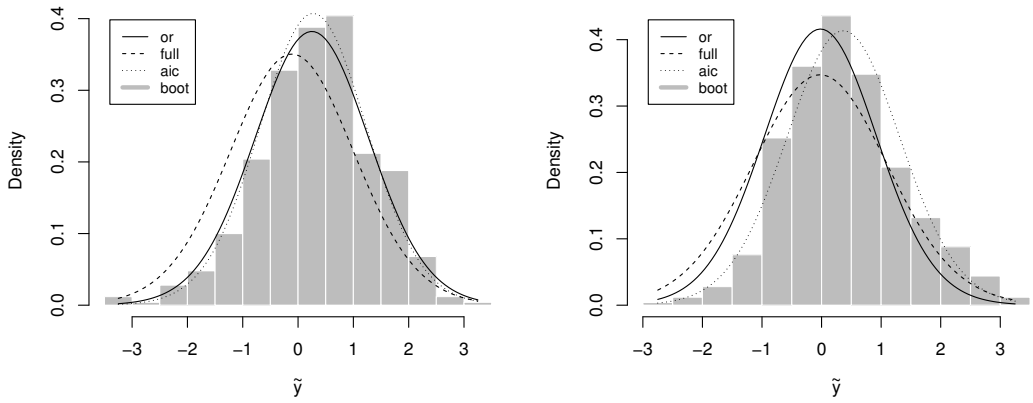
Given the apparently damaging effects that model selection can have on the validity of statistical inference, it is imperative to ask if these effects carry over to the insurer's prediction problem. That is, are actuaries safe to base their predictions on (3) when the data are used to first select a model? Despite the surge of interest in statistics on post-selection inference, as discussed above, the effect of selection on prediction has not received much attention; the only work along these lines that we are aware of is Leeb (2009), but see Section 5.

When only a subset S of the predictor variables are to be considered, then the prediction methodology described above can be modified in an obvious way. Indeed, the $100(1-\alpha)\%$ prediction interval becomes

$$\tilde{x}_S^\top \hat{\beta}_S \pm t_{n-|S|}(\alpha)\, \hat{\sigma}_S \left\{1 + \tilde{x}_S^\top (X_S^\top X_S)^{-1} \tilde{x}_S\right\}^{1/2},$$

and we can proceed to define a corresponding predictive distribution, as in (3), which we will denote by $\pi_Y(\tilde{y} \mid S)$ to highlight the dependence on the model S. If $S^\star$ is the "true" model, i.e., $\beta_j = 0$ for all $j \notin S^\star$, then all the distributional properties of the prediction interval, etc, carry over to this case. But what happens if data is used to select a model $\hat{S}$? We will investigate the effect of selection on prediction by looking at the corresponding predictive density $\pi_Y(\tilde{y} \mid \hat{S})$, which depends on data in two different ways—one is direct, just like in (3), and the other is indirect, through $\hat{S}$. For the discussion that follows, again we will focus on $\hat{S}$ chosen via AIC because this is the preferred method in prediction applications, and convenient R functions are available for doing best subset selection via AIC, e.g., `regsubsets`. In some limited experiments using other selection methods, such as BIC and lasso, we found similar results to those presented here.

To see the effect of selection on prediction, we consider three alternative predictive distributions besides the AIC predictive distribution. The first is the oracle predictive distribution, $\pi_Y(\tilde{y} \mid S^\star)$, based on knowledge of the true model $S^\star$. This is the "gold standard" predictive, ideal for comparison purposes, but, unfortunately, it is not available to the actuary in practice because he/she typically will not know $S^\star$. The second is the predictive distribution, $\pi_Y(\tilde{y} \mid \mathcal{S})$, based on the full model that includes all the rating variables. This is a simple conservative choice that may be inefficient, for a variety of reasons, compared to the other more sophisticated methods. Finally, following our discussion above, we consider a predictive



(a) AIC and oracle are similar (b) AIC and oracle are different

Figure 2: Displays of the four predictive distributions in two simulation experiments.

distribution that accounts for the possible departure from the usual least-squares distribution theory caused by selection. The exact distribution theory, accounting for selection, is not available in closed-form, but we can easily get a bootstrap approximation via resampling (e.g., Davison and Hinkley book, Sec. 6.3.3), which we denote by $\hat{\pi}_Y(\tilde{y} \mid \hat{S})$. Here and throughout this paper, we take the bootstrap sample size to be $B = 500$. Of course, among these four methods (oracle, full, AIC and boostrap-AIC), the oracle predictive distribution is the best. Prediction based on the full model should also be reasonable but its inefficiency will manifest in having a wider predictive density than the oracle. It is not clear, however, what to expect from the two AIC-based predictive distributions.

To build some intuition about the performance of the various predictive distributions, we revisit that example from Section 3. Suppose the loss variables $X_1, \ldots, X_p$, with $p = 10$, are iid $\mathsf{N}(0, 1)$, and all the β coefficients are zero. We then fit the various models and evaluate the corresponding predictive distributions based on another set $\tilde{x}$ of iid $\mathsf{N}(0, 1)$ values of the ten rating variables. This process was carried out for a number of simulated data sets and there were two distinct cases that emerged: one where the AIC-based predictive distributions were similar to that of the oracle, and one where they were different. Panels (a) and (b) of Figure 2, respectively, are representative of these two cases. Note that, in each panel, both AIC-based predictive densities have slightly smaller spread than the oracle, and the main difference between the two panels is an apparently not-so-substantial location shift of the AIC-based predictive densities away from the oracle in Panel (b).

4.2 Further investigations

A number of interesting and important questions arise from the limited results described above. In particular:

1. Do the prediction intervals derived from the AIC selection-based predictive distribution

have adequate prediction coverage?

2. If so, then how does its length compare to the oracle?
3. If not, then why, and does the bootstrap adjustment do better somehow?

In this section we carry out some further simulation studies to address these questions and, ultimately, to make a recommendation for what method practitioners ought to use based on the currently available statistical tools.

The experiments carried out above is somewhat unrealistic in the sense that the rating variables were independent and, in fact, none of them contributed to the loss variable distribution because the β coefficients were all zero. Here we consider a more realistic scenario in which there is some dependence between rating variables and there are some non-zero coefficients, with varying magnitudes.

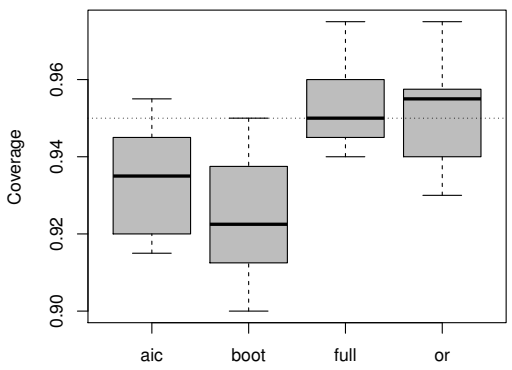
- We consider $X_1, \ldots, X_p$ to be multivariate normal, with standard normal marginals, and with first-order autoregressive correlation structure, i.e., the correlation between X_j and X_k is $0.5^{|j-k|}$, for $j, k = 1, \ldots, p$.
- In an effort to reach some conclusions independent of a fixed choice of the non-zero values of β, we consider three classes—*weak*, *moderate*, and *strong*—and then randomly sample from these classes. In particular, we first sample 3 of the 10 rating variables to include and, for each of those three, the corresponding β_j's are sampled iid from $\mathsf{N}(\mu, 1)$; the other 7 all have $\beta_j = 0$. The weak, moderate, and strong classes correspond to $\mu = 1$, $\mu = 3$, and $\mu = 5$, respectively.

For each of the weak, moderate, and strong cases, and for each "true" β sampled from the class, we simulate 200 data sets according to the model for rating variables above and (1), and evaluate 95% prediction intervals based on the various methods above. The average lengths and prediction coverage proportion of these predictive intervals can be computed for each β, and Figure 3 gives a summary of these results over 20 β's sampled from the respective classes. Throughout, we keep $n = 50$, $p = 10$, and $\sigma = 1$ fixed.

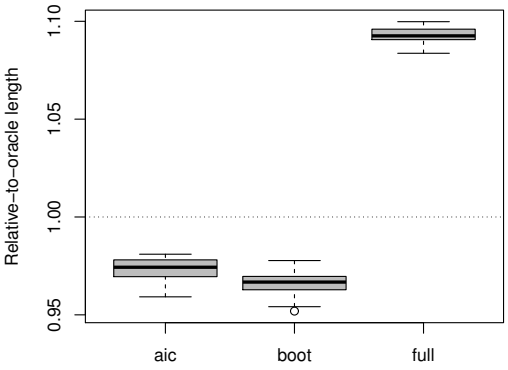
The first observation is that both the oracle and the full model produce prediction intervals with the right coverage, but the full model intervals tend to be longer, by up to 10%, confirming the claims we made previously. Second, we see that the AIC selection-based intervals are both a bit shorter than the oracle, on average, and therefore, tend to under-cover, especially the bootstrap version.

Any explanation for AIC's less-than-fully-satisfactory performance? First, it is known that AIC tends to overfit; that is, if $\hat{S}$ is the set of rating variables selected by AIC, then AIC tends to overfit in the sense that, typically, $\hat{S} \supset S^\star$ (Hurvich and Tsai 1989; Zheng and Loh 1995). It follows from Theorem 1 in Hong et al (2017) that AIC overfitting implies variance underestimation, i.e., $\hat{S} \supset S^\star$ implies $\hat{\sigma}^2_{\hat{S}} < \hat{\sigma}^2_{S^\star}$. And for relatively large n, the width of the prediction intervals is chiefly determined by these variance estimates. Indeed, for n appreciably larger than p:

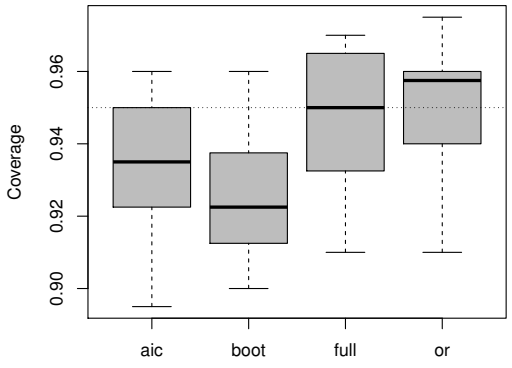
- The critical value $t_{n-|S|}$ in (2) will not differ much for $S = S^\star$ or $S = \hat{S}$.



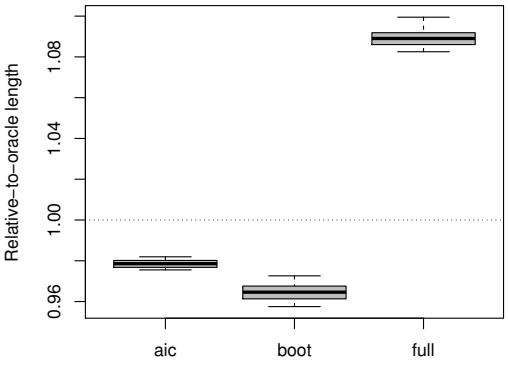
(a) Coverage proportion; weak class



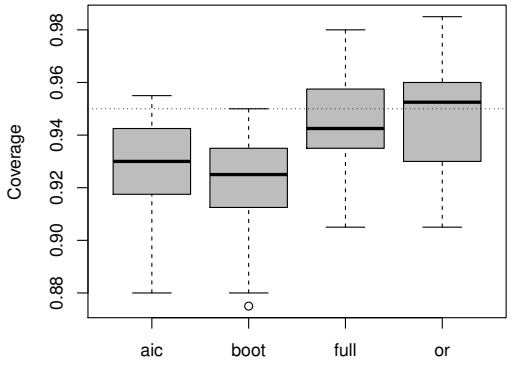
(b) Average length; weak class



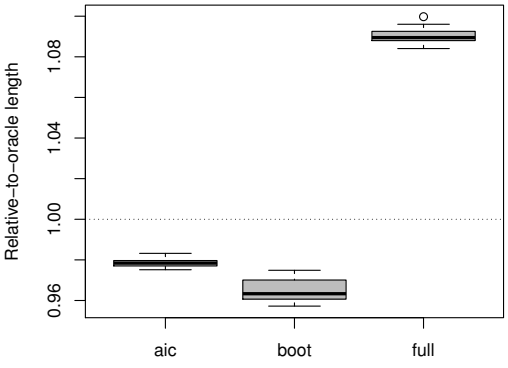
(c) Coverage proportion; moderate class



(d) Average length; moderate class



(e) Coverage proportion; strong class



(f) Average length; strong class

Figure 3: Summary of coverage proportion and average length of the various prediction intervals over 20 samples from each of the weak, moderate, and strong classes of signals.

- Since n is large, $X_S^\top X_S \approx n\Sigma_S$, where Σ_S is the corresponding sub-matrix of Σ, when the rows of X are $\mathsf{N}(0, \Sigma)$. Therefore,

$$\tilde{x}_S^\top (X_S^\top X_S)^{-1} \tilde{x}_S \approx \frac{\tilde{x}_S^\top \Sigma_S^{-1} \tilde{x}_S}{n}.$$

  Since the numerator on the right-hand side, as a function of $\tilde{x}_S \sim \mathsf{N}(0, \Sigma_S)$, is a chi-square random variable with $|S|$ degrees of freedom, it should be small compared to the n in the denominator. Therefore, the the AIC and oracle prediction intervals will not be affected by the $\{1 + \tilde{x}_S^\top (X_S^\top X_S)^{-1} \tilde{x}_S\}^{1/2}$ term either.

Therefore, any noticeable difference in the prediction intervals for AIC versus the oracle must be due to the variances $\hat{\sigma}^2_{S^\star}$ and $\hat{\sigma}^2_{\hat{S}}$. Since AIC tends to overfit, and overfitting implies variance underestimation, the shorter length and under-coverage of the AIC-based prediction intervals is to be expected. And given the under-coverage of the AIC selection-based prediction interval, a possible explanation for the bootstrap's slightly worse coverage is the dilation phenomenon described in Efron (2003).

Based on the results presented here, and keeping in mind that incorrect predictions can jeopardize an insurer's solvency, we have to recommend that practitioners use the full model to derive their predictions. This is, indeed, a conservative recommendation, but solvency is an important concern and none of the other methods seem to provide valid prediction intervals in general. Our recommendation would surely change if the problems being considered by actuarial scientists where "high-dimensional," i.e., where p is large compared to n, or if new statistical methodology for valid post-selection prediction were developed.

5 Conclusions and further questions

In this paper, we considered the effect of model selection on both inference and prediction. The material in Section 3 is mostly a review of some recent work in the statistics literature, but bringing the concerns about inference after model selection to the attention of practicing actuaries is important and one of our main goals. On the other hand, so the observations made in Section 4 are, to the best of our knowledge, new. Our conclusion is that post-selection prediction—based on either a naive application of the usual least-squares theory or a bootstrap adjustment—is unsatisfactory in the sense that the corresponding prediction intervals tend to under-cover. Therefore, we make the following recommendation:

> *Actuaries should not carry out a selection step if valid prediction is the goal, that is, they should construct their predictive distribution, intervals, etc based on the full model and the usual least-squares distribution theory.*

This is indeed a conservative recommendation since the full model might not be the most efficient one. But since prediction errors may jeopardize the insurer's solvency, we argue that conservatism is a prudent position to take. Naturally, our recommendation could potentially

change if the problem setting were different, if something other than prediction were the goal, or if new statistical methodology were developed.

The high-dimensional regression problem, where $p \gg n$, has received considerable attention in the statistics literature, but this situation is still rare in actuarial science applications. However, as new technology develops, one would expect that, eventually, the high-dimensional problem would be one that actuarial scientists would be interested in and, naturally, the question of how to make valid predictions in such cases would be relevant. When $p \gg n$, our recommendation to use the full model for prediction is no longer feasible, so some entirely new considerations would be needed.

One important question, for future research, is *how to correct for the selection effect when the goal is valid prediction?* To really answer the question, we would need to derive the form of an optimal prediction interval correction which makes the intervals asymptotically honest. This is a substantial endeavor, a focus of our ongoing work.

Acknowledgments

The authors thank the three anonymous referees for their comments and suggestions, which led to significant improvements in this article. This work is supported by the Society of Actuaries through the 2016 Individual Grant.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest.

Benjamini, Y. Yekutieli, D., Edwards, D. Shaffer, J. P., Tamhane, A. C., Westfall, P. H., and Holland, B. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters, *Journal of the American Statistical Association*, 100, 71–93.

Bignozzi, V. and Tsanakas, A. (2016). Parameter uncertainty and residual estimation risk, *Journal of Risk and Insurance*, 83, 949–978.

Berk, R. Brown, L., Buja, A. Zhuang, K. and Zhao, L. (2013). Valid post-selection inference, *Annals of Statistics*, 41, 802–837.

Cairns, A. J. G. (2000). A discussion of parameter and model uncertainty in insurance, *Insurance: Mathematics and Economics*, 27, 313–330.

Clyde, M. and George, E. I. (2004). Model uncertainty, *Statistical Science*, 19, 81–94.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*, Cambridge: Cambridge University Press.

de Jong, P. and Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*, Cambridge University Press.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B*, 57, 45–97.

Duncan, I., Loginov, M., and Ludkovski, M. (2016). Testing alternative regression frameworks for predictive modeling of healthcare costs, *North American Actuarial Journal*, 20, 65–87.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1–26.

Efron, B. (2003). Second thoughts on the boostrap, *Statistical Science*, 18, 135–140.

Efron, B. (2014). Estimation and accuracy after model selection, *Journal of the American Statistical Association*, 109, 991–1007.

Fithian, W. S. (2015). *Topics in Adaptive Inference*, Ph.D. thesis, Stanford University Department of Statistics.

Frees, E. W. (2010). *Regression Modeling with Actuarial and Financial Applications*, Cambridge: Cambridge University Press.

Frees, E. W , Derrig, R. A. and Meyers, G. (2014). *Predictive Modeling Applications in Actuarial Science, Vol. I: Predictive Modeling Techniques*, Cambridge: Cambridge University Press.

Hartman, B. and Groendyke, C. (2013). Model selection and averaging in financial risk management, *North American Actuarial Journal*, 17, 216–228.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Essentials of Statistical Learning*, second edition. Springer.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science*, 14, 382–417.

Hong, L., Kuffner, T. A. and Martin, R. (2017). On overfitting and post-selection uncertainty assessments, *Biometrika*, to appear, `http://arxiv.org/abs/1712.02379`.

Huang, S., Hartman, B. and Brazauskas, V. (2016). Model selection and averaging of health costs in episode treatment groups, *ASTIN Bulletin*, 47(1), 153–167.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* 76(2), 297–307.

Kabaila, P. (1995). The effect of model selection on confidence regions and prediction regions, *Econometric Theory*, 537–549.

Klugman, S.A., Panjer, H.H. and Willmont, G.E. (2012). *Loss Models: From Data To Decisions*, Fourth Edition, Hoboken, NJ: Wiley.

Leeb, H. (2009). Conditional predictive inference post model selection, *Annals of Statistics*, 37, 2838–2876.

Leeb, H. and Pötscher B.M. (2005). Model selection and inference: facts and fiction, *Econometric Theory*, 21, 21–59.

Leeb, H. and Pötscher B.M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics*, 34, 2554–2591.

Leeb, H. and Pötscher B.M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24, 338–376.

Lockhart, R. , Taylor, J. , Tibshirani, R. J. , and Tibshirani, R. (2014). A significance test for the lasso, *Annals of Statistics* 42 (2), 413–468.

Lumley, T. (2009). leaps: regression subset selection. R package version 2.9; using Fortran code by Alan Miller. `http://CRAN.R-project.org/package=leaps`.

McCullogh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, second edition, Boca Raton: Chapman & Hall/CRC.

Peters, G. W., Shevchenko, P. V., and Wüthrich, M. V. (2008). Model uncertainty in claims reserving within Tweedie's compound Poission models, *ASTIN Bulletin*, 39, 1–33.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

Taylor J. and Tibshirani, R.J. (2015). Statistical learning and selective inference, *PNAS*, 112, 7629–7634.

Taylor J. and Tibshirani, R.J. (2017). Post-selection inference for l_1-penalized likelihood models, *Canadian Journal of Statistics*, `doi/10.1002/cjs.11313`.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

Venter, G. and Sahasrabuddhe, R. (2016). A note on parameter risk, *Variance*, 9, 54–63.

Werner, G. and Modlin, C. (2010). *Basic Ratemaking*, Arlington: Casualty Actuarial Society.

Zheng, X. and Loh, W.-Y. (1995). Consistent variable selection in linear models. *Journal of the American Statistical Association* 90(429), 151–156.