Article

Resampling Plans and the Estimation of Prediction Error

Bradley Efron [1,†]

[1] Stanford University; brad@stat.stanford.edu
† Current address: Department of Statistics, Sequoia Hall, Mail Code 4065, Stanford, CA 94305

Abstract: This article was prepared for the special issue on *Resampling methods for statistical inference of the 2020s*. Modern algorithms such as random forests and deep learning are automatic machines for producing prediction rules from training data. Resampling plans have been the key technology for evaluating a rule's prediction accuracy. After a careful description of the measurement of prediction error the article discusses the advantages and disadvantages of the principal methods: cross-validation, the nonparametric bootstrap, covariance penalties (Mallows' C_p and the Akaike Information Criterion), and conformal inference. The emphasis is on a broad overview of a large subject, featuring examples, simulations, and a minimum of technical detail.

Keywords: cross-validation, C_p, AIC, Q-class, conformal inference, random forests, bagging

1. Introduction

Modern prediction algorithms such as random forests and deep learning use *training sets*, often very large ones, to produce rules for predicting new responses from a set of available predictors. A second question — right after "How should the prediction rule be constructed?" — is "How accurate are the rule's predictions?" Resampling methods have played a central role in the answer. This paper is intended to provide an overview of what are actually several different answers, while trying to keep technical complications to a minimum.

This is a special issue of *STATS* devoted to resampling, and before beginning on prediction rules it seems worthwhile to say something about the general effect of resampling methods on statistics and statisticians. Table 1 shows the *law school data* [1], a small data set but one not completely atypical of its time. The table reports average scores of the 1973 entering class at 15 law schools on two criteria: undergraduate grade point average (GPA) and result on the "LSAT", a national achievement test. The observed Pearson correlation coefficient between GPA and LSAT score is

$$\hat{\rho} = 0.776. \tag{1}$$

How accurate is $\hat{\rho}$?

Suppose that Dr. Jones, a 1940s statistician, was given the data in Table 1 and asked to attach a standard error to $\hat{\rho} = 0.776$; let's say a nonparametric standard error since a plot of (LSAT,GPA) looks definitely non-normal. At his disposal is the *nonparametric delta method*, which gives a first-order Taylor series approximation formula for $\text{se}(\hat{\rho})$. For the Pearson correlation coefficient this turns out to be

$$\widehat{\text{se}}(\hat{\rho}) = \left\{ \frac{\hat{\rho}^2}{4n} \left[\frac{\hat{\mu}_{40}}{\hat{\mu}_{20}^2} + \frac{\hat{\mu}_{04}}{\hat{\mu}_{02}^2} + \frac{2\hat{\mu}_{22}}{\hat{\mu}_{20}\hat{\mu}_{02}} + \frac{4\hat{\mu}_{22}}{\hat{\mu}_{11}^2} - \frac{4\hat{\mu}_{31}}{\hat{\mu}_{11}\hat{\mu}_{20}} - \frac{4\hat{\mu}_{13}}{\hat{\mu}_{11}\hat{\mu}_{02}} \right] \right\}^{1/2}, \tag{2}$$

where $n = 15$ and $\hat{\mu}_{ij}$ is the mean of $(\text{GPA} - \overline{\text{GPA}})^i(\text{LSAT} - \overline{\text{LSAT}})^j$, the bars indicating averages.

Citation: Efron, B. Resampling Plans and the Estimation of Prediction Error. *Stats* **2021**, *1*, 1–26. https://doi.org/

Received:
Accepted:
Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Table 1. Average scores for admitees to 15 American law schools, 1973. GPA is undergraduate grade point average, LSAT "law boards" score. Pearson correlation coefficient between GPA and LSAT is 0.776.

GPA	LSAT
3.39	576
3.30	635
2.81	558
3.03	578
3.44	666
3.07	580
3.00	555
3.43	661
3.36	651
3.13	605
3.12	653
2.74	575
2.76	545
2.88	572
2.96	594

Jones either looks up or derives (2), evaluates the six terms on his mechanical calculator, and reports

$$\widehat{\mathrm{se}}(\hat{\rho}) = 0.124, \tag{3}$$

after which he goes home with the feeling of a day well spent.

Jones' daughter, a 1960s statistician, has a much easier go of it. Now she doesn't have to look up or derive formula (2). A more general resampling algorithm, the Tukey–Quenouille *jackknife* is available, and can be almost instantly evaluated on her university's mainframe computer. It gives her the answer

$$\widehat{\mathrm{se}}(\hat{\rho}) = 0.143. \tag{4}$$

Dr. Jones is envious of his daughter:

1. She doesn't need to spend her time deriving arduous formulas like (2).
2. She isn't restricted to traditional estimates like $\hat{\rho}$ that have closed-form Taylor series expansions.
3. Her university's mainframe computer is a million times faster than his old Marchant calculator (though it *is* across the campus rather than on her desk).

If now, 60 years later, the Jones family is still in the statistics business they'll have even more reason to be grateful for resampling methods. Faster, cheaper, and more convenient computation combined with more aggressive methodology have pushed the purview of resampling applications beyond the assignment of standard errors.

Figure 1 shows 2000 nonparametric bootstrap replications $\hat{\rho}^*$ from the law school data.[1] Their empirical standard deviation is the nonparametric bootstrap estimate of standard error for $\hat{\rho}$,

$$\widehat{\mathrm{se}}(\hat{\rho}) = 0.138. \tag{5}$$

[1] Each $\hat{\rho}^*$ is the correlation from a bootstrapped data matrix obtained by resampling the 15 rows of the 15×2 matrix in Table 1 15 times with replacement. See chapter 11 of [2].

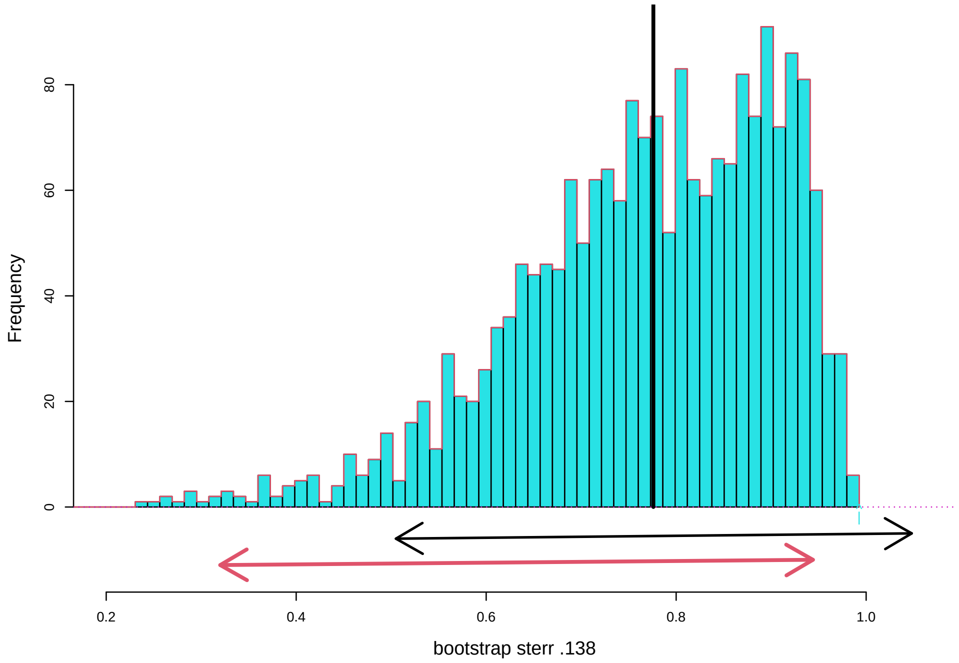

Figure 1. 2000 nonparametric bootstraps, law school correlation. 95% confidence limits: bootstrap (red), standard (black).

Two thousand is 10 times too many replications needed for a standard error, but it isn't too many for a bootstrap confidence interval. The arrowed segments in Figure 1 compare the standard approximate 95% confidence limit

$$\rho \in \hat{\rho} \pm 1.96\widehat{\text{se}} = [0.505, 1.048], \tag{6}$$

$\widehat{\text{se}} = 0.138$, with the nonparametric bootstrap interval[2]

$$\rho \in [0.320, 0.944]. \tag{7}$$

The standard method does better if we first make Fisher's z transformation

$$\hat{\xi} = \frac{1}{2} \log \frac{1+\hat{\rho}}{1-\hat{\rho}}, \tag{8}$$

compute the standard interval on the ξ scale, and transform the endpoints back to the ρ scale. This gives 0.95 interval

$$\rho \in [0.275, 0.946], \tag{9}$$

not so different from the bootstrap interval (7), and at least not having its upper limit above 1.00!

This is the kind of trick Dr. Jones would have known. Resampling, here in the form of the bca algorithm, automates devices like (8) without requiring Fisher-level insight for each new application.

If there was a statistics' "word of the year" it would be two words: *deep learning*. This is one of a suite of prediction algorithms that input data sets, often quite massive ones, and output prediction rules. Others in the suite include *random forests*, *support vector machines*, and *gradient boosting*.

[2] This is the *bca interval*, constructed using program `bcajack` from the CRAN package `bcaboot` [3]. Chapter 11 of [2] shows why bca's "second-order corrections" (here very large) improve on the standard method.

Having produced a prediction rule, it is natural to wonder how accurately it will predict future cases, our subject in what follows:

- Section 2 gives a careful definition of the prediction problem, and describes a class of loss functions (the "Q class") that apply to discrete as well as continuous response variables.
- Section 3 concerns nonparametric estimates of prediction loss (cross-validation and the bootstrap "632 rule") as well as Breiman's *out-of-bag* error estimates for random forests.
- *Covariance penalties*, including Mallows' C_p and the Akaike Information Criterion, are parametric methods discussed in Section 4 along with the related concept of *degrees of freedom*.
- Section 5 briefly discusses *conformal inference*, the most recent addition to the resampling catalog of prediction error assessments.

2. The prediction problem

Statements of the prediction problem are often framed as follows:

- A data set $\boldsymbol{d}$ of n pairs is observed,

$$\boldsymbol{d} = \{(x_i, y_i),\ i = 1, 2, \ldots, n\}, \tag{10}$$

  where the x_i are p-dimensional predictor vectors and the y_i are one-dimensional responses.
- The (x, y) pairs are assumed to be independent and identically distributed draws from an unknown $(p+1)$-dimensional distribution F,

$$(x_i, y_i) \overset{\text{iid}}{\sim} F, \qquad i = 1, 2, \ldots, n. \tag{11}$$

- Using some algorithm $\mathcal{A}$, the statistician constructs a prediction rule $f(x, \boldsymbol{d})$ that provides a prediction $\hat{\mu}(x)$,

$$\hat{\mu}(x) = f(x, \boldsymbol{d}) \tag{12}$$

  for any vector x in the space of possible predictors.
- A new pair (x, y) is independently drawn from F,

$$(x, y) \sim F \qquad \text{independent of } \boldsymbol{d}, \tag{13}$$

  but with only x observable.
- The statistician predicts the unseen y by $\hat{\mu}(x) = f(x, \boldsymbol{d})$ and wishes to assess prediction error. Later the prediction error will turn out to directly relate to the estimation error of $\hat{\mu}(x)$ for the true conditional expectation of y given x,

$$\mu(x) = E_F\{y \mid x\}. \tag{14}$$

- Prediction error is assessed as the expectation of loss under distribution F,

$$\mathrm{Err}^{(u)} = E_F\{Q(y, \hat{\mu}(x))\}, \tag{15}$$

  for a given loss function such as squared error: $Q(y, \hat{\mu}) = (y - \hat{\mu})^2$.

Here E_F indicates expectation over the random choice of all $p+1$ pairs (x_i, y_i) and (x, y) in (11) and (13). The u in $\mathrm{Err}^{(u)}$ reflects the unconditional definition of error in (15). The resampling algorithms we will describe calculate an estimate of $\mathrm{Err}^{(u)}$ from the observed data. (One might hope for a more conditional error estimate, say one applying to the observed set of predictors $\boldsymbol{x}$, a point discussed in what follows.)

Naturally the primary concern within the prediction community has been with the choice of the algorithm $\mathcal{A}$ that produces the rule $\hat{\mu}(x) = f(x, \boldsymbol{d})$. Elaborate computer-intensive algorithms such as *random forests* and *deep learning* have achieved star status, even in the popular press. Here however, the “prediction problem” will focus on the estimation of prediction error. To a large extent the prediction problem has been a contest of competing resampling methods, as discussed in the next three sections.

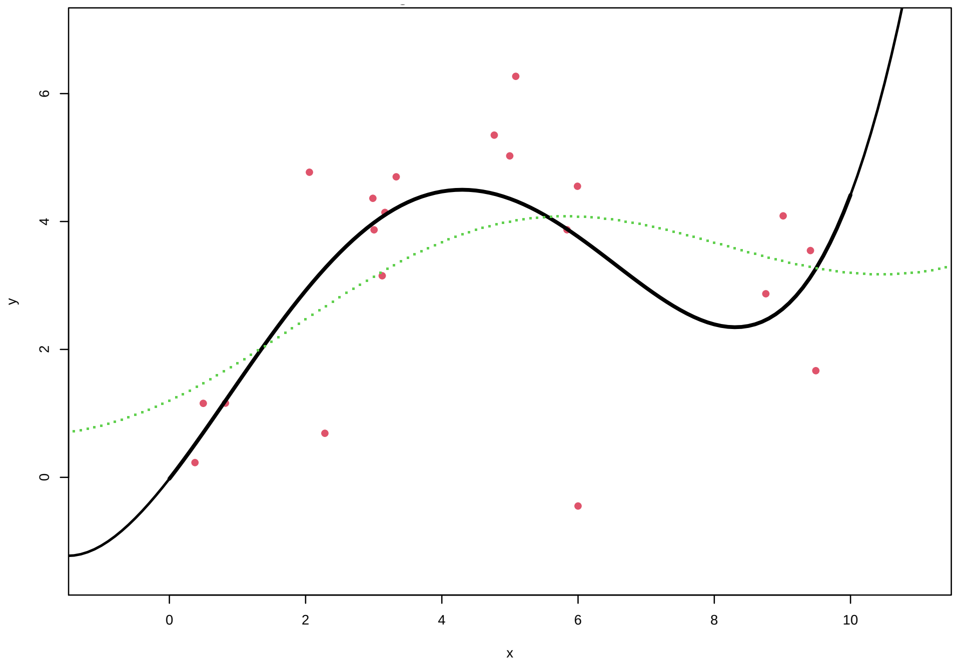

Figure 2. Points (x_i, y_i), $i = 1, 2, \ldots, 20$, and fitted 4th-degree polynomial curve; light dotted curve is true mean.

Figure 2 illustrates a simple example: $n = 20$ pairs (x_i, y_i) have been observed, in this case with x real. A fourth-degree polynomial $f(x, \boldsymbol{d})$ has been fit by ordinary least squares applied to $\boldsymbol{d} = \{(x_i, y_i),\ i = 1, 2, \ldots, 20\}$ with the heavy curve tracing out $\hat{\mu}(x) = f(x, \boldsymbol{d})$.

In the usual OLS notation, we've observed $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ from the notional model

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}, \tag{16}$$

where $\boldsymbol{X}$ is the 20×5 matrix with ith row $(1, x_i, x_i^2, x_i^3, x_i^4)$, β the unknown 5-dimensional vector of regression coefficients, and $\boldsymbol{\epsilon}$ a vector of 20 uncorrelated errors having mean 0 and variance σ^2,

$$\epsilon_i \sim (0, \sigma^2) \qquad \text{for } i = 1, 2, \ldots, n. \tag{17}$$

The fitted curve $\hat{\mu}(x) = f(x, \boldsymbol{d})$ is given by

$$\hat{\mu}(x) = X(x)^\top \hat{\beta}, \tag{18}$$

for $X(x)^\top = (1, x_i, x_i^2, x_i^3, x_i^4)$ and $\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ (the OLS estimate), this being algorithm $\mathcal{A}$.

The *apparent error*, what will be called err in what follows, is

$$\text{err} = \sum_{i=1}^{n} \frac{Q(y_i, \hat{\mu}_i)}{n} \qquad (\hat{\mu}_i = f(x_i, \boldsymbol{d})) \tag{19}$$

which equals 1.99 for $Q(y, \hat{\mu}) = (y - \hat{\mu})^2$. The usual unbiased estimate for the noise parameter σ^2, not needed here, modifies the denominator in (19) to take account of fitting a 5-vector $\hat{\beta}$ to $\boldsymbol{d}$,

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{Q(y_i, \hat{\mu}_i)}{(n-5)} = 2.65. \tag{20}$$

Dividing the sum of squared errors by $n - p$ rather than n can be thought of as a classical prediction error adjustment; err usually underestimates future prediction error since the $\hat{\mu}_i$ have been chosen to fit the observations y_i.

Because this is a simulation, we know the true function $\mu(x)$ (14), the light dotted curve in Figure 2:

$$\mu(x) = 2 + 0.2x + \cos\left(\frac{x-5}{2}\right); \tag{21}$$

the points y_i were generated with normal errors, variance 2,

$$y_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu(x_i), 2), \qquad i = 1, 2, \ldots, n. \tag{22}$$

Given model (22), we can calculate the true prediction error for estimate $\hat{\mu}(x)$ (18). If $y^* \sim \mathcal{N}(\mu(x), 2)$ is a new observation, independent of the original data $\boldsymbol{d}$ which gave $\hat{\mu}(x)$, then the true prediction error Err_x is

$$\text{Err}_x = E_*(y^* - \mu(x))^2 = 2 + (\hat{\mu}(x) - \mu(x))^2, \tag{23}$$

the notation E_* indicating expectation over y^*. Let Err denote the average true error over the $n = 20$ observed values x_i,

$$\text{Err} = \frac{1}{n}\sum_{i=1}^{n} \text{Err}_{x_i} = 2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}(x_i) - \mu(x_i))^2, \tag{24}$$

equaling 2.40 in this case. Err figures prominently in the succeeding sections. Here it exceeds the apparent error err $= 1.99$ (19) by 21%. (Err is not the same as $\text{Err}^{(u)}$ (15).)

Prediction algorithms are often, perhaps most often, applied to situations where the responses are dichotomous, $y_i = 1$ or 0; that is, they are *Bernoulli* random variables, binomials of sample size 1 each,

$$y_i \overset{\text{ind}}{\sim} \text{Bi}(1, \mu(x_i)) \qquad \text{for } i = 1, 2, \ldots, n. \tag{25}$$

Here $\mu(x)$ is the probability that $y = 1$ given prediction vector x,

$$\mu(x) = \Pr\{y = 1 \mid x\}. \tag{26}$$

The probability model F in (11) can be thought of in two steps: as first selecting x according to some p-dimensional distribution G and then "flipping a biased coin" to generate $y \sim \text{Bi}(1, \mu(x))$.

Squared error isn't appropriate for dichotomous data. Two loss (or "error") functions $Q(\mu, \hat{\mu})$ are in common use for measuring the discrepancy between μ and $\hat{\mu}$, the true and estimated probability that $y = 1$ in (25). The first is *counting error*,

$$Q(\mu, \hat{\mu}) = \begin{cases} 0 & \text{if } \mu \text{ and } \hat{\mu} \text{ are on same side of } 1/2 \\ 2 \cdot \left|\mu - \frac{1}{2}\right| & \text{if not.} \end{cases} \tag{27}$$

For $y = 0$ or 1, $Q(y, \hat{\mu})$ equals 0 or 1 if y and $\hat{\mu}$ are on the same or different sides of $1/2$.

The second error function is *binomial deviance* (or twice the Kullback–Leibler divergence),

$$Q(\mu, \hat{\mu}) = 2\left\{\mu \log\left(\frac{\mu}{\hat{\mu}}\right) + (1-\mu)\log\left(\frac{1-\mu}{1-\hat{\mu}}\right)\right\}. \tag{28}$$

Binomial deviance plays a preferred role in maximum likelihood estimation. Suppose that $\boldsymbol{\mu}(\beta)$ is a p-parameter family for the true vector of means in model (25),

$$\boldsymbol{\mu}(\beta) = (\mu(x_1, \beta), \mu(x_2, \beta), \ldots, \mu(x_n, \beta))^\top \qquad (\beta \in \Omega \subset \mathcal{R}^p). \tag{29}$$

Then the maximum likelihood estimate (MLE) $\hat{\beta}$ is the minimizer of the average binomial deviance (28) between $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ and $\boldsymbol{\mu}(\beta)$,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} Q(y_i, \mu(x_i, \beta)) \right\}; \tag{30}$$

see Chapter 8 of [2]. Most of the numerical examples in the following sections are based on binomial deviance (30).[3]

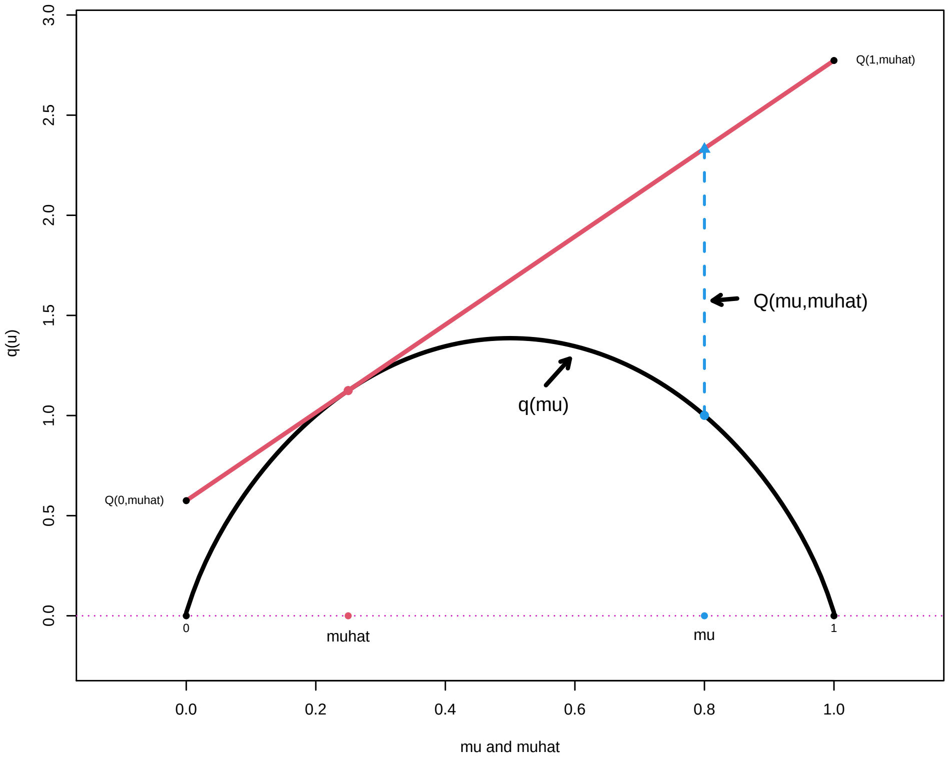

Figure 3. The Q class of error measures.

Squared error, counting error, and binomial deviance are all members of the Q-*class*, a general construction illustrated in Figure 3 [4, Sect. 3]. The construction begins with some concave function $q(\mu)$; for the dichotomous cases considered here, $\mu \in [0, 1]$ with $q(0) = q(1) = 0$. The error $Q(\mu, \hat{\mu})$ between a true value μ and an estimate $\hat{\mu}$ is defined by the illustrated tangency calculation:

$$Q(\mu, \hat{\mu}) = q(\hat{\mu}) + \dot{q}(\hat{\mu})(\mu - \hat{\mu}) - q(\mu), \tag{31}$$

$\dot{q}(\mu) = dq(\mu)/d\mu$ (equivalent to the "Bregman divergence" [5]).

The entropy function

$$q(\mu) = -2\{\mu \log \mu + (1 - \mu) \log(1 - \mu)\} \tag{32}$$

makes $Q(\mu, \hat{\mu})$ equal binomial deviance. Two other common choices are $q(\mu) = \min\{\mu, 1 - \mu\}$ for counting error and $q(\mu) = \mu(1 - \mu)$ for squared error.

[3] If $\hat{\mu}$ equals 0 or 1 then (30) is infinite. To avoid infinities, our numerical examples truncate $\hat{\mu}$ to $[0.005, 0.995]$.

Working within the Q-class (31), it is easy to express the *true error* of prediction $\hat{\mu}(x)$ at predictor value x where the true mean is $\mu(x) = E_F\{y \mid x\}$. Letting y^* be an independent realization from the distribution of y given x, the true error at x is, by definition,

$$\operatorname{Err}(\mu(x), \hat{\mu}(x)) = E_*\{Q(y^*, \hat{\mu}(x))\}, \tag{33}$$

only y^* being random in the expectation.

Lemma 1. *The true error at x (33) is*

$$\operatorname{Err}(\mu(x), \hat{\mu}(x)) = Q(\mu(x), \hat{\mu}(x)) + q(\mu(x)) - c(x), \tag{34}$$

with $c(x) = 0$ in the dichotomous case.

Proof. From definition (31) of the Q-class,

$$\begin{aligned} E_*\{Q(y^*, \hat{\mu}(x))\} &= E_*\{q(\hat{\mu}(x)) + \dot{q}(\hat{\mu}(x))(y^* - \hat{\mu}(x)) - q(y^*)\} \\ &= q(\hat{\mu}(x)) + \dot{q}(\hat{\mu}(x))(\mu(x) - \hat{\mu}(x)) - E_* q(y^*) \\ &= Q(\mu(x), \hat{\mu}(x)) + q(\mu(x)) - E_* q(y^*), \end{aligned} \tag{35}$$

giving (34) with $c(x) = E_* q(y^*)$. In the dichotomous case $q(y^*) = 0$ for y^* equal 0 or 1, so $c(x) = 0$. □

To simplify notation, let $\mu_i = \mu(x_i)$ and $\hat{\mu}_i = \hat{\mu}(x_i)$, with $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_n)^\top$ and $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_n)^\top$. The *average true error* $\operatorname{Err}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ is defined to be

$$\operatorname{Err}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \frac{1}{n} \sum_{i=1}^{n} \operatorname{Err}(\mu_i, \hat{\mu}_i) = \frac{1}{n} \sum_{i=1}^{n} [Q(\mu_i, \hat{\mu}_i) + q(\mu_i) - c(x_i)]. \tag{36}$$

It is "true" in the desirable sense of applying to the given prediction rule $\hat{\boldsymbol{\mu}}$. If we average $\operatorname{Err}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ over the random choice of $\boldsymbol{d}$ in (11), we get the less desirable unconditional error $\operatorname{Err}^{(u)}$ (15).

$\operatorname{Err}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ is minimized by $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}$,

$$\operatorname{Err}(\boldsymbol{\mu}, \boldsymbol{\mu}) = \frac{1}{n} \sum_{i=1}^{n} (q(\mu_i) - c(x_i)). \tag{37}$$

Subtraction from (36) gives

$$\operatorname{Err}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) - \operatorname{Err}(\boldsymbol{\mu}, \boldsymbol{\mu}) = \frac{1}{n} \sum_{i=1}^{n} Q(\mu_i, \hat{\mu}_i). \tag{38}$$

This is an exact analogue of the familiar squared error relationship. Suppose $\boldsymbol{y}$ and $\boldsymbol{y}^*$ are independently $\mathcal{N}(\mu, \sigma^2 I)$ vectors, and that $\boldsymbol{y}$ produces an estimate $\hat{\boldsymbol{\mu}}$. Then

$$\frac{1}{n} E_* \|\boldsymbol{y}^* - \hat{\boldsymbol{\mu}}\|^2 - \frac{1}{n} E_* \|\boldsymbol{y}^* - \boldsymbol{\mu}\|^2 = \frac{1}{n} \sum_{i=1}^{n} (\mu_i - \hat{\mu}_i)^2, \tag{39}$$

which is (38) if $Q(\mu, \hat{\mu}) = (\mu - \hat{\mu})^2$.

At a given value of x, say x_0, a prediction $\hat{\mu}(x_0)$ can also be thought of as an estimate of $\mu(x_0) = E_F\{y_0 \mid x_0\}$. Lemma 1 shows that the optimum choice of a prediction rule $\hat{\mu}(x_0) = f(x_0, \boldsymbol{d})$ is also the optimum choice of an estimation rule: for any rule $f(x, \boldsymbol{d})$,

$$E_F\{\operatorname{Err}(\mu(x_0), \hat{\mu}(x_0))\} = E_F\{Q(\mu(x_0), \hat{\mu}(x_0))\} + q(\mu(x_0) - c(x_0)) \tag{40}$$

(E_F as in (11)), so that the rule that minimizes the expected prediction error also minimizes the expected estimation error $E_F\{Q(\mu(x_0), \hat{\mu}(x_0))\}$. *Predicting y and estimating its mean μ are equivalent tasks within the Q-class.*

3. Cross-validation and its bootstrap competitors

Resampling methods base their inferences on recomputations of the statistic of interest derived from systematic modifications of the original sample. This isn't a very precise definition but it can't be if we want to cover the range of methods used in estimating prediction error. There are intriguing differences among the methods concerning just what the modifications are and how the inferences are made, as discussed in this and the next two sections.

Cross-validation has a good claim to being the first resampling method. The original idea was to randomly split the sample $\boldsymbol{d}$ into two halves, the *training* and *test* sets, $\boldsymbol{d}_{\text{train}}$ and $\boldsymbol{d}_{\text{test}}$. A prediction model is developed using only $\boldsymbol{d}_{\text{train}}$, and then validated by its performance on $\boldsymbol{d}_{\text{test}}$. Even if we cheated in the training phase, say by throwing out "bad" points, etc., the validation phase guarantees an honest estimate of prediction error.

One drawback is that inferences based on $n/2$ data points are likely to be less accurate than those based on all n, a concern if we are trying to accurately assess prediction error. "One-at-a-time" cross-validation almost eliminates this defect: let $\boldsymbol{d}_{(i)}$ be data set (10) with point (x_i, y_i) removed, and define[4]

$$\hat{\mu}_{(i)} = f(x_i, \boldsymbol{d}_{(i)}), \tag{41}$$

the prediction for case i based on x_i using the rule $f(x_i, \boldsymbol{d}_{(i)})$, constructed using only the data in $\boldsymbol{d}_{(i)}$. The cross-validation estimate of prediction error is then

$$\widehat{\text{Err}}_{\text{cv}} = \frac{1}{n} \sum_{i=1}^{n} Q(y_i, \hat{\mu}_{(i)}). \tag{42}$$

Because y_i is not involved in $\hat{\mu}_{(i)}$, overfitting is no longer a concern. Under the independent-draws model (11), $\widehat{\text{Err}}_{\text{cv}}$ is a nearly unbiased estimate of $\text{Err}^{(u)}$ (15) ("nearly" because it applies to samples of size $n-1$ rather than n).

The little example in Figure 2 has $n = 20$ points (x_i, y_i). Applying (42) with $Q(y, \hat{\mu}) = (y - \hat{\mu})^2$ gave $\widehat{\text{Err}}_{\text{cv}} = 3.59$, compared with apparent error $\text{err} = 1.99$ (19) and true error $\text{Err} = 2.40$ (24). This not-very-impressive result has much to do with the small sample size resulting in large differences between the original estimates $\hat{\mu}_i$ and their cross-validated counterparts $\hat{\mu}_{(i)}$. The single data point at $x_i = 6.0$ accounted for 40% of $\widehat{\text{Err}}_{\text{cv}}$.

Figure 4 concerns a larger data set that will serve as a basis for simulations comparing cross-validation with its competitors: 200 transplant recipients were followed to investigate the subsequent occurrence of anemia; 138 did develop anemia (coded as $y_i = 1$) while 62 did not ($y_i = 0$). The goal of the study was to predict y from x, a vector of $p = 17$ baseline variables[5] (including the intercept term).

A standard logistic regression analysis gave estimated values of the anemia probability $\Pr\{y_i = 1 \mid x_i\}$ for $i = 1, 2, \ldots, n = 200$ that we will denote as

$$\boldsymbol{\mu}^{\dagger} = (\mu_1^{\dagger}, \mu_2^{\dagger}, \ldots, \mu_{200}^{\dagger}). \tag{43}$$

Figure 4 shows a histogram of the 200 $\mu_i^{\dagger}$ values. Here we will use $\boldsymbol{\mu}^{\dagger}$ as the "ground truth" for a simulation study (rather than analyzing the transplant study itself). The $\mu_i^{\dagger}$ will play the role of the true mean $\mu(x_i)$ in Lemma 1 (34), enabling us to calculate true errors for the various prediction error estimates.

A 200×100 matrix $\boldsymbol{Y}$ of dichotomous responses y_{ij} was generated as independent Bernoulli variables (that is, binomials of sample size 1),

[4] This assumes that we know how to apply the construction rule $\mathcal{A}$ to subsets of size $n-1$.

[5] The predictor variables were body mass index, sex, race, patient and donor age, four measures of matching between patient and donor, three baseline medicine indicators, and four baseline general health measures.

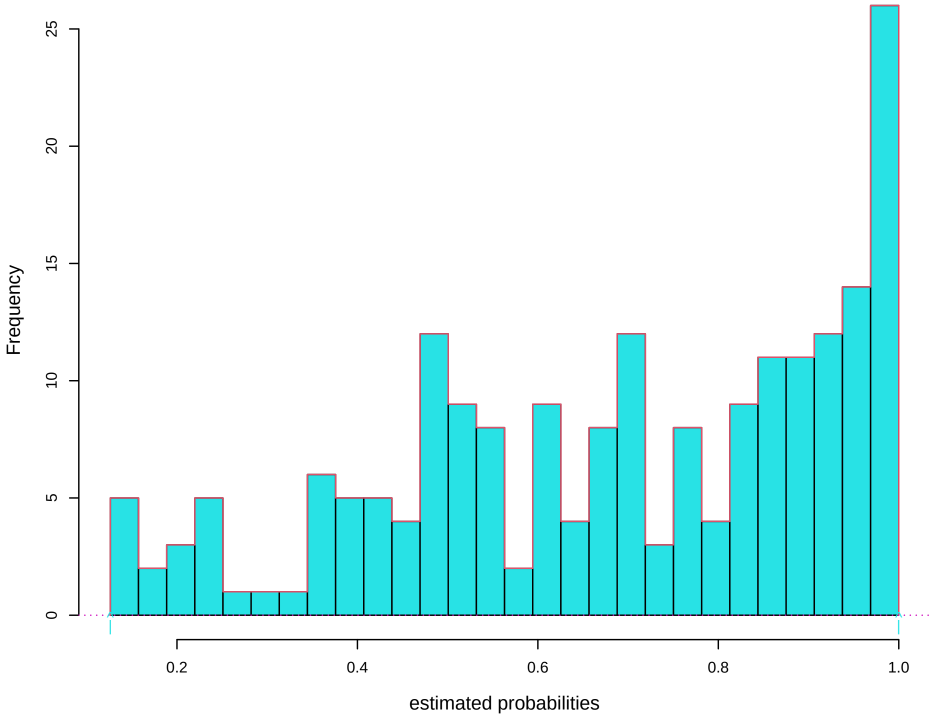

Figure 4. Logistic regression estimated anemia probabilities for the 200 transplant patients.

$$y_{ij} \overset{\text{ind}}{\sim} \text{Bi}(1, \mu_i^\dagger) \tag{44}$$

for $i = 1, 2, \ldots, 200$ and $j = 1, 2, \ldots, 100$. The jth column of $\boldsymbol{Y}$,

$$\boldsymbol{y}_j = (y_{1j}, y_{2j}, \ldots, y_{200j})^\top \tag{45}$$

is a simulated binomial response vector (25) having true mean vector $\boldsymbol{\mu}^\dagger$. $\boldsymbol{Y}$ provides 100 such response vectors.

For each one, a logistic regression was run,

$$\texttt{glm}(\boldsymbol{y}_j \sim \boldsymbol{X}, \texttt{binomial}), \tag{46}$$

in the language R, with $\boldsymbol{X}$ the 200×17 matrix of predictors from the transplant study. Cross-validation[6] gave an estimate of prediction error for the jth simulation,

$$\widehat{\text{Err}}_{\text{cv}}(j) = \frac{1}{n} \sum_{i=1}^{200} Q(y_{ij}, \hat{\mu}_{(i)j}), \tag{47}$$

while (36) gave true error

$$\text{Err}(j) = \frac{1}{n} \sum_{i=1}^{200} \left[Q(\mu_i^\dagger, \hat{\mu}_{ij}) + q(\mu_i^\dagger) \right], \tag{48}$$

where $\hat{\mu}_{ij}$ was the estimated mean from (46).

In terms of mean±standard deviation the 100 simulations gave

$$\text{Err}_{\text{cv}} \sim 1.16 \pm 0.14 \quad \text{and} \quad \text{Err} \sim 1.07 \pm 0.12. \tag{49}$$

[6] "10-at-a-time" cross-validation rather than one-at-a-time: the 200 (x,y) pairs were randomly split into 20 groups of 10 each; each group was removed from the prediction set in turn and its 10 estimates obtained by logistic regression based on the other 190.

$\mathrm{Err_{cv}}$ averaged 8% more than Err (a couple percent of which came from having sample size 190 rather than 200). The standard deviation of $\mathrm{Err_{cv}}$ isn't much bigger than the standard deviation of Err, which might suggest that $\mathrm{Err_{cv}}$ was tracking Err as it varied across the simulations.

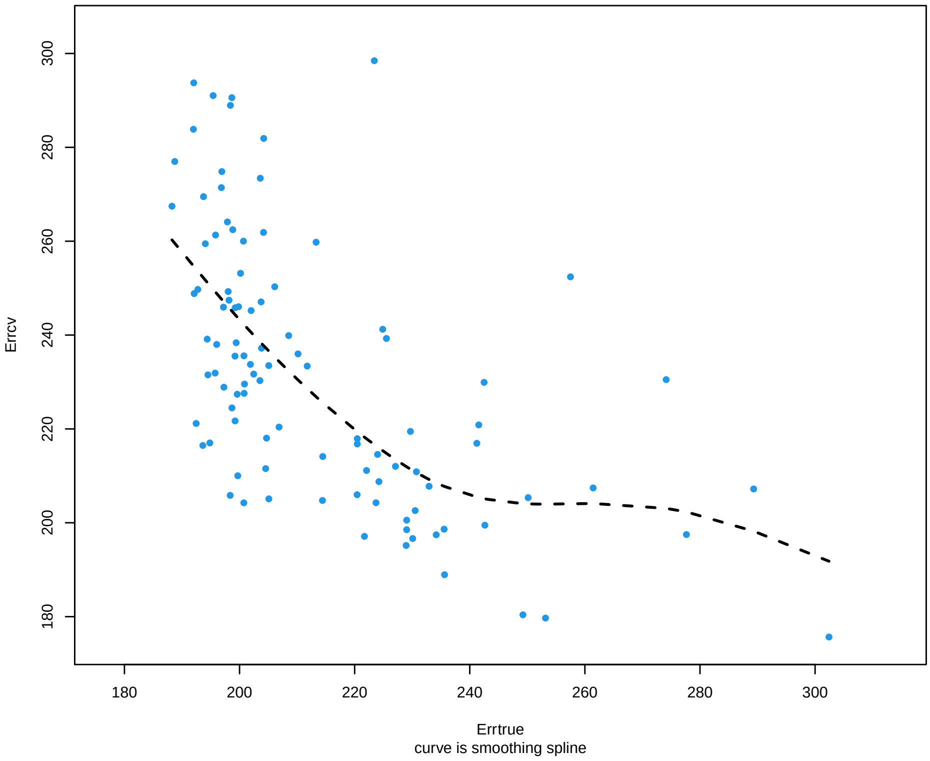

Figure 5. CV estimate of prediction error versus true prediction error, transplant data.

Sorry to say, that wasn't at all the case. The pairs $(\mathrm{Err}(j), \mathrm{Err_{cv}}(j)), j = 1, 2, \ldots, 100$, are plotted in Figure 5. It shows $\mathrm{Err_{cv}}$ actually *decreasing* as the true error Err *increases*. The unfortunate implication is that $\mathrm{Err_{cv}}$ isn't estimating the true error, but only its expectation $\mathrm{Err}^{(u)}$ (15). This is not a particular failing of cross-validation. It is habitually observed for all prediction error estimates — see for instance Figure 9 of [4] — though the phenomenon seems unexplained in the literature.

Cross-validation tends to pay for its low bias with high variability. Efron [1] proposed bootstrap estimates of prediction error intended to decrease variability without adding much bias. Among the several proposals the most promising was the *632 rule*,[7] described as follows:

- Nonparametric bootstrap samples $\boldsymbol{d}^*$ are formed by drawing n pairs (x_i, y_i) *with replacement* from the original data set $\boldsymbol{d}$ (10).
- Applying the original algorithm $\mathcal{A}$ to $\boldsymbol{d}^*$ gives prediction rule $f(x, \boldsymbol{d}^*)$ and predictions

$$\hat{\mu}(x)^* = f(x, \boldsymbol{d}^*), \tag{50}$$

  as in (12).
- B bootstrap data sets

$$\boldsymbol{d}^*(1), \boldsymbol{d}^*(2), \ldots, \boldsymbol{d}^*(B) \tag{51}$$

[7] An improved version *632+* was introduced in Efron and Tibshirani [6], designed for reduced bias in overfit situations where err (19) equals zero. The calculations here use only *632*.

are independently drawn, giving predictions

$$\hat{\mu}^*_{ij} = f(x_i, \boldsymbol{d}^*(j)), \tag{52}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, B$.

- Two numbers are recorded for each choice of i and j, the error of $\hat{\mu}^*_{ij}$ as a prediction of y_i,

$$Q^*_{ij} = Q(y_i, \hat{\mu}^*_{ij}) \tag{53}$$

and

$$N^*_{ij} = \#\{(x_i, y_i) \in \boldsymbol{d}^*(j)\}, \tag{54}$$

the number of times (x_i, y_i) occurs in $\boldsymbol{d}^*(j)$.

- The *zero bootstrap* $\widehat{\mathrm{Err}}_0$ is calculated as the average value of Q^*_{ij} for those cases having $N_{ij} = 0$,

$$\widehat{\mathrm{Err}}_0 = \frac{\sum_{(i,j):N_{ij}=0} Q^*_{ij}}{\sum_{(i,j):N_{ij}=0} 1} \tag{55}$$

- Finally, the 632 estimate of prediction error is defined to be

$$\widehat{\mathrm{Err}}_{632} = 0.632 \cdot \widehat{\mathrm{Err}}_0 + 0.368 \cdot \mathrm{err}, \tag{56}$$

err being the apparent error rate (19).

$\widehat{\mathrm{Err}}_{632}$ was calculated for the same 100 simulated response vectors $\boldsymbol{y}_j$ (45) used for $\widehat{\mathrm{Err}}_{\mathrm{cv}}$ (each using $B = 400$ replications), the 100 simulations giving

$$\widehat{\mathrm{Err}}_{632} \sim 1.15 \pm 0.10, \tag{57}$$

an improvement on $\widehat{\mathrm{Err}}_{\mathrm{cv}} \sim 1.16 \pm 0.14$ at (49). This is in line with the 24 sampling experiments reported in [6].[8]

Table 2. Mean of $Q(y, \hat{\mu})$ given the number of times a case (x_i, y_i) occurred in bootstrap sample; first simulation. True Error 0.97, apparent err 0.90.

# times:	**0**	**1**	**2**	**3**	**4**	**5**
mean Q:	1.44	.96	.81	.69	.64	.57

Table 2 concerns the rationale for the 632 rule. The $n \times B = 80,000$ values Q^*_{ij} for the first of the 100 simulated $\boldsymbol{y}$ vectors were averaged according to how many times (x_i, y_i) appeared in $\boldsymbol{d}^*(j)$,

$$\widehat{\mathrm{Err}}_k = \frac{\sum_{(i,j):N_{ij}=k} Q^*_{ij}}{\sum_{(i,j):N_{ij}=k} 1}, \tag{58}$$

for $k = 0, 1, 2, 3, 4, 5$. Not surprisingly, $\widehat{\mathrm{Err}}_k$ decreases with increasing k. $\widehat{\mathrm{Err}}_0$ (55), which would seem to be the bootstrap analogue of $\widehat{\mathrm{Err}}_{\mathrm{cv}}$, is seen to exceed the true error Err, while err is below Err. The intermediate linear combination $0.632\widehat{\mathrm{Err}}_0 + 0.368\mathrm{err}$ is motivated in Section 6 of [1], though in fact the argument is more heuristic than compelling. The 632 rules *do* usually reduce variability of the error estimates compared to cross-validation, but bias can be a problem.

[8] There rule 632+ was used, with loss function *counting error* (27) rather than binomial deviance. $Q(y, \hat{\mu})$ is discontinuous for counting error, which works to the advantage of 632 rules.

The 632 rule recomputes a prediction algorithm by nonparametric bootstrap resampling of the original data. *Random forests* [7], a widely popular prediction algorithm, carries this further: the algorithm itself as well as estimates of its accuracy depend on bootstrap resampling calculations.

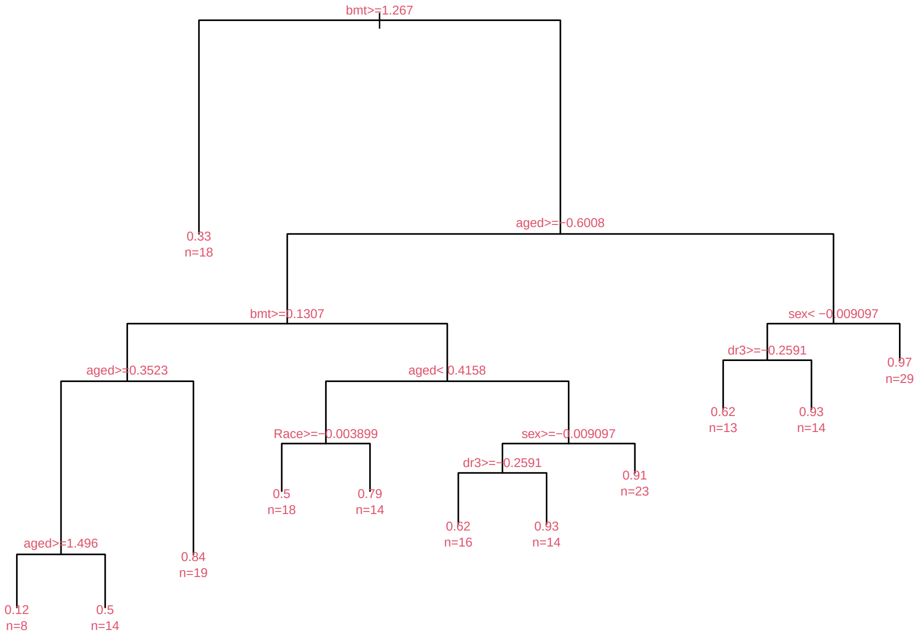

Figure 6. Regression tree for transplant data, simulation 1.

Regression trees are the essential component of random forests. Figure 6 shows one such tree,[9] as applied to $\boldsymbol{y}_1$, the first of the 100 response vectors for the transplant data simulation (44); $\boldsymbol{y}_1$ consists of 57 0s and 143 1s, average value $\bar{y} = 0.72$. The tree-making algorithm performs successive splits of the data, hoping to partition it into bins that are mostly 0s or 1s. The bin at the far right — comprising $n = 29$ cases having low body mass index, low age, and female gender — has just one 0 and 28 1s, for an average of 0.97. For a new transplant case (x, y), with only x observable, we could follow the splits down the tree and use the terminal bin average as a quantitative prediction of y.

Random forests improves the predictive accuracy of any one tree by *bagging* ("bootstrap aggregation"), sometimes also called *bootstrap smoothing*: B bootstrap data sets $\boldsymbol{d}^*(1), \boldsymbol{d}^*(2), \ldots, \boldsymbol{d}^*(B)$ are drawn at random (51), each one generating a tree like that in Figure 6.[10] A new x is followed down each of the B trees, with the random forest prediction being the average of x's B terminal values. Letting $\hat{\mu}_{ij} = f(x_i, \boldsymbol{d}^*(j))$, the prediction at x_i for the tree based on $\boldsymbol{d}^*(j)$, the random forest prediction at x_i is

$$\hat{\mu}_i = \sum_{j=1}^{B} \hat{\mu}_{ij} / B. \tag{59}$$

The predictive accuracy for $\hat{\mu}_i$ uses a device like that for $\widehat{\text{Err}}_0$ (55): Let $\hat{\mu}_{(i)}$ be the average value of $\hat{\mu}_{ij}$ for bootstrap samples $\boldsymbol{d}_j$ *not* containing (x_i, y_i),

$$\hat{\mu}_{(i)} = \frac{\sum_{j:N_{ij}=0} \hat{\mu}_{ij}}{\sum_{j:N_{ij}=0} 1}, \tag{60}$$

called the "out-of-bag" (oob) estimate of μ_i. The oob error estimate is then

[9] Constructed using `rpart`, the R version of *CART* [8]. Chapters 9 and 15 of [9] describe CART and random forests.

[10] Some additional variability is added to the tree-building process: only a random subset of the p predictors is deemed eligible at each splitting point.

$$\widehat{\mathrm{Err}}_{\mathrm{oob}}(x_i) = Q(y_i, \hat{\mu}_{(i)}). \tag{61}$$

The overall oob estimate of prediction error is

$$\widehat{\mathrm{Err}}_{\mathrm{oob}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathrm{Err}}_{\mathrm{oob}}(x_i). \tag{62}$$

(Notice that the leave-out calculations here are for the estimates $\hat{\mu}_i$, while those for $\widehat{\mathrm{Err}}_{632}$ (56) are for the errors Q^*_{ij}.) Calculated for the 100 simulated response vectors $\boldsymbol{y}_j$ (45), this gave

$$\widehat{\mathrm{Err}}_{\mathrm{oob}} \sim 1.12 \pm 0.07, \tag{63}$$

a better match to the true error Err than either $\widehat{\mathrm{Err}}_{\mathrm{cv}}$ (49) or $\widehat{\mathrm{Err}}_{632}$ (57). In fact the actual match was even better than (63) suggests, as shown in Table 3 of Section 4. This is all the more surprising given that, unlike $\widehat{\mathrm{Err}}_{\mathrm{cv}}$ and $\widehat{\mathrm{Err}}_{632}$, $\widehat{\mathrm{Err}}_{\mathrm{oob}}$ is fully nonparametric: it makes no use of the logistic regression model $\mathtt{glm}(y \sim \boldsymbol{X}, \mathtt{binomial})$ (46), which was involved in generating the simulated response vectors $\boldsymbol{y}_j$ (44). (It has to be added that $\widehat{\mathrm{Err}}_{\mathrm{oob}}$ is not an estimate for the prediction error of the logistic regression model $\hat{\mu}_{\mathtt{glm}}(x)$ (46), but rather for the random forest estimates $\hat{\mu}_{\mathrm{rf}}(x)$.)

4. Covariance penalties and degrees of freedom

A quite different approach to the prediction problem was initiated by Mallows' C_p *formula* [10]. An observed n-dimensional vector $\boldsymbol{y}$ is assumed to follow the homoskedastic model

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \qquad \text{with } \boldsymbol{\epsilon} \sim (\mathbf{0}, \sigma^2, \boldsymbol{I}), \tag{64}$$

the notation indicating uncorrelated errors of mean 0 and variance σ^2 as in (17); σ^2 is known. A linear rule

$$\hat{\boldsymbol{\mu}} = \boldsymbol{M}\boldsymbol{y} \tag{65}$$

is used to estimate $\boldsymbol{\mu}$ with $\boldsymbol{M}$ a fixed and known matrix. How accurate is $\hat{\boldsymbol{\mu}}$ as a predictor of future observations?

The apparent error

$$\mathrm{err} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 \tag{66}$$

is likely to underestimate the true error of $\hat{\boldsymbol{\mu}}$ given a hypothetical new observation vector $\boldsymbol{y}^* = \boldsymbol{\mu} + \boldsymbol{\epsilon}^*$ independent of $\boldsymbol{y}$,

$$\mathrm{Err} = \mathrm{Err}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \frac{1}{n}E_* \sum_{i=1}^{n}(y_i^* - \hat{\mu}_i)^2, \tag{67}$$

the E_* notation indicating that $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \dots, \hat{\mu}_n)$ is fixed in (67). Mallows' C_p formula says that

$$\widehat{\mathrm{Err}}_{\mathrm{cp}} = \mathrm{err} + \frac{2\sigma^2}{n}\,\mathrm{trace}(M) \tag{68}$$

is an unbiased estimator of Err,

$$E\left\{\widehat{\mathrm{Err}}_{\mathrm{cp}}\right\} = E\{\mathrm{Err}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})\}; \tag{69}$$

that is, $\widehat{\mathrm{Err}}_{\mathrm{cp}}$ *isn't* unbiased for $\mathrm{Err}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ but *is* unbiased for $E\{\mathrm{Err}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})\}$ ($\boldsymbol{\mu}$ fixed in the expectation) under model (64)–(65).

One might wonder what has happened to the covariates x_i in $\boldsymbol{d}$ (10)? The answer is that they are still there but no longer considered random–rather as fixed ancillary quantities like the sample size n. In the OLS model $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$ for Figure 3 (16)–(18) the covariates $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ determine $\boldsymbol{X}$, $\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$ and

$$\hat{\boldsymbol{\mu}} = \boldsymbol{M}\boldsymbol{y} \qquad \left(\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right), \tag{70}$$

(65). We could, but don't, write $\boldsymbol{M}$ as $\boldsymbol{M}_{\boldsymbol{X}}$.

Mallows' C_p formula (68) can be extended to the Q class of error measures, Figure 4. An unknown probability model[11] $\boldsymbol{f}$ is assumed to have produced $\boldsymbol{y}$ and its true mean vector $\boldsymbol{\mu}$,

$$\boldsymbol{\mu} = E_{\boldsymbol{f}}\{\boldsymbol{y}\}; \tag{71}$$

an estimate $\hat{\boldsymbol{\mu}} = m(\boldsymbol{y})$ has been calculated using some algorithm $m(\cdot)$,

$$\boldsymbol{f} \longrightarrow \boldsymbol{y} \longrightarrow \hat{\boldsymbol{\mu}} = m(\boldsymbol{y}); \tag{72}$$

the apparent error err (19) and true error $\text{Err}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ (33) are defined as before,

$$\text{err} = \frac{1}{n}\sum_{i=1}^{n} Q(y_i, \hat{\mu}_i) \quad \text{and} \quad \text{Err} = \frac{1}{n}E_*\left\{\sum_{i=1}^{n} Q(y_i^*, \hat{\mu}_i)\right\}, \tag{73}$$

with the $\hat{\mu}_i$ fixed and $\boldsymbol{f} \to \boldsymbol{y}^* = (y_1^*, y_2^*, \dots, y_n^*)$ independently of $\boldsymbol{y}$. Lemma 1 for the true error (36) still applies,

$$\text{Err} = \frac{1}{n}\sum_{i=1}^{n}[Q(\mu_i, \hat{\mu}_i) + q(\mu_i) - c(x_i)]. \tag{74}$$

The Q-class version of Mallows' formula (68) is derived as *Optimism Theorem 1* in Section 3 of [4]:

Theorem 1. *Define*

$$\hat{\lambda}_i = -\dot{q}(\hat{\mu}_i)/2 \qquad \left(\dot{q}(\mu) = \frac{d}{d\mu}q(\mu)\right). \tag{75}$$

Then

$$\widehat{Err} = err + \frac{2}{n}\sum_{i=1}^{n} \text{cov}_{\boldsymbol{f}}(\hat{\lambda}_i, y_i) \tag{76}$$

where $\text{cov}_{\boldsymbol{f}}$ *indicates covariance under model* (72), *is an unbiased estimate of Err in the same sense as* (69),

$$E_{\boldsymbol{f}}\left\{\widehat{Err}\right\} = E_{\boldsymbol{f}}\{Err\}. \tag{77}$$

The covariance terms in (76) measure how much each y_i affects its own estimate. They sum to a *covariance penalty* that must be added to the apparent error to account for the fitting process. If $Q(\mu, \hat{\mu})$ is binomial deviance then

$$\hat{\lambda}_i = \log[\hat{\mu}_i/(1 - \hat{\mu}_i)], \tag{78}$$

the logistic parameter; the theorem still applies as stated whether or not $\hat{\boldsymbol{\mu}} = m(\boldsymbol{y})$ is logistic regression.

[11] Notice that $\boldsymbol{f}$ is not the same as f in (12).

$\widehat{\mathrm{Err}}$ as stated in (76) is not directly usable since the covariance terms $\mathrm{cov}_f(\hat{\lambda}_i, y_i)$ are not observable statistics. This is where resampling comes in.

Suppose that the observed data $\boldsymbol{y}$ provides an estimate $\hat{\boldsymbol{f}}$ of $\boldsymbol{f}$. For instance in a normal regression model $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2, \boldsymbol{I})$ we could take $\hat{\boldsymbol{f}}$ to be $\boldsymbol{y}^* \sim \mathcal{N}(\hat{\boldsymbol{\mu}}, \sigma^2 \boldsymbol{I})$ for some estimate $\hat{\boldsymbol{\mu}}$. We replace (72) with the *parametric bootstrap model*

$$\hat{\boldsymbol{f}} \longrightarrow \boldsymbol{y}^* \longrightarrow \hat{\boldsymbol{\mu}}^* = m(\boldsymbol{y}^*) \tag{79}$$

and generate B independent replications $\boldsymbol{y}^*(1), \boldsymbol{y}^*(2), \ldots, \boldsymbol{y}^*(B)$, from which are calculated B pairs,

$$\left(y_i^*(j), \hat{\lambda}_i^*(j)\right), \qquad j = 1, 2, \ldots, B, \tag{80}$$

for $i = 1, 2, \ldots, n$ as in (75). The covariances in (76) can then be estimated as

$$\widehat{\mathrm{cov}}(\hat{\lambda}_i, y_i) = \frac{1}{B} \sum_{j=1}^{B} \left(\hat{\lambda}_i^*(j) - \hat{\lambda}_i^*(\cdot)\right)\left(y_i^*(j) - y_i^*(\cdot)\right), \tag{81}$$

$\hat{\lambda}_i^*(\cdot) = \sum_j \hat{\lambda}_i^*(j)/B$ and $y_i^*(\cdot) = \sum_j y_i^*(j)/B$, yielding a useable version of (76),

$$\widehat{\mathrm{Err}}_{\mathrm{cp}} = \mathrm{err} + \frac{2}{n} \sum_{i=1}^{n} \widehat{\mathrm{cov}}(\hat{\lambda}_i, y_i), \tag{82}$$

“cp” standing for “covariance penalty”.

Table 3. Transplant data simulation experiment. *Top:* First 5 of 100 estimates of total prediction error using cross-validation, covariance penalties (cp), the 632 rule, the out-of-bag random forest results, and the apparent error; degrees of freedom (df) estimate from cp. *Bottom:* Mean of the 100 simulations, their standard deviations, correlations with Errtrue, and root mean square differences from Errtrue. Cp and 632 rules used $B = 400$ bootstrap replications per simulation.

	Errtrue	**ErrCv**	**ErrCp**	**Err632**	**Erroob**	**err**	**df**
	194	216.5	232.9	260.6	240.3	180.4	26.2
	199	245.8	220.9	241.0	230.9	173.8	23.5
	192	221.2	225.9	247.6	232.7	180.6	22.7
	234	197.4	200.9	215.2	219.7	162.2	19.4
	302	175.7	178.3	187.1	211.2	144.6	16.9

	Errtrue	**ErrCv**	**ErrCp**	**Err632**	**Erroob**	**err**	**df**
mean	214.0	231.1	216.1	229.4	223.8	169.9	23.1
stdev	23.0	28.7	15.8	19.0	14.6	13.5	4.0
cor.true		−.58	−.61	−.64	−.26	−.41	−.52
rms		48.8	34.8	40.7	31.6	54.2	

Table 3 compares the performances of cross-validation, covariance penalties, the 632 rule, and the random forest out-of-bag estimates on the 100 transplant data simulations. The results are given in terms of *total* prediction error, rather than average error as in (47). The bottom line shows their root-mean-square differences from true error Err,

$$\mathrm{rms} = \left[\sum_{i=1}^{100} \left(\widehat{\mathrm{Err}}_i - \mathrm{Err}_i\right)^2 \Big/ 100\right]^{1/2}. \tag{83}$$

$\widehat{\mathrm{Err}}_{\mathrm{cv}}$ is highest, with $\widehat{\mathrm{Err}}_{\mathrm{cp}}$, $\widehat{\mathrm{Err}}_{632}$, and $\widehat{\mathrm{Err}}_{\mathrm{oob}}$ respectively 71%, 83%, and 65% as large.

For the jth simulation vector $\boldsymbol{y}_j$ (45) the parametric bootstrap replications (79) were generated as follows: the logistic regression estimate $\hat{\boldsymbol{\mu}}_j$ was calculated from (46),

$\hat{\boldsymbol{\mu}}_j = \texttt{glm}(\boldsymbol{y}_j \sim \boldsymbol{X}, \texttt{binomial})\texttt{\$,fit}$; the bootstrap replications $\boldsymbol{y}_j^*$ had independent Bernoulli components

$$y_{ij}^* \overset{\text{ind}}{\sim} \text{Bi}(1, \hat{\mu}_{ij}) \qquad \text{for } i = 1, 2, \ldots, 200. \tag{84}$$

$B = 400$ independent replications $\boldsymbol{y}_j^*$ were generated for each $j = 1, 2, \ldots, 100$, giving $\widehat{\text{Err}}_{\text{cp}}$ according to (81)–(82).

Because the resamples were generated by a model of the same form as that which originally gave the $\boldsymbol{y}_j$'s (43)–(44), $\widehat{\text{Err}}_{\text{cp}}$ is unbiased for Err. In practice our resampling model $\hat{f}$ in (79) might not match f in (72), causing $\widehat{\text{Err}}_{\text{cp}}$ to be downwardly biased (and raising its rms). $\widehat{\text{Err}}_{\text{cv}}$'s nonparametric resamples make it nearly unbiased for the unconditional error rate $\text{Err}^{(u)}$ (15) irrespective of the true model, accounting for the overwhelming popularity of cross-validation in applications.

A few comments

- In computing $\widehat{\text{Err}}_{\text{cp}}$ it isn't necessary for $\hat{f}$ in (79) to be constructed from the original estimate $\hat{\boldsymbol{\mu}}$. We might base $\hat{f}$ on a bigger model than that which led to the choice of $m(\boldsymbol{y})$ for $\hat{\boldsymbol{\mu}}$; in the little example of Figure 3, for instance, we could take $\hat{f}$ to be $\mathcal{N}(\hat{\boldsymbol{\mu}}(6), \sigma^2 \boldsymbol{I})$ where $\hat{\boldsymbol{\mu}}(6)$ is the OLS sixth degree polynomial fit, while still taking $\hat{\boldsymbol{\mu}} = m(\boldsymbol{y})$ to be fourth degree as in (18). This reduces possible model-fitting bias in $\widehat{\text{Err}}_{\text{cp}}$, while increasing its variability.
- A major conceptual difference between $\widehat{\text{Err}}_{\text{cv}}$ and $\widehat{\text{Err}}_{\text{cp}}$ concerns the role of the covariates $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, considered as random in model (10) but fixed in (72). Classical regression problems have usually been analyzed in a fixed-$\boldsymbol{x}$ framework for three reasons:

  1. mathematical tractability;
  2. not having to specify $\boldsymbol{x}$'s distribution;
  3. inferential relevance.

  The reasons come together in the classic covariance formula for the linear model $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$,

$$\text{cov}(\hat{\beta}) = \sigma^2 \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}. \tag{85}$$

  A wider dispersion of the x_i's in Figure 3 would make $\hat{\beta}$ more accurate, and conversely.
- It can be argued that because $\widehat{\text{Err}}_{\text{cp}}$ is estimating the conditional error *given* $\boldsymbol{x}$ it is more relevant to what is being estimated. See [11] and [12] for a lot more on this question.
- On the other hand, "fixed-$\boldsymbol{x}$" methods such as Mallows' C_p can be faulted for missing some of the variability contributing to the unconditional prediction error $\text{Err}^{(u)}$ (15). Let $\text{Err}(x, \boldsymbol{d})$ be the conditional error given $\boldsymbol{d}$ for predicting y given x,

$$\text{Err}(x, \boldsymbol{d}) = Q(\mu(x), f(x, \boldsymbol{d})) + q(\mu(x)) - c(x) \tag{86}$$

  according to Lemma 1 (34). Then

$$\text{Err}^{(u)} = E\left\{\int_{\mathcal{X}} g(x)\, \text{Err}(x, \boldsymbol{d})\, dx\right\}, \tag{87}$$

  where $g(x)$ is the marginal density of x and E indicates expectation over the choice of the training data $\boldsymbol{d}$.
- In the fixed-$\boldsymbol{x}$ framework of (36), Err replaces the integrand in (87) with its $\boldsymbol{x}$ average $\sum_1^n \text{Err}(x_i, \boldsymbol{d})/n$. We expect

$$\text{Err}^{(u)} > E_f\{\text{Err}\} \tag{88}$$

since $\text{Err}(x, \boldsymbol{d})$ typically increases for values of x farther away from $\boldsymbol{x}$. Rosset and Tibshirani [12] give an explicit formula for the difference in the case of normal-theory ordinary least squares when they show that the factor 2 for the penalty term in Mallows' C_p formula should be increased to $2 + (p+1)/(n-p-1)$. Cross-validation effectively estimates $\text{Err}^{(u)}$ while C_p estimates the fixed-$\boldsymbol{x}$ version of $E\{\text{Err}\}$.

- With (88) in mind, $\widehat{\text{Err}}_{\text{cp}}$ and $\widehat{\text{Err}}_{\text{cv}}$ are often contrasted as

$$\textit{insample} \quad \text{versus} \quad \textit{outsample}. \tag{89}$$

This is dangerous terminology if it's taken to mean that $\widehat{\text{Err}}_{\text{cv}}$ applies to prediction errors $\text{Err}(x, \boldsymbol{d})$ at individual points x outside of $\boldsymbol{x}$. In Figure 3 for instance it seems likely that $\text{Err}(11, \boldsymbol{d})$ exceeds $\widehat{\text{Err}}_{\text{cv}}$, but this is a fixed-$\boldsymbol{x}$ question and beyond the reach of the random-$\boldsymbol{x}$ assumptions underlying (88). See the discussion of Figure 8 in Section 5.

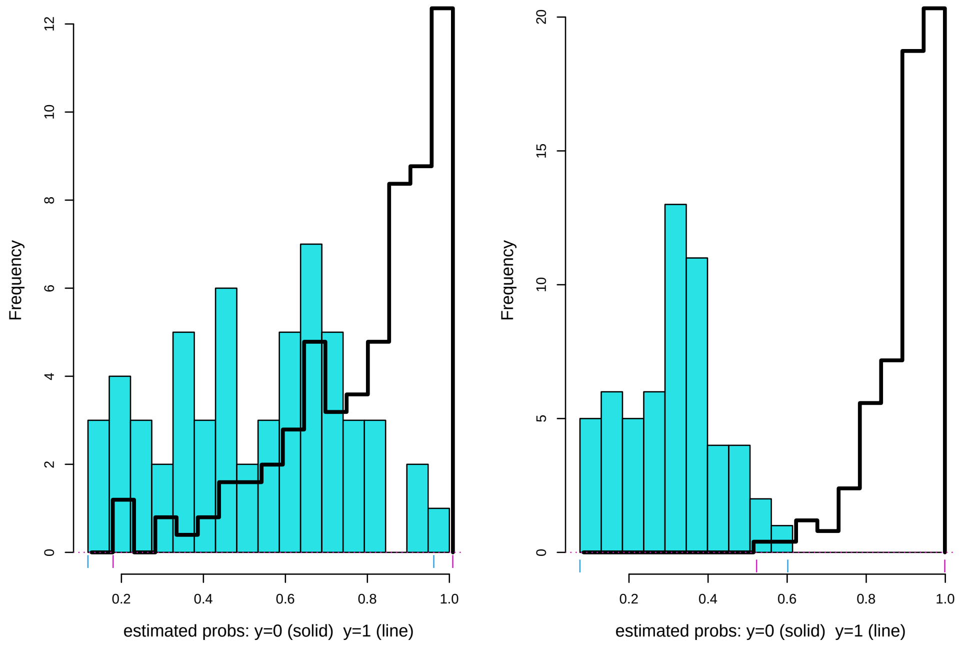

Figure 7. Left: Estimated probabilities, logistic regression simulation 1. **Right:** Now for random forests, simulation 1. The increased random forests separation suggests increased degrees of freedom.

- The sad story told in Figure 5 shows $\widehat{\text{Err}}_{\text{cv}}$ *negatively* correlated with the true Err. The same is the case for $\widehat{\text{Err}}_{\text{cp}}$ and $\widehat{\text{Err}}_{632}$, as can be seen from the negative correlations in the *cor.true* row of Table 3. $\widehat{\text{Err}}_{\text{oob}}$ is also negatively correlated with Err, but less so. In terms of *rms*, the bottom row shows that the fully nonparametric $\widehat{\text{Err}}_{\text{oob}}$ estimates beat even the parametric $\widehat{\text{Err}}_{\text{cp}}$ ones.
- Figure 7 compares the 200 $\hat{\mu}_i$ estimates from the logistic regression estimates (46) with those from random forests, for y the first of the 100 transplant simulations. Random forests is seen to better separate the $y_i = 0$ from the $y_i = 1$ cases. $\widehat{\text{Err}}_{\text{oob}}$ relates to error prediction for random forest estimates, *not* for logistic regression, but this doesn't explain how $\widehat{\text{Err}}_{\text{oob}}$ could provide excellent estimates of the true error Err–which in fact were based on the logistic regression model (43)–(44). If this is a fluke it's an intriguing one.
- There is one special case where the covariance penalty formula (76) can be unbiasedly estimated without recourse to resampling: if f is the normal model

$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I)$, and $Q(y_i, \mu_i) = (y_i - \mu_i)^2$ — so $\hat{\lambda}_i = \hat{\mu}_i$ — then *Stein's unbiased risk estimate* (SURE) is defined to be

$$\widehat{\text{Err}}_{\text{SURE}} = \text{err} + \frac{2\sigma^2}{n} \sum_{i=1}^{n} \frac{\partial \hat{\mu}_i}{\partial y_i}, \tag{90}$$

where the partial derivatives are calculated directly from the functional form of $\hat{\boldsymbol{\mu}} = m(\boldsymbol{y})$. Section 2 of [4] gives an example comparing $\widehat{\text{Err}}_{\text{SURE}}$ with $\widehat{\text{Err}}_{\text{cp}}$. Each term $\partial \hat{\mu}_i / \partial y_i$ measures the influence of y_i on its own estimate.

Degrees of freedom

The OLS model $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$ yields the familiar estimate $\hat{\boldsymbol{\mu}} = \boldsymbol{M}\boldsymbol{y}$ of $\boldsymbol{\mu} = E\{\boldsymbol{y}\}$, where $\boldsymbol{M}$ is the projection matrix $\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ as in (70); $\boldsymbol{M}$ has

$$\text{trace}(\boldsymbol{M}) = \text{trace}\left[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\right] = p, \tag{91}$$

where $p = \text{rank}(\boldsymbol{X})$. Mallows' C_p formula $\widehat{\text{Err}}_{\text{cp}} = \text{err} + (2\sigma^2/n)\,\text{trace}(\boldsymbol{M})$ becomes

$$\widehat{\text{Err}}_{\text{cp}} = \text{err} + \frac{2\sigma^2}{n} p \tag{92}$$

in this case. In other words, the covariance penalty that must be added to the apparent error err is directly proportional to p, the degrees of freedom of the OLS model.

Suppose that $\hat{\boldsymbol{\mu}} = \boldsymbol{M}\boldsymbol{y}$ with matrix $\boldsymbol{M}$ not necessarily a projection. It has become common to define $\hat{\boldsymbol{\mu}}$'s degrees of freedom as

$$\text{df} = \text{degrees of freedom} = \text{trace}(M), \tag{93}$$

trace(M) playing the role of p in the C_p formula (92). In this way, trace(M) becomes a *lingua franca* for comparing linear estimators $\hat{\boldsymbol{\mu}} = \boldsymbol{M}\boldsymbol{y}$ of different forms.[12]

A nice example is the *ridge regression* estimator

$$\hat{\boldsymbol{\mu}}_\gamma = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X} + \gamma \boldsymbol{I}\right)^{-1}\boldsymbol{X}^\top y, \tag{94}$$

γ a fixed non-negative constant; $\hat{\boldsymbol{\mu}}_0$ is the usual OLS estimator while $\hat{\boldsymbol{\mu}}_\gamma$ "shrinks" $\hat{\boldsymbol{\mu}}$ toward $\mathbf{0}$, more so as γ increases. Some linear algebra gives the degrees of freedom for (94) as

$$\text{df}_\gamma = \sum_{i=1}^{p} \frac{e_i}{e_i + \gamma}, \tag{95}$$

where the e_i are the eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$.

The generalization of Mallows' formula (68) in Theorem 1 (76) has penalty term

$$\frac{2}{n} \sum_{i=1}^{n} \text{cov}_f(\hat{\lambda}_i, y_i), \tag{96}$$

which again measures the self-influence of each y_i on its own estimate $\hat{\mu}_i$. The choice of

$$q(\mu) = \mu(1 - \mu) \tag{97}$$

in Figure 3 results in $Q(\mu, \hat{\mu}) = (\hat{\mu} - \mu)^2$, squared error, and $\hat{\lambda}_i = \hat{\mu}_i - 1/2$, in which case $\text{cov}_f(\hat{\lambda}_i, y_i)$ equals $\sigma^2 m_{ii}$, and (96) becomes $(2\sigma^2/n)\,\text{trace}(\boldsymbol{M})$, as in Mallows' formula (68). This suggests using

[12] The referee points out that formulas like (92) are more often used for model selection rather than error rate prediction. Zhang and Yang [13] consider model selection applications, as does Remark B of [4].

$$\mathbf{Cov} = \sum_{i=1}^{n} \operatorname{cov}_f(\hat{\lambda}_i, y_i) \tag{98}$$

as a measure of degrees of freedom (or its estimate $\widehat{\mathbf{Cov}}$ from (81)) for a general estimator $\hat{\boldsymbol{\mu}} = m(\boldsymbol{y})$.

Some support comes from the following special situation: suppose $\boldsymbol{y}$ is obtained from a p-parameter generalized linear model, with prediction error measured by the appropriate deviance function.[13] Theorem 2 of [14], Section 6, then gives the asymptotic approximation

$$\mathbf{Cov} \doteq p, \tag{99}$$

as in (91)–(92), the intuitively correct answer.

Approximation (99) leads directly to *Akaike's information criterion* (AIC). In a generalized linear model, the total deviance from the MLE $\hat{\boldsymbol{\mu}}$ is

$$n \cdot \mathrm{err} = \sum_{i=1}^{n} Q(y_i, \hat{\mu}_i) = 2\left[\sum_{i=1}^{n} \log\left(g_{y_i}(y_i)/g_{\hat{\mu}_i}(y_i)\right)\right], \tag{100}$$

$g_{\hat{\mu}_i}(y_i)$ denoting the density function [2, Hoeffding's Lemma]. Suppose we have glm's of different sizes p we wish to compare. Minimizing $\mathrm{err} + (2/n)\mathbf{Cov}$ over the choice of model is then equivalent to maximizing the total log likelihood $\log g_{\hat{\boldsymbol{\mu}}}(\boldsymbol{y})$ minus a dimensionality penalty,

$$\log\left(g_{\hat{\boldsymbol{\mu}}}(\boldsymbol{y})\right) - \mathrm{Cov} \doteq \log\left(g_{\hat{\boldsymbol{\mu}}}(\boldsymbol{y})\right) - p, \tag{101}$$

which is the AIC.

Approximation (99) isn't razor-sharp: $p = 17$ is the transplant simulation logistic regression but the 100 estimates $\widehat{\mathbf{Cov}}$ averaged 23.08 with standard error 0.40. Degrees of freedom play a crucial role in model selection algorithms. Resampling methods allow us to assess $\mathbf{Cov}$ (98) even for very complicated fitting algorithms $\hat{\boldsymbol{\mu}} = m(\boldsymbol{y})$.

5. Conformal inference

If there is a challenger to cross-validation for "oldest resampling method" it is *permutation testing*, going back to Fisher in the 1930s. The newest prediction error technique, *conformal inference*, turns out to have permutation roots, as briefly reviewed next.

A clinical trial of an experimental drug has yielded independent real-valued responses for control and treatment groups:

$$\text{Control}: \quad \boldsymbol{u} = (u_1, \ldots, u_n) \quad \text{and} \quad \text{Treatment}: \quad \boldsymbol{v} = (v_1, \ldots, v_m). \tag{102}$$

Student's t-test could be used to see if the new drug was giving genuinely larger responses but Fisher, reacting to criticism of normality assumptions, proposed what we would now call a nonparametric two-sample test.

Let $\boldsymbol{z}$ be the combined data set,

$$\boldsymbol{z} = (\boldsymbol{u}, \boldsymbol{v}) = (z_1, \ldots, z_{n+m}), \tag{103}$$

and choose some score function $S(\boldsymbol{z})$ that contrasts the last m z-values with the first n, for example the difference of means,

$$S(\boldsymbol{z}) = \sum_{i=n+1}^{n+m} \frac{z_i}{m} - \sum_{i=1}^{n} \frac{z_i}{n}. \tag{104}$$

[13] Binomial deviance for logistic regression as in the transplant example.

Define $\mathcal{S}$ as the set of scores for all permutations of $\boldsymbol{z}$,

$$\mathcal{S} = \{S(\boldsymbol{z}^*)\}, \tag{105}$$

$\boldsymbol{z}^*$ ranging over the $(m+n)!$ permutations.

The permutation p-value for the treatment's efficacy in producing larger responses is defined to be

$$p = \#\{S(\boldsymbol{z}^*) \geq S(\boldsymbol{z})\}/(m+n)!, \tag{106}$$

the proportion of $\mathcal{S}$ having scores $S(\boldsymbol{z}^*)$ exceeding the observed score $S(\boldsymbol{z})$. Fisher's key idea was that if in fact all the observations came from the same distribution F,

$$z_i \overset{\text{ind}}{\sim} F \qquad (i = 1, 2, \ldots, m+n) \tag{107}$$

(implying that Treatment is the same as Control), then all $(m+n)!$ permutations would be equally likely. Rejecting the null hypothesis of No Treatment Effect if $p \leq \alpha$ has null probability (nearly) α.

Usually $(m+n)!$ is too many for practical use. This is where the sampling part of resampling comes in. Instead of all possible permutations, a randomly drawn subset of B of them, perhaps $B = 1000$, is selected for scoring, giving an estimated permutation p-value

$$\hat{p} = \#\{S(\boldsymbol{z}^*) \geq S(\boldsymbol{z})\}/B. \tag{108}$$

In 1963, Hodges and Lehmann considered an extension of the null hypothesis (107) to cover location shifts; in terms of cumulative distribution functions (cdfs), they assumed

$$u_i \overset{\text{iid}}{\sim} F(u) \quad \text{and} \quad v_i \overset{\text{iid}}{\sim} F(v - \Delta), \tag{109}$$

where Δ is a fixed but unknown constant that translates the v's distribution by Δ units to the right of the u's.

For a given trial value of Δ let

$$\boldsymbol{z}(\Delta) = (u_1, \ldots, u_n, v_1 - \Delta, \ldots, v_m - \Delta) \tag{110}$$

and compute its permutation p-value

$$\hat{p}(\Delta) = \#\{S(\boldsymbol{z}^*(\Delta)) \geq S(\boldsymbol{z}(\Delta))\}. \tag{111}$$

A 0.95 two-sided nonparametric confidence interval for Δ is then

$$\Delta : \quad 0.025 \leq \hat{p}(\Delta) \leq 0.975. \tag{112}$$

The only assumption is that, for the true value of Δ, the $m+n$ components of $\boldsymbol{z}(\Delta)$ are i.i.d.[14] from some distribution F.

Vovk has proposed an ingenious extension of this argument applying to prediction error estimation, a much cited reference being [15]; see also [16]. Returning to the statement of the prediction problem at the beginning of Section 2, $\boldsymbol{d} = \{(x_i, y_i), i = 1, 2, \ldots, n\}$ is the observed data and (x_0, y_0) a new (predictor, response) pair, all $n+1$ pairs assumed to be random draws from the same distribution F,

$$(x_i, y_i) \overset{\text{iid}}{\sim} F \qquad \text{for } i = 0, 1, \ldots, n; \tag{113}$$

x_0 is observed but not y_0, and it is desired to predict y_0. Vovk's proposal, *conformal inference*, produces an exact nonparametric distribution of the unseen y_0.

[14] More generally, we only need the components to be exchangeable.

Let Y_0 be a proposed trial value for y_0, and define $\boldsymbol{D}$ as the data set $\boldsymbol{d}$ augmented with (x_0, Y_0),

$$\boldsymbol{D} = \{\boldsymbol{d}, (x_0, Y_0)\}. \tag{114}$$

A prediction rule $f(x, \boldsymbol{D})$ gives estimates

$$\hat{\mu}_i = f(x_i, \boldsymbol{D}) \qquad \text{for } i = 0, 1, \dots, n. \tag{115}$$

(It is required that $f(x, \boldsymbol{D})$ be invariant under reordering of $\boldsymbol{D}$'s elements.)

For some score function $s(y, \hat{\mu})$ let

$$s_i = s(y_i, \hat{\mu}_i) \qquad \text{for } i = 1, 2, \dots, n, \tag{116}$$

where $s(y, \hat{\mu})$ measures disagreement between y and $\hat{\mu}(x)$, larger values of $|s|$ indicating less conformity between observation and prediction.[15]

If the proposed trial value Y_0 were in fact the unobserved y_0 then $s_1, s_2, \dots, s_n$ and $s_0 = s(Y_0, \hat{\mu}_0)$ would be exchangeable random variables, because of the i.i.d. assumption (113). Let $s_{(1)}, s_{(2)}, \dots, s_{(n)}$ be the ordered values of $s_1, s_2, \dots, s_n$. Assuming no ties, the n values

$$s_{(1)} < s_{(2)} < \dots < s_{(n)} \tag{117}$$

partition the line into $(n+1)$ intervals, the first and last of which are semi-infinite. Exchangeability implies that s_0 has probability $1/(n+1)$ of falling into any one of the intervals.

A conformal interval for the unseen y_0 consists of those values of Y_0 for which $s_0 = s(Y_0, \hat{\mu}_0)$ "conforms" to the distribution (117). To be specific, for a chosen miscoverage level α, say 0.05, let I_0 and I_1 be integers approximately proportion $\alpha/2$ from the endpoints of $1, 2, \dots, n$,

$$I_0 = [n\alpha/2] \quad \text{and} \quad I_1 = [n(1 - \alpha/2)] + 1. \tag{118}$$

The conservative two-sided level $1 - \alpha$ conformal prediction interval $\mathcal{C}$ for y_0 is

$$\mathcal{C} = \left\{ Y_0 : s(Y_0, \boldsymbol{D}) \in \left[s_{(I_0)}, s_{(I_1)} \right] \right\}. \tag{119}$$

The argument is the same as for the Hodges–Lehmann interval (112), now with $m = 1$ and $\Delta = Y_0$.

Interval (119) is computationally expensive since all of the s_i, not just s_0, change with each choice of trial value Y_0. The *jackknife conformal interval* begins with the jackknife estimates

$$\hat{\mu}_{(i)} = f\left(x_i, \boldsymbol{d}_{(i)}\right), \qquad i = 1, 2, \dots, n, \tag{120}$$

where $\boldsymbol{d}_{(i)} = \{(x_j, y_j), j \neq i\}$, that is $\boldsymbol{d}$ (10) with (x_i, y_i) deleted. The scores s_i (116) are taken to be

$$s_i = s\left(y_i, \hat{\mu}_{(i)}\right), \qquad i = 1, 2, \dots, n, \tag{121}$$

for some function s, for example $s_i = y_i - \hat{\mu}_{(i)}$. These are compared with

$$s_0 = s(Y_0, \hat{\mu}_0), \tag{122}$$

$\hat{\mu}_0 = f(x_0, \boldsymbol{d})$, and $\mathcal{C}$ is computed as at (119). Now the score distribution (117) does not depend on Y_0 (nor does $\hat{\mu}_0$), greatly reducing the computational burden.

[15] More generally s_i can be any function $s(y_i, \boldsymbol{D})$.

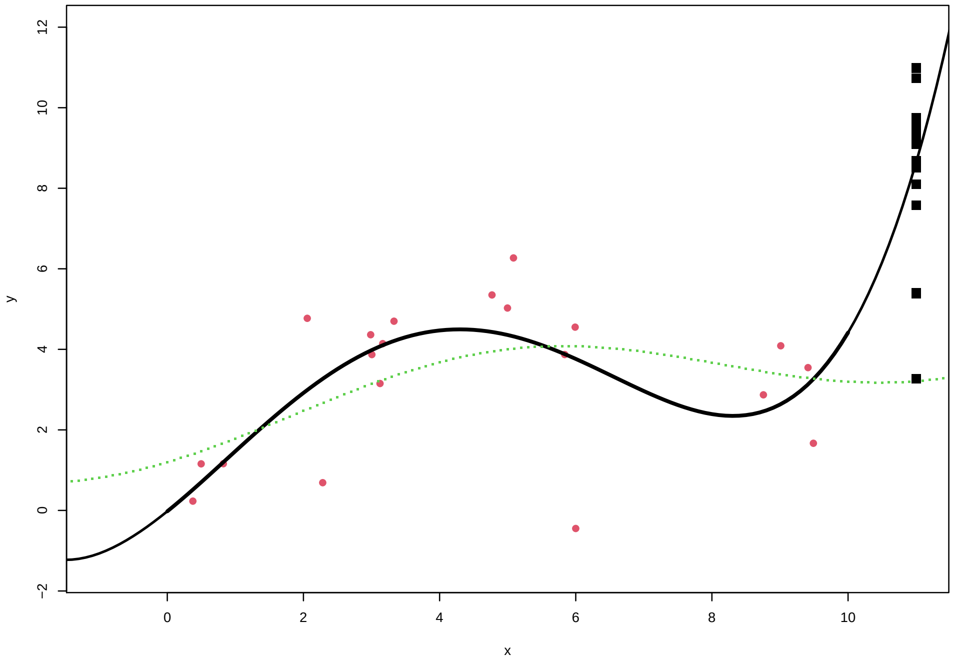

Figure 8. Square points: jackknife conformal predictions at $x = 11$ for example in Figure 3; each interval Pr $= 1/21$.

The jackknife conformal interval at $x_0 = 11$ was calculated for the small example of Figure 2 using

$$s_i = y_i - \hat{\mu}_{(i)} \tag{123}$$

for $i = 1, 2, \ldots, n = 20$. For this choice of scoring function, interval (119) is

$$\mathcal{C} = \left[\hat{\mu}_0 + s_{(I_0)}, \hat{\mu}_0 + s_{(I_1)})\right]; \tag{124}$$

$y_0 \in \mathcal{C}$ with conformal probability $(I_{(1)} - I_{(0)})/(n+1)$. The square dots in Figure 8 are the values $\hat{\mu}_0 + s_{(i)}$ for $i = 1, 2, \ldots, 20$, with y_0 having probability $1/21$ of falling into each of the 21 intervals. Conformal probability for the full range

$$\left[\hat{\mu}_0 + s_{(1)}, \hat{\mu}_0 + s_{(20)}\right] = [3.27, 10.99] \tag{125}$$

is $19/21 = 0.905$.

Some comments

- Even when Y_0 equals y_0, $s_0 = s(Y_0, \hat{\mu}_0)$ is not perfectly exchangeable with the $s_i = s(y_i, \hat{\mu}_{(i)})$ (121): each $\hat{\mu}_{(i)}$ is based on $n-1$ other cases, while $\hat{\mu}_0$ is based on n. Other stand-ins for the full conformal intervals (119) are favored in the literature but these have their own disadvantages. Barber *et al.* [17] offer a version of the jackknife intervals, "jackknife +", with more dependable inferential performance.
- The jackknife scores s_i (121) are also the one-at-a-time cross-validation scores if s is taken to be the prediction loss Q in (42). In this sense, conformal inference can be thought of as a more ambitious version of prediction error estimation, where we try to estimate the entire error distribution rather than just its expectation. The conformal point estimate $\bar{s} = \sum_1^n s_i/n$ is the same as $\widehat{\mathrm{Err}}_{\mathrm{cv}}$ (42) if s equals Q (which is why "conformal" wasn't included in Table 3).
- Figure 8 is misleading in an important sense: the 95% coverage claimed in (125) is a marginal inference following from the i.i.d. assumption $(x_i, y_i) \overset{\text{iid}}{\sim} F$ for $i = 0, 1, \ldots, n$ (113), and doesn't apply conditionally to the particular configuration

of x's and y's seen in the figure. (See Remark 3 and Section 3 of [16].) The same complaint was leveled against cross-validation in Section 3 — for estimating the unconditional error $\text{Err}^{(u)}$ rather than prediction error for the rule at hand — but conformal inference leans even harder on the i.i.d. assumption.

- Classic parametric prediction intervals *do* apply conditionally. The normal-theory version of model (16)–(17) gives 95% interval

$$\hat{\mu}_0 \pm 1.96\,\sigma\sqrt{1+\gamma_0} \tag{126}$$

for response y_0 at $x_0 = 11$, where

$$\gamma_0 = X(x_0)^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} X(x_0) \tag{127}$$

in notation (18).[16] The nonparametric random sampling assumption (113) destroys the geometry seen in Figure 8.

- Rather than $s_i = y_i - \hat{\mu}_{(i)}$ at (123), we might use scores

$$s_i = \frac{y_i - \hat{\mu}_{(i)}}{\gamma_i}, \tag{128}$$

where $\gamma_i = \gamma(x_i)$ is some measure of prediction difficulty at x_i. Conformal inference continues to apply here since the s_i are still exchangeable. Boström *et al.* [18] give several random forest examples, using out-of-bag estimates for the γ_i.

- *Covariate shift* estimation offers a more ambitious approach to broadening the reach of conformal prediction; see [19] and [20]. The underlying probability distribution F in (113) can be thought of in two stages, first choosing x according to say $g(x)$ and then y given x according to $p(y \mid x)$,

$$F: \quad x \sim g(x) \quad \text{and} \quad y \mid x \sim p(y \mid x). \tag{129}$$

It is assumed that (129) holds in a training set, but in the test set where predictions are to be made $g(x)$ is shifted to $g_{\text{test}}(x)$,

$$F_{\text{test}}: \quad x \sim g_{\text{test}}(x) \quad \text{and} \quad y \mid x \sim p(y \mid x). \tag{130}$$

With sufficiently large training and test sets available, the ratio $g_{\text{test}}(x)/g(x)$ can be estimated, allowing a suitably weighted version of conformal interval (124) to be constructed.

- Conformal prediction is less appealing for dichotomous response data but can still be informative. Figure 9 shows its application to the transplant data of Section 3 and Section 4. The score function s is taken to be the deviance residual

$$s_i = \text{sign}\Big(y_i - \hat{\mu}_{(i)}\Big) Q\Big(y_i, \hat{\mu}_{(i)}\Big)^{1/2}, \tag{131}$$

$\hat{\mu}_{(i)}$ the jackknife logistic regression estimate (120) and $Q(y,\mu)$ binomial deviance (28). The left side of Figure 9 shows the histogram of the 200 s_i values. Any given value of $\hat{\mu}_0$ corresponds to two possible values of $s_0 = \text{sign}(y_0 - \hat{\mu}_0)Q(y_0, \hat{\mu}_0)^{1/2}$, for y_0 equal 0 or 1, and two values of the conformal p-value,

$$p(\hat{\mu}_0) = \#\{s_i \geq s_0\}/201. \tag{132}$$

The right side of Figure 9 graphs $p(\hat{\mu}_0)$ as a function of $\hat{\mu}_0$ for the two cases:

[16] At $x_0 = 11$, (126) gives 95% prediction interval $[-15.6, 32.9]$, reflecting the hopelessly large extrapolation variability of the fourth-degree polynomial model; the standard deviation of $\hat{\mu}_0$ at $x_0 = 11$ is 12.22.

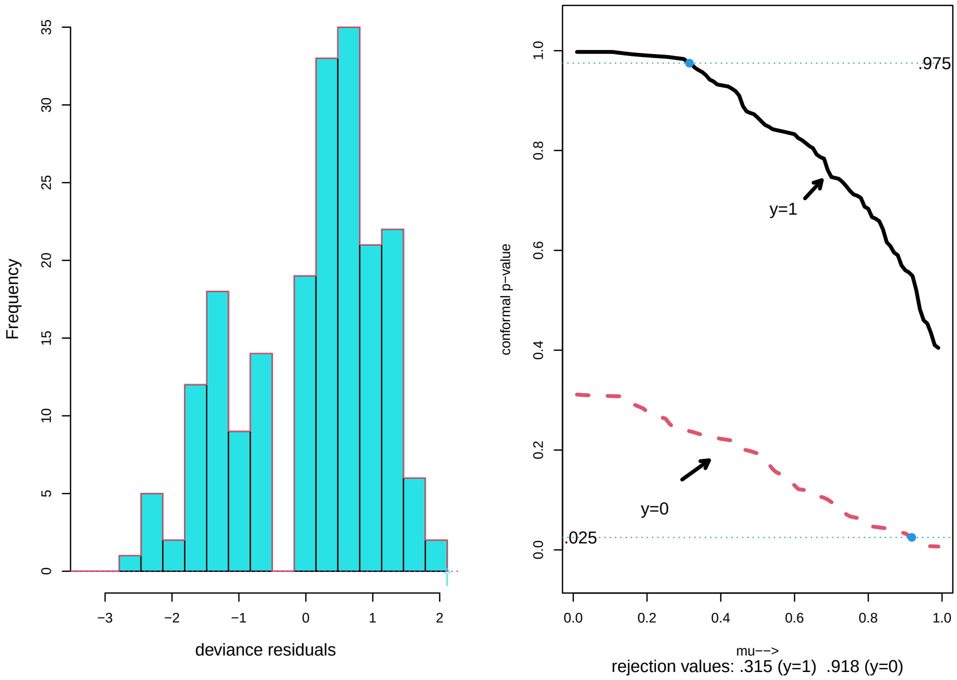

Figure 9. Left: Signed deviance residuals for transplant data. **Right:** Attained conformal p-value given μ; solid $y = 1$, dashed $y = 0$.

$$\begin{aligned} \hat{\mu}_0 &< 0.315 \text{ gives } \hat{p} \geq 0.975 \text{ for } y_0 = 1, \\ \hat{\mu}_0 &> 0.918 \text{ gives } \hat{p} \geq 0.025 \text{ for } y_0 = 0. \end{aligned} \tag{133}$$

For $\hat{\mu}_0$ in $[0.315, 0.918]$, neither $y_0 = 0$ or $y_0 = 1$ can be rejected at the 0.025 level.

Conclusion

As far as prediction error is concerned, cross-validation and covariance penalties are established subjects backed up by a substantial theoretical and applied literature. Conformal prediction, as the new kid on the block, is still in its formative stage, with at least a promising hint of moving beyond complete reliance on the random sampling model (15). All three approaches rely on resampling methodology, very much in the spirit of statistical inference in the 2020s.

Funding: This research was funded in part by the National Science Foundation grant number DMS 1608182.

References

1. Efron, B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* **1983**, *78*, 316–331.
2. Efron, B.; Hastie, T. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*; Cambridge University Press: Cambridge, 2016; p. 476. Institute of Mathematical Statistics Monographs (Book 5).
3. Efron, B.; Narasimhan, B. The automatic construction of bootstrap confidence intervals. *J. Comput. Graph. Stat.* **2020**, *29*, 608–619. doi:10.1080/10618600.2020.1714633.
4. Efron, B. The estimation of prediction error: Covariance penalties and cross-validation. *J. Amer. Statist. Assoc.* **2004**, *99*, 619–642.
5. Brègman, L.M. A relaxation method of finding a common point of convex sets and its application to the solution of problems in convex programming. *Ž. Vyčisl. Mat i Mat. Fiz.* **1967**, *7*, 620–631.
6. Efron, B.; Tibshirani, R. Improvements on cross-validation: The .632+ bootstrap method. *J. Amer. Statist. Assoc.* **1997**, *92*, 548–560.
7. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32. doi:10.1023/A:1010933404324.
8. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wadsworth Statistics/Probability Series, Wadsworth Advanced Books and Software, Belmont, CA, 1984; pp. x+358.
9. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second ed.; Springer Series in Statistics, Springer, New York, 2009; pp. xxii+745. doi:10.1007/978-0-387-84858-7.

10. Mallows, C.L. Some Comments on C_p. *Technometrics* **1973**, *15*, 661–675. doi:10.2307/1267380.
11. Bates, S.; Hastie, T.; Tibshirani, R. Cross-validation: What does it estimate and how well does it do it?, 2021. arXiv eprint 2104.00673v2.
12. Rosset, S.; Tibshirani, R.J. From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *J. Amer. Statist. Assoc.* **2020**, *115*, 138–162. doi:10.1080/01621459.2018.1424632.
13. Zhang, Y.; Yang, Y. Cross-validation for selecting a model selection procedure. *J. Econometrics* **2015**, *187*, 95–112. doi:10.1016/j.jeconom.2015.02.006.
14. Efron, B. How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **1986**, *81*, 461–470.
15. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer, New York, 2005; pp. xvi+324.
16. Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R.J.; Wasserman, L. Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.* **2018**, *113*, 1094–1111. doi:10.1080/01621459.2017.1307116.
17. Barber, R.F.; Candès, E.J.; Ramdas, A.; Tibshirani, R.J. Predictive inference with the jackknife+. *Ann. Statist.*, *49*, 486–507. doi:10.1214/20-AOS1965.
18. Boström, H.; Linusson, H.; Löfström, T.; Johansson, U. Accelerating difficulty estimation for conformal regression forests. *Ann. Math. Artif. Intell.* **2017**, *81*, 125–144. doi:10.1007/s10472-017-9539-9.
19. Tibshirani, R.J.; Foygel Barber, R.; Candès, E.J.; Ramdas, A. Conformal prediction under covariate shift. Advances in Neural Information Processing Systems 32 (NIPS 2019); Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché Buc, F.; Fox, E.; Garnett, R., Eds. Curran Associates, Inc., 2019, pp. 2526–2536.
20. Sugiyama, M.; Krauledat, M.; Müller, K.R. Covariate Shift Adaptation by Importance Weighted Cross Validation. *J. Mach. Learn. Res.* **2007**, *8*, 985—1005.