Inference Overview

Recall that random variables are characterized by their distributions, which often involve unknown parameters.

- By inference, we mean to use **data** to
  1. **estimate** parameters of interest
  2. find **confidence intervals** for parameters
  3. **test hypotheses** about parameters

- To estimate parameters, we may use method of moments (MME) or maximum likelihood estimator (MLE) etc. Some simple estimators can be intuitively derived, eg. use sample mean to estimate population mean.

- To achieve 2 and 3, we need to know the **distribution** of the test statistic or parameter estimate

- In some cases, we are able to get the **exact** distribution, such as Fisher's Exact Test.

- In other cases, we must use large sample theory to get an **asymptotic**, or approximate, distribution.

We can often use large sample theory to show that some parameter estimates and test statistics are **normally distributed**. In these situations, we really only need to know the MEAN and VARIANCE of the random variables (since the mean and variance completely specify the Normal distribution).

The Normal Distribution

The normal distribution is an example of a **probability distribution** for a continuous random variable.

- It is specified by its **density:**

$$f(y) \;=\; \frac{1}{\sqrt{2\,\pi\,\sigma^2}}\, e^{\frac{-(y-\mu)^2}{2\,\sigma^2}}$$

- There is a whole family of normal distributions, denoted by $N(\mu, \sigma^2)$, specified by values of the parameters μ and σ.

  μ = mean of the population distribution

  σ = standard deviation of the population distribution

- The **Standard Normal Distribution** is defined by $\mu = 0$ and $\sigma = 1$. The density can thus be simplified to:

$$f(y) \;=\; \frac{1}{\sqrt{2\,\pi}}\, e^{\frac{-y^2}{2}}$$

Normal distribution review:

- To calculate the probability that a $N(0,1)$ r.v. Z falls in the interval from a to b, we could use calculus:

$$Pr(a \leq Z \leq b) \;=\; \int_a^b \frac{1}{\sqrt{2\,\pi}}\, e^{\frac{-y^2}{2}} dy$$

- The probabilities $Pr(Z \leq z)$ for various values of z are also given by statistical tables or software [R: pnorm()]. Then we can calculate the above probability such as:

$$Pr(a \leq Z \leq b) \;=\; Pr(Z \leq b) - Pr(Z \leq a)$$

- For normal r.v.'s with mean μ and standard deviation σ, traditionally we use the following **standardization** to calculate probabilities:

$$\begin{aligned} Pr(a \leq Y \leq b) &= Pr\left(\frac{a-\mu}{\sigma} \leq \frac{Y-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &= Pr(a^* \leq Z \leq b^*) \\ &= Pr(Z \leq b^*) - Pr(Z \leq a^*) \end{aligned}$$

  where Z is $N(0,1)$, $a^* = (a-\mu)/\sigma$, and $b^* = (b-\mu)/\sigma$.

  Note: you don't need this step using software.

- All normal distributions have 95% of their area between $(\mu - 1.96\sigma)$ and $(\mu + 1.96\sigma)$.

Large Sample Theory

Central Limit Theorem (CLT) :

Let Y be the sum of n independent, identically distributed (i.i.d.) random variables $Y_1, Y_2, ..., Y_n$:

$$Y = \sum_{i=1}^{n} Y_i$$

Then, for large n,

$$Z = \left(\frac{Y - E(Y)}{\sqrt{\text{Var}(Y)}} \right) \overset{approx}{\sim} N(0, 1),$$

- There are certain "regularity" conditions that must be satisfied, such as
$$0 < \text{Var}(Y_i) < \infty.$$

- Most of the statistical tests we perform are based on the Central Limit Theorem.

Another form of the CLT:

Let $\overline{Y}$ be the sample mean,

$$\overline{Y} = \frac{\Sigma_{i=1}^{n} Y_i}{n},$$

of n i.i.d. random variables $Y_1, Y_2, ..., Y_n$ with

$$E(Y_i) = \mu \quad \text{and} \quad \text{Var}(Y_i) = \sigma^2$$

Then, for large n,

$$Z = \left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \right) \dot{\sim} N(0, 1).$$

Here, we have stated the Central Limit Theorem in terms of the sample mean, instead of the sum.

Example: Binomial Data $Y \sim Bin(n, p)$

- A usual estimator for p (why):

$$\hat{p} = \overline{Y} = \frac{Y}{n} = \frac{\Sigma_{i=1}^{n} Y_i}{n},$$

  where the Y_i are **i.i.d.** Bernoulli random variables.

- We know that the "exact" distribution of Y is

$$Y = n\hat{p} \sim Bin(n, p)$$

  What is $P(Y = k) =$?

- Note that $\hat{p}$ is just the sample mean of the Bernoulli r.v.'s

- To apply CLT, we have n i.i.d Bernoulli r.v.'s with

$$E(Y_i) = \mu = p$$

  and

$$\text{Var}(Y_i) = \sigma^2 = p(1 - p)$$

- Substituting $\overline{Y} = \hat{p}$, $\mu = p$ and $\sigma^2 = p(1-p)$ in the CLT, we get:

$$\begin{aligned} Z &= \frac{\overline{Y} - \mu}{\sqrt{\sigma^2/n}} \\ &= \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \\ &\dot{\sim} N(0,1) \end{aligned}$$

  when n is large.

- In other words, for large n we can say:

$$\hat{p} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

Notes:

- For large n, confidence intervals and test statistics based on the exact distribution,

$$n\hat{p} \sim Bin(n, p)$$

can be cumbersome (computationally intensive)

- Furthermore, they are almost identical to those based on

$$\hat{p} \dot{\sim} N\left(p, \frac{p(1-p)}{n}\right),$$

so the latter is usually used for large n.

- If n is large, and neither p nor $(1-p)$ is close to 0, then the normal approximation works well.

- The closer p or $(1-p)$ is to 0, the worse the normal approximation (you need larger n for the normal approximation to be OK).

- The typical assumption for the normal approximation to be good is

$$0 < p - 3\sqrt{\frac{p(1-p)}{n}} < p + 3\sqrt{\frac{p(1-p)}{n}} < 1.$$

Transformations of r.v. (Delta Method)

When we looked at the CLT, we stated it in terms of both the sum and the sample mean.

Now suppose we want the approximate distribution of the estimated 'logit': $\text{logit}(\hat{p}) = \log\{\hat{p}/(1-\hat{p})\}$. (why)

The Delta Method: Suppose we have an asymptotically normal r.v. Y:

$$Y \dot{\sim} N(\mu, \sigma^2),$$

then

$$g(Y) \dot{\sim} N\left(g(\mu), [g'(\mu)]^2\, \sigma^2\right)$$

- Two regularity conditions are:

  $g(y)$ is differentiable

  $g'(\mu) \neq 0$

Example: The "Logit"

- Suppose $Y \sim B(n, p)$.

- Based on the CLT, we have the following large sample distribution:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

- Suppose we want to find the approximate distribution of the estimated 'logit':

$$g(\hat{p}) = \text{logit}(\hat{p}) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \log(\hat{p}) - \log(1-\hat{p})$$

- Using the Delta Method,

$$g(\hat{p}) \dot{\sim} N\left(g(p),\, [g'(p)]^2\, \frac{p(1-p)}{n}\right)$$

or, equivalently,

$$\text{logit}(\hat{p}) \sim N\left(\text{logit}(p),\ \frac{1}{np(1-p)}\right)$$

Confidence Intervals

- An estimate such as $\hat{p}$ comes with variability or uncertainly [see plot: OI biostat page 176], and we refer to it as a 'point estimate'.

- A confidence interval (CI) for a parameter θ (our general notation) is a **random interval**, computed from the data (hence random), that contains θ with pre-specified probability, eg. 95%.

  Interpretation of CI:
  This relies on the abstract construct of *repeated sampling*; i.e. if we take repeated random samples of the data (say $X_1, ..., X_n$) from the same population, and form a CI from each sample, then about 95% of them should contain θ. [see plot: OI biostat page 181]

- More generally, instead of 95%, we may consider a $100(1-\alpha)$% CI, and $(1-\alpha)$ is sometimes called the coverage probability.

- A CI gives a plausible *range of values* for θ with a *margin of error*.

CATEGORICAL OUTCOMES
Analysis of 2×2 contingency tables

Categorical vs. continuous variables

For many analysis purposes, for example:

- descriptive statistics ('Table 1'):
  - continuous variables are summarized by mean (SD)
  - categorical variables are summarized by count (%)

- regression models:
  - continuous outcomes are often modeled using linear regression;
  - categorical outcomes are often modeled using some type of logistic regression.

Cold incidence among French Skiers:

		OUTCOME			
		COLD	NO COLD	Total	
TREATMENT	VITAMIN C	17	122	139	<-- (fixed by design)
	NO VITAMIN C	31	109	140	<-- (fixed by design)
	Total	48	231	279	

- Number on each treatment fixed by design.
- Usually the design for experimental studies/clinical trials
- Individuals are followed to assess response

Questions:

1. what is the research question of interest?
   what is the outcome?
2. what are the distributions involved?
   (what are the random variables?)
3. what would be your hypotheses?
   (what are the parameters?)

In general, we can form the following 2×2 table:

		Outcome 1	Outcome 2	
TREATMENT	1	Y_1	$n_1 - Y_1$	n_1
	2	Y_2	$n_2 - Y_2$	n_2

- Individuals are given (or sometimes randomized to) treatment 1 or treatment 2

- The measured outcome is success or failure.

Facts about the distribution of outcomes

- n_1 and n_2 are fixed by design

- Y_1 and Y_2 are *independent* with distributions:

$$Y_1 \sim B(n_1, p_1)$$

$$Y_2 \sim B(n_2, p_2)$$

- We want to estimate p_1, p_2 and compare them.

- The *likelihood*, i.e. probability of observed data, is the product of 2 independent binomials:

$$\begin{aligned} L(p_1, p_2) &= P(Y_1 = y_1|p_1)\ P(Y_2 = y_2|p_2) \\ &= \left(\begin{array}{c} n_1 \\ y_1 \end{array} \right) p_1^{y_1}(1-p_1)^{n_1-y_1} \left(\begin{array}{c} n_2 \\ y_2 \end{array} \right) p_2^{y_2}(1-p_2)^{n_2-y_2} \end{aligned}$$

- Research question(s) of interest:
  1. Does treatment affect outcome? (causal)
  2. Are treatment and outcome associated? (observational)
  3. Is the success probability the same on both treatments?

- These are often assessed via a null hypothesis:

$$\mathrm{H}_0\text{: } p_1 = p_2 = p$$

  and an alternative hypothesis

$$\mathrm{H}_A\text{: } p_1 \neq p_2$$

- Two-sided alternatives are often considered more rigorous, because they are harder to reject (**why**).

- We are interested in
  1. describing treatment differences
  2. testing for a treatment effect.

Q: How can we quantify treatment differences?

Measuresof treatment differences

1. **Risk Difference**

$$\Delta = p_1 - p_2, \qquad -1 \leq \Delta \leq 1$$

2. **Relative Risk or Risk Ratio**

$$RR = \frac{p_1}{p_2}, \qquad 0 \leq RR \leq \infty$$

The log-relative risk is often used to get around the restriction that the relative risk must be positive:

$$\log RR = \log\left(\frac{p_1}{p_2}\right) = \log(p_1) - \log(p_2)$$

where

$$-\infty \leq \log RR \leq \infty.$$

3. **Odds Ratio or Relative Odds**

- The odds of success versus failure on treatment i is:

$$\frac{p_i}{(1-p_i)}, \qquad i = 1, 2$$

- The ratio of the odds for treatment 1 to treatment 2 is:

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{p_1(1-p_2)}{p_2(1-p_1)}, \qquad 0 \leq OR \leq \infty,$$

- Again, the log-odds ratio is often used, to avoid the restriction that the odds ratio must be positive, i.e.,

$$\begin{aligned} \log OR &= \log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right) \\ &= \log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) \\ &= \text{logit}(p_1) - \text{logit}(p_2) \end{aligned}$$

  where $-\infty \leq \log OR \leq \infty$

- Note that the $\log(OR)$ is the difference in logits.

Relationship between OR and RR

- Recall, from the definition of an **Odds Ratio**

$$\begin{aligned} OR &= \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \\ &= \left(\frac{p_1}{p_2}\right)\left[\frac{1-p_2}{1-p_1}\right] \\ &= RR\left[\frac{1-p_2}{1-p_1}\right] \end{aligned}$$

- When the disease is **rare:**

$$\left[\frac{1-p_2}{1-p_1}\right] \approx \frac{1}{1} = 1; \quad \text{and so} \quad OR \approx RR.$$

- In the Vitamin C example, $\hat{p}_1$ and $\hat{p}_2$ were 0.12 and 0.22, respectively. These are not small enough to be considered rare, and so the estimated OR, $\widehat{OR} = 2.041$ is not that close to the estimated RR, $\widehat{RR} = 1.811$.

- In the example below, aspirin use and heart attacks, $\hat{p}_1$ and $\hat{p}_2$ are both $< 2\%$, so the estimates of the Odds Ratio and Relative Risk are very similar.
  Ex. verify that $\widehat{OR} = 1.832$ and $\widehat{RR} = 1.818$.

Example: Clinical trial for Aspirin Use and Heart Attack in Doctors

		OUTCOME		
		Heart Attack	NO Heart Attack	Total
TREATMENT	Placebo	189	10845	11034
	Aspirin	104	10933	11037
	Total	293	21778	22071

Questions for thought

- Write down the estimates of risk difference, risk ratio, and odds ratio.

- What do we need in order to make inference?

Variance of a treatment difference, in general

- Our treatment differences can be written

$$\theta = g(p_1) - g(p_2).$$

  The estimator is then

$$\hat{\theta} = g(\hat{p}_1) - g(\hat{p}_2)$$

  (Recall that the MLE of $g(\beta)$ is $g(\hat{\beta})$, where $\hat{\beta}$ is MLE of β.)

  Q: What is g in the above for risk difference, risk ratio, and odds ratio?

- Also, since $\hat{p}_1$ and $\hat{p}_2$ are independent (**why**), so are any functions of $\hat{p}_1$ and $\hat{p}_2$. Therefore

$$\text{Var}[g(\hat{p}_1) - g(\hat{p}_2)] = \text{Var}[g(\hat{p}_1)] + \text{Var}[g(\hat{p}_2)]$$

- **Q:** What is $\text{Var}[g(\hat{p}_i)]$?

Example: logit

- We know that
$$\hat{p} = \frac{Y}{n}$$
-
$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$
-
$$g(\hat{p}) = \text{logit}(\hat{p}) = \log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \log(\hat{p}) - \log(1-\hat{p})$$
-
$$\begin{aligned}\frac{\partial[\log(p) - \log(1-p)]}{\partial p} &= \frac{1}{p} - \frac{-1}{1-p} \\ &= \frac{1}{p(1-p)}\end{aligned}$$
- By Delta method, $\text{Var}[g(\hat{p})]$ is approximately
$$\begin{aligned}[g'(\mu)]^2\ \sigma^2 &= \left[\frac{1}{p(1-p)}\right]^2 \frac{p(1-p)}{n} \\ &= \frac{1}{np(1-p)}\end{aligned}$$

The results are summarized in the following table:

Treatment Difference	Estimate	Var(Estimate)
Δ	$\widehat{p}_1 - \widehat{p}_2$	$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
log (RR)	$\log\left(\frac{\widehat{p}_1}{\widehat{p}_2}\right)$	$\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}$
log (OR)	$\log\left(\frac{\widehat{p}_1/(1-\widehat{p}_1)}{\widehat{p}_2/(1-\widehat{p}_2)}\right)$	$\frac{1}{n_1 p_1(1-p_1)} + \frac{1}{n_2 p_2(1-p_2)}$

Ex. Derive these variances if you have not done it before.

We use the log scale, because

- it gets rid of restrictions on the ranges of the parameters, and makes computation a lot easier.
- Also, the **normal approximation** works better.

Confidence Intervals

A 95% **confidence interval** for a parameter β is (L, U), so that

$$P(L < \beta < U) = 0.95,$$

where U and L are calculated from the data. This is the so-called 'interval estimate' ($\hat{\beta}$ is called point estimate).

To find confidence intervals, we often use the fact that for $Y \sim N(\mu, \sigma^2)$,

$$\begin{aligned} & P(-1.96 < \frac{Y - \mu}{\sigma} < 1.96) \\ = \; & P(Y - 1.96\sigma < \mu < Y + 1.96\sigma) = 0.95 \end{aligned}$$

- First of all, we need to estimate the variances – we replace p_1 and p_2 in Var(Estimate) with $\hat{p}_1$ and $\hat{p}_2$.

- Then the 95% confidence intervals for treatment differences can be obtained as

$$(\hat{p}_1 - \hat{p}_2) \ \pm \ 1.96 \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$\log(\widehat{RR}) \ \pm \ 1.96 \sqrt{\frac{1-\hat{p}_1}{n_1\hat{p}_1} + \frac{1-\hat{p}_2}{n_2\hat{p}_2}}$$

$$\log(\widehat{OR}) \ \pm \ 1.96 \sqrt{\frac{1}{y_1} + \frac{1}{n_1 - y_1} + \frac{1}{y_2} + \frac{1}{n_2 - y_2}}$$

Ex. verify the last one.

Confidence Intervals for OR and RR

- You might want a confidence interval for RR or OR instead of $\log RR$ or $\log OR$.

- **Q:** what would they be?

Hypothesis Testing

Q: What is the null hypothesis?

Q: What would be a test statistic?

- Under the **null** $\mathrm{H}_0 : p_1 = p_2$, we know that

$$(1) \quad p_1 - p_2 = 0$$

$$(2) \quad \log(RR) = 0$$

$$(3) \quad \log(OR) = 0$$

- Then to test this null hypothesis (against either one or two-sided alternatives), we can use any of the following statistics:

$$(1) \quad Z_1 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\widehat{\mathrm{Var}}(\hat{p}_1 - \hat{p}_2)}}$$

$$(2) \quad Z_2 = \frac{\log(\widehat{RR})}{\sqrt{\widehat{\mathrm{Var}}[\log(\widehat{RR})]}}$$

$$(3) \quad Z_3 = \frac{\log(\widehat{OR})}{\sqrt{\widehat{\mathrm{Var}}[\log(\widehat{OR})]}}$$

- All three are approximately $N(0,1)$ **under the null**.

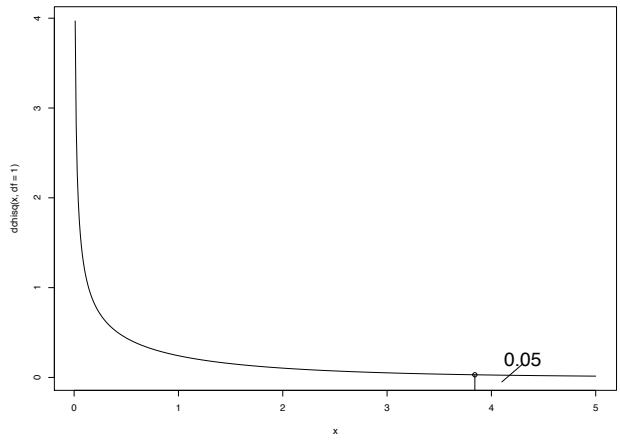
- Suppose that we use Z_i, then for a two-sided test at 0.05 **significance level**, we would reject the null if $|Z_i| > 1.96$

- Note that $|Z_i| > 1.96$ is equivalent to the fact that the corresponding 95% confidence interval does not contain treatment difference 0.

- $Z_1^2 \overset{asymp.}{\sim} \chi_1^2$ under the null gives the *chi-squared test* for a 2×2 contingency table. [R: chisq.test()]

```
> FrenchSkier <-
+ matrix(c(17, 31, 122, 109),
+        nrow = 2,
+        dimnames = list(Treatment = c("V_c", "No V_c"),
+                        Outcome = c("Cold", "No Cold")))

> FrenchSkier
         Outcome
Treatment Cold No Cold
   V_c      17     122
   No V_c   31     109

> chisq.test(FrenchSkier)

Pearson's Chi-squared test with Yates' continuity correction

data:  FrenchSkier
X-squared = 4.1407, df = 1, p-value = 0.04186
```

Q: What is the cutoff for χ^2_1 and significance level 0.05?
Is it one- or two-sided?



Fisher's exact test

		Outcome 1	Outcome 2	
Treatment	1	Y_1	$n_1 - Y_1$	n_1
	2	Y_2	$n_2 - Y_2$	n_2
		$Y_1 + Y_2 = Y$	$n - Y$	$n_1 + n_2 = n$

- When data are sparse, i.e. the *expected* count of any cell is < 5, Fisher's exact test is typically used.
  - Here expected refers to under the null hypothesis that $p_1 = p_2 = p$, $\widehat{E}(Y_i) = n_i\hat{p}$ where $\hat{p} = (Y_1 + Y_2)/(n_1 + n_2)$;
  - Exact inference is based on the fact that, given all 4 marginal totals $(n_1, n_2, Y, n - Y)$ fixed, the first element Y_1 of the contingency table has a *hypergeometric* distribution under the null (Fisher, 1935):

$$P(Y_1 = k) = \frac{\binom{n_1}{k}\binom{n_2}{Y-k}}{\binom{n}{Y}}$$

- R: fisher.test()

```
## Agresti (1990, p. 61f; 2002, p. 91) Fisher's Tea Drinker
## A British woman claimed to be able to distinguish whether milk or
##  tea was added to the cup first.  To test, she was given 8 cups of
##  tea, in four of which milk was added first.  The null hypothesis
##  is that there is no association between the true order of pouring
##  and the woman's guess, the alternative that there is a positive
##  association (that the odds ratio is greater than 1).

TeaTasting <- matrix(c(3, 1, 1, 3), nrow = 2,
       dimnames = list(Guess = c("Milk", "Tea"),
                       Truth = c("Milk", "Tea")))

> TeaTasting
      Truth
Guess  Milk Tea
  Milk    3   1
  Tea     1   3

> fisher.test(TeaTasting)

Fisher's Exact Test for Count Data

data:  TeaTasting
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.2117329 621.9337505
sample estimates:
odds ratio
  6.408309
```

Q: what is the conclusion here?

$r \times c$ **contingence tables**

Contingence tables can be extended to more than 2 categories for each of the two variables.

- It will no longer be comparing two probabilities.

- There is still chi-squared test for association of the two categorical variables.

- There is also Fisher's exact test when the expected count of any cell is < 5, under the null hypothesis that the two variables are independent.

```
## A r x c table  Agresti (2002, p. 57) Job Satisfaction

Job <- matrix(c(1,2,1,0, 3,3,6,1, 10,10,14,9, 6,7,12,11), 4, 4,
dimnames = list(income = c("< 15k", "15-25k", "25-40k", "> 40k"),
satisfaction = c("VeryD", "LittleD", "ModerateS", "VeryS")))

> Job
        satisfaction
income   VeryD LittleD ModerateS VeryS
  < 15k      1       3        10     6
  15-25k     2       3        10     7
  25-40k     1       6        14    12
  > 40k      0       1         9    11

> fisher.test(Job)

Fisher's Exact Test for Count Data

data:  Job
p-value = 0.7827
alternative hypothesis: two.sided
```

Logistic Regression for a 2×2 table

	Outcome (Y)		
Treatment(X)	1	0	Total
1	Y_1	$n_1 - Y_1$	n_1
0	Y_0	$n_0 - Y_0$	n_0

- Previously, the second row of the table had subscripts of "2", which are now changed to subscripts of "0".

- We considered Y_1 and Y_0 as two separate variables, each of which followed a binomial distribution.

- Now we will define a binary variable X for treatment assignment, with values
  - 0 for placebo or standard treatment
  - 1 for new treatment

- Similarly, we will define a binary variable Y for outcome, with values
  - 0 for failure/no response
  - 1 for success/response

- The number of successes on the new treatment ($X = 1$) is Y_1, with success probability p_1

- The number of successes on the placebo ($X = 0$) is Y_0, with success probability p_0

- We can also write the success probabilities in terms of the conditional probabilities of Y given X:

$$P(Y = 1|X = 1) = p_1$$

$$P(Y = 1|X = 0) = p_0$$

Introduction to Logistic Models

The logistic regression model for the probability of success is

$$\text{pr}[Y = 1|X = x] \quad = \quad p_x \quad = \quad \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$$

where β_0 and β_1 are parameters, and $x = 0$ or 1. This model can also be written in terms of the logit:

$$\text{logit}(p_x) \quad = \quad \beta_0 + \beta_1 x$$

- If $x = 1$, then

$$\text{pr}[Y = 1|X = 1] = p_1 = \frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}}$$

- If $x = 0$, then

$$\text{pr}[Y = 1|X = 0] = p_0 = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$

- A test of equality of the success probabilities on the two treatments, is therefore equivalent to a test of $\beta_1 = 0$, i.e.,

$$\mathrm{H}_0 : p_1 = p_0 \;\Leftrightarrow\; \mathrm{H}_0 : \beta_1 = 0$$

- We will show that

$$\beta_1 = \log(OR).$$

Properties of the Logistic Regression Model

- The parameters have no restrictions,

$$-\infty < \beta_0 < \infty$$

$$-\infty < \beta_1 < \infty$$

  and for $x = 0, 1$,

$$0 < \left(\frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \right) < 1 \qquad \Rightarrow \qquad 0 < p_x < 1$$

- p_x is the probability of success on treatment x, but to compute the odds we also need to know the failure probability:

$$\begin{aligned} \text{pr}[Y = 0|X = x] &= 1 - p_x \\ &= 1 - \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \\ &= \frac{1 + e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} - \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \\ &= \frac{1}{1 + e^{\beta_0+\beta_1 x}} \end{aligned}$$

The Odds for Treatment x

- The **ODDS** for treatment x ($x = 0$ or 1) is equal to:

$$\begin{aligned} \frac{p_x}{1-p_x} &= \frac{[e^{\beta_0+\beta_1 x}]/[1+e^{\beta_0+\beta_1 x}]}{1/[1+e^{\beta_0+\beta_1 x}]} \\ &= e^{\beta_0+\beta_1 x} \end{aligned}$$

- In many cases, we are interested in the **logit**, or log-odds:

$$\begin{aligned} \text{logit}(p_x) &= \log\left(\frac{p_x}{1-p_x}\right) \\ &= \log\left(e^{\beta_0+\beta_1 x}\right) = \beta_0 + \beta_1 x \end{aligned}$$

- Based on the general formula above, we get:

$$\begin{aligned} \text{logit}(p_1) &= \beta_0 + \beta_1 \cdot 1 = \beta_0 + \beta_1 \\ \text{logit}(p_0) &= \beta_0 + \beta_1 \cdot 0 = \beta_0 \end{aligned}$$

The Odds Ratio from a Logistic Model

- The log-odds ratio is the difference in logits:

$$\begin{aligned} \log(OR) &= \text{logit}(p_1) - \text{logit}(p_0) \\ &= [\beta_0 + \beta_1] - \beta_0 \\ &= \beta_1 \end{aligned}$$

- Equivalently, the **odds ratio** for the new treatment versus placebo ($x = 1$ versus $x = 0$) is therefore:

$$OR = e^{\beta_1}$$

Example: Lung cancer and smoking

- Suppose you are comparing lung cancer and smoking with

$$X = \begin{cases} 1 \text{ if ever smoked (row 1)} \\ 0 \text{ if never smoked (row 2)} \end{cases}$$

$$Y = \begin{cases} 1 \text{ if lung cancer (column 1)} \\ 0 \text{ if no lung cancer (column 2)} \end{cases}$$

- Then we have the following logits for each outcome:

$$\text{logit}(p_1) = \beta_0 + \beta_1$$

$$\text{logit}(p_0) = \beta_0$$

- The log-odds ratio for lung cancer for smokers versus non-smokers is

$$\begin{aligned} \beta_1 &= \text{logit}(p_1) - \text{logit}(p_0) \\ &= \log[OR(\text{smoke} : \text{no smoke})] \end{aligned}$$

Exact test :

Let us consider the problem of testing the value of the parameter p for a binomial random variable with $n = 10$ trials. We wish to test

$$H_0: p = .5$$

versus

$$H_A: p > .5$$

We will use the number of successes, X, as a test statistic; the rejection region will consist of large values of X, those values that are relatively unlikely under H_0 and more likely under H_A. To determine the precise rejection region for a given value of α, we can use this table of cumulative binomial probabilities $[P(X \leq n)]$:

p	0	1	2	3	4	5	6	7	8	9	10
.7	.0000	.0001	.0016	.0106	.0474	.1503	.3504	.6172	.8507	.9718	1.0000
.6	.0001	.0017	.0123	.0548	.1662	.3669	.6177	.8327	.9536	.9940	1.0000
.5	.0010	.0107	.0547	.1719	.3770	.6230	.8281	.9453	.9893	.9990	1.0000

Suppose that the rejection region consists of the points $\{8, 9, 10\}$. The significance level of the test, α, is the probability of rejecting H_0 when it is true; from the last row of the table ($p = .5$), we see that

$$\alpha = P(X > 7) = 1 - P(X \leq 7) = .0547$$

If the rejection region consists of $\{7, 8, 9, 10\}$, the significance level of the test is $\alpha = .172$.

The Neyman–Pearson approach sets a value for α first; suppose that we choose to set $\alpha = .0547$. If the true value of p is .6, the power of the test is the probability that X is greater than or equal to 8; that is, the power is .1673. If the true value is .7, the power is .3828. The power is thus a function of p, and it is not difficult to see that the power tends to 1 as p approaches 1 and that the power tends to α as p approaches .5. □

researcher wishes to test

$$H_0: p = 0.85$$
versus
$$H_1: p \neq 0.85$$

The decision will be based on the magnitude of k, the total number in the sample for whom the durg is effective—that is, on

$$k = k_1 + k_2 + \ldots + k_{19}$$

where

$$k_i = \begin{cases} 0 & \text{if the new drug fails to relieve } i\text{th patient's pain} \\ 1 & \text{if the new drug does relieve } i\text{th patient's pain} \end{cases}$$

What should the decision rule be if the intention is to keep α somewhere near 10%? [Note that Theorem 6.3.1 does not apply here because Inequality 6.3.1 is not satisfied—specifically, $np_o + 3\sqrt{np_o(1 - p_o)} = 19(0.85) + 3\sqrt{19(0.85)(0.15)} = 20.8$ is not less than $n(= 19)$.]

If the null hypothesis is true, the expected number of successes would be $np_o = 19(0.85)$, or 16.2. It follows that values of k to the extreme right or extreme left of 16.2 should constitute the critical region.

```
MTB > pdf;
SUBC > binomial 19 0.85.
```

Probability Density Function

```
Binomial with n = 19 and p = 0.850000

      X        P(X = x)
      8        0.0000
      9        0.0001
     10        0.0007
     11        0.0032      → P(X ≤ 13) = 0.0536
     12        0.0122
     13        0.0374
     14        0.0907
     15        0.1714
     16        0.2428
     17        0.2428
     18        0.1529
     19        0.0456      → P(X = 19) = 0.0456
```

For significance level 0.1, the rejection region is $X<=13$ or $X=19$.

FIGURE 6.3.1

Hypothesis Testing

Basic idea:

1. make (given) an assumption about the underlying distribution of data

   eg. 1) $\mathrm{N}(\mu, \sigma^2)$: $\mu = 1$

   2) two samples from $\mathrm{N}(\mu_1, \sigma^2)$ and $\mathrm{N}(\mu_2, \sigma^2)$: $\mu_1 > \mu_2$

2. look at the data, and decide if the data are consistent with the assumption

This process is formulated as statistical hypothesis testing.

Neyman-Pearson Paradigm

In the framework of statistical hypothesis testing, we make a choice between two mutually exclusive hypotheses:

1. *null* hypothesis H_0, eg. $\mu = 1$, or $\mu_1 = \mu_2$
2. *alternative* hypothesis H_1 (some books use H_A)

- According to H_1, we have
  - one-sided test, eg. H_1: $\mu > 1$, or $\mu_1 < \mu_2$
  - two-sided test, eg. H_1: $\mu \neq 1$, or $\mu_1 \neq \mu_2$

- There is *asymmetry* between H_0 and H_1: we want to answer whether we can reject H_0 or not.

- A decision as to whether or not to reject H_0 in favor of H_1, is made based on the value of a test statistic T.

- The set of values of T for which H_0 is rejected, is called the *rejection region*;

- it is often of the form $T > C$ say, where C is called the *critical value*.

- How to choose C is based on error probabilities.

Two types of error

There are two types of error:

1. type I error: H_0 is true but rejected;
2. type II error: H_1 is true, but H_0 is *not rejected*.

The corresponding error rates are:

- $\alpha = P_{H_0}(\text{“}H_1\text{”}) = P(\text{type I error})$, also called *significance level*, or the *size* of a test;
- $\beta = P_{H_1}(\text{“}H_0\text{”}) = P(\text{type II error})$;
- $1 - \beta = P_{H_1}(\text{“}H_1\text{”})$ is called the *power* of the test.

Truth	Fail to reject H_0	Reject H_0
H_0	no error (1-α)	type I error (α)
H_1	type II error (β)	no error (1-β)

The critical value C is chosen to meet some pre-specified α.

Example

In phase II clinical trials for cancer treatment, we often study the response rate of a drug, i.e. if it shrinks a solid tumor by certain amount. This response rate is often compared to the rate that is achieved by the current standard therapy.

- Data: $X_1, ..., X_n$ from n patients, $X_i = 1$ if patient i achieved response, 0 otherwise.

- Distribution or model: $\sum X_i \sim B(n, p)$

- H_0: $p \leq 0.2$ or $p = 0.2$, where 20% is the response rate achieved by standard therapy

- H_0: $p > 0.2$

- **Q:** what would be a test statistic?

We may use $T = \sum X_i/n = \widehat{p}$

- reject H_0 if $\widehat{p} >$ some value C

- we want $\alpha = P_{p=0.2}(\widehat{p} > C)$

- What is the distribution of $\widehat{p}$ when $p = 0.2$?

$$\widehat{p} \overset{asymp.}{\sim} N(0.2,\, 0.2 \times 0.8/n)$$

- So

$$\begin{aligned} \alpha &= P_{p=0.2}(\widehat{p} > C) \\ &= P\left(\frac{\widehat{p} - 0.2}{\sqrt{0.16/n}} > C_1 = \frac{C - 0.2}{\sqrt{0.16/n}}\right) \\ &\approx P(Z > C_1) \end{aligned}$$

where $Z \sim N(0, 1)$.

- If $\alpha = 0.05$, then $C_1 = 1.65$, because $P(Z > 1.65) = 0.05$.

- Then $C = 0.2 + 0.66/\sqrt{n}$.

- Therefor reject H_0 if $\widehat{p} > 0.2 + 0.66/\sqrt{n}$.

These are mathematical derivations, which provide a *decision rule* that can be applied to any n, and any $\widehat{p}$.

One can think of this as what goes on inside a software.

```
> prop.test(x=40, n=100, p=0.2, alternative="greater")

1-sample proportions test with continuity correction

data:  40 out of 100, null probability 0.2
X-squared = 23.766, df = 1, p-value = 5.44e-07
alternative hypothesis: true p is greater than 0.2
95 percent confidence interval:
 0.3183752 1.0000000
sample estimates:
  p
0.4
```

p-value

Suppose that T is the test statistic, and we reject H_0 if $T > C$.

Now given data $X_1, ..., X_n$, we have calculated the value of T to be T_{obs}. Consider T_{obs} to be a fixed value for now.

p-value $= P_{H_0}(T \geq T_{obs})$.

- p-value is a measure of evidence against H_0; i.e. under H_0, how extreme is the observed data in the direction of H_1.

Eg. (cont'd) Previously we derived the test to reject $H_0 : p = 0.2$ if $\widehat{p} > 0.2 + 0.66/\sqrt{n}$.

Suppose $n = 25$, and $\widehat{p} = 0.3$. **Ex.** do we reject H_0 based on the decision rule above?

Now

$$\begin{aligned} p - \text{value} &= P_{H_0}(\widehat{p} \geq 0.3) \\ &= P\left(Z > \frac{0.3 - 0.2}{0.4/\sqrt{25}} = 1.25\right) \\ &= 0.106 \end{aligned}$$

How exactly do we use p-value?

Note that:

$$\begin{aligned} p - \text{value} < \alpha &\Leftrightarrow P_{H_0}(T \geq T_{obs}) < P_{H_0}(T > C) \\ &\Leftrightarrow T_{obs} > C \\ &\Leftrightarrow \text{reject } H_0 \text{ at } \alpha \text{ level.} \end{aligned}$$

What would be the conclusion of our example?

Note also

- p-value is a random variable, as it is a function of the data;
- it can be shown that, under H_0 the p-value has a Uniform (0, 1) distribution.

Summary of procedure for hypothesis testing

- Based on the research question, set up null and alternative hypotheses, and probabilities of error;
- choose an appropriate statistical model;
- find a test statistic;
- find the null distribution of the test statistic;
- find the rejection region or critical value based on the type I error rate α;
- for data analysis, compute the test statistic or p-value based on data, and decide whether H_0 is rejected;
- for study design, compute the sample size n based on desired power $1 - \beta$.

How to find a test statistic?

- Should have distinguishable values under H_0 versus H_1; eg. tends to be around zero under H_0, and have large values under H_1.

- Need to know its null distribution
  - exact
  - asymptotic

- To develop a test statistic is often a topic of statistical research.

Duality between CI and Hypothesis Test

For a parameter θ in general, $H_0 : \theta = \theta_0$

- for a two-sided alternative $H_1 : \theta \neq \theta_0$, H_0 is rejected if and only if the $100(1-\alpha)\%$ CI does not contain θ_0.

- Similar duality holds for one-sided alternatives. One-sided CI is also referred to as lower/upper confidence bounds:
  - for $H_1 : \theta > \theta_0$, H_0 is rejected if and only if the $100(1-\alpha)\%$ CI of the form (L, ∞) does not contain θ_0.
  - for $H_1 : \theta < \theta_0$, H_0 is rejected if and only if the $100(1-\alpha)\%$ CI of the form $(-\infty, U)$ does not contain θ_0.

Exact Tests

Study questions (refer to the texts for the each of the two examples):

- What are the hypotheses?
- What is the significance level α? one- or two-sided?
- What is the test statistic?
- What is the decision rule corresponding to the α above?
- How is the decision rule derived?
- Can you derive a decision rule for $\alpha = 0.05$?

4.1 Variability in estimates

A natural way to estimate features of the population, such as the population mean weight, is to use the corresponding summary statistic calculated from the sample.[6] The mean weight in the sample of 60 adults in `cdc.samp` is $\bar{x}_{\text{weight}} = 173.3$ lbs; this sample mean is a **point estimate** of the population mean, μ_{weight}. If a different random sample of 60 individuals were taken from `cdc`, the new sample mean would likely be different as a result of **sampling variation**. While estimates generally vary from one sample to another, the population mean is a fixed value.

Guided Practice 4.1 How would one estimate the difference in average weight between men and women? Given that $\bar{x}_{\text{men}} = 185.1$ lbs and $\bar{x}_{\text{women}} = 162.3$ lbs, what is a good point estimate for the population difference?[7]

Point estimates become more accurate with increasing sample size. Figure 4.3 shows the sample mean weight calculated for random samples drawn from `cdc`, where sample size increases by 1 for each draw until sample size equals 500. The red dashed horizontal line in the figure is drawn at the average weight of all adults in `cdc`, 169.7 lbs, which represents the population mean weight.[8]

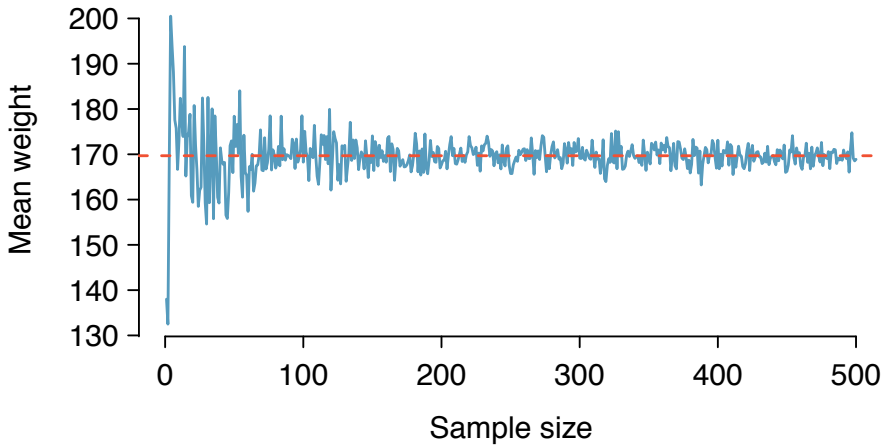


Figure 4.3: The mean `weight` computed for a random sample from `cdc`, increasing sample size one at a time until $n = 500$. The sample mean approaches the population mean (i.e., mean weight in `cdc`) as sample size increases.

Note how a sample size around 50 may produce a sample mean that is as much as 10 lbs higher or lower than the population mean. As sample size increases, the fluctuations around the population mean decrease; in other words, as sample size increases, the sample mean becomes less variable and provides a more reliable estimate of the population mean.

[6] Other population parameters, such as population median or population standard deviation, can also estimated using sample versions.

[7] Given that $\bar{x}_{\text{men}} = 185.1$ lbs and $\bar{x}_{\text{women}} = 162.3$ lbs, the difference of the two sample means, $185.1 - 162.3 = 22.8$lbs, is a point estimate of the difference. The data in the random sample suggests that adult males are, on average, about 23 lbs heavier than adult females.

[8] It is not exactly the mean weight of all US adults, but will be very close since `cdc` is so large.

4.2 Confidence intervals

4.2.1 Interval estimates for a population parameter

While a point estimate consists of a single value, an interval estimate provides a plausible range of values for a parameter. When estimating a population mean μ, a **confidence interval** for μ has the general form

$$(\bar{x} - m,\ \bar{x} + m) = \bar{x} \pm m,$$

where m is the **margin of error**. Intervals that have this form are called **two-sided confidence intervals** because they provide both lower and upper bounds, $\bar{x} - m$ and $\bar{x} + m$, respectively. One-sided sided intervals are discussed in Section 4.2.3.

The standard error of the sample mean is the standard deviation of its distribution; additionally, the distribution of sample means is nearly normal and centered at μ. Under the normal model, the sample mean $\bar{x}$ will be within 1.96 standard errors (i.e., standard deviations) of the population mean μ approximately 95% of the time.[9] Thus, if an interval is constructed that spans 1.96 standard errors from the point estimate in either direction, a data analyst can be 95% **confident** that the interval

$$\bar{x} \pm 1.96 \times \text{SE} \tag{4.2}$$

contains the population mean. The value 95% is an approximation, accurate when the sampling distribution for the sample mean is close to a normal distribution. This assumption holds when the sample size is sufficiently large (guidelines for 'sufficiently large' are given in Section 4.4).

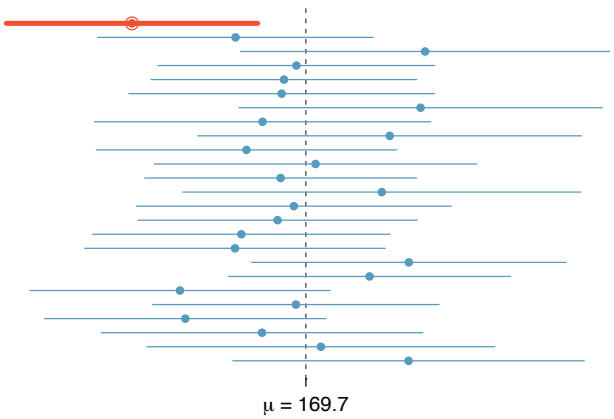


Figure 4.7: Twenty-five samples of size $n = 60$ were taken from cdc. For each sample, a 95% confidence interval was calculated for the population average adult weight. Only 1 of these 25 intervals did not contain the population mean, $\mu = 169.7$ lbs.

The phrase "95% confident" has a subtle interpretation: if many samples were drawn from a population, and a confidence interval is calculated from each one using Equation 4.2, about 95% of those intervals would contain the population mean μ. Figure 4.7

[9] In other words, the Z-score of 1.96 is associated with 2.5% area to the right (and Z = -1.96 has 2.5% area to the left); this can be found on normal probability tables or from using statistical software.

If we denote the interval (L, U), then

$$P(L < \theta < U) = 1-\alpha.$$

(L, U: random; θ: deterministic)

How to construct a c.i.

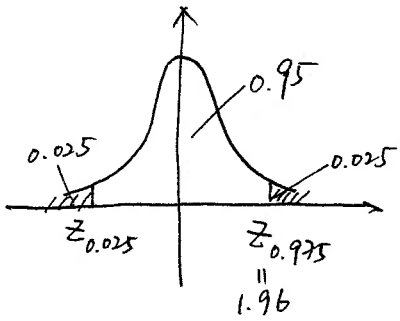
Eg 1. $X_1, \cdots X_n \sim N(\mu, \sigma^2)$ iid, σ^2 known

Fact: $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$ ind

then $X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$

Exercise: use the above to show that $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Then $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$



$$P(|Z| < 1.96) = 0.95$$

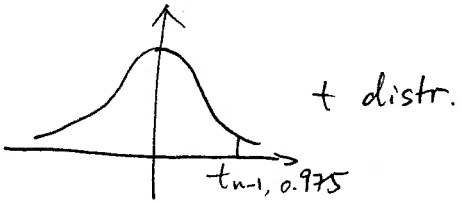
$$= P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right)$$

$$= P\left(\bar{X} - 1.96\sigma/\sqrt{n} < \mu < \bar{X} + 1.96\sigma/\sqrt{n}\right)$$

Sometimes write the c.i: $\bar{X} \pm 1.96\sigma/\sqrt{n}$

Eg 2. $X_1, \cdots X_n \sim N(\mu, \sigma^2)$ iid, σ^2 unknown

Fact: $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$



Eg 3. $X_1, \cdots X_n$ iid Poisson (λ) $(EX = \mathrm{Var}X = \lambda)$

then $\dfrac{\bar{X}-\lambda}{\sqrt{\lambda/n}} \to N(0,1)$

To find a 95% c.i. for λ:

① est. λ/n by $\bar{X}/n$, then c.i $\bar{X} \pm 1.96\sqrt{\bar{X}/n}$

eg. study involving asbestos fiber counts

we had $\hat{\lambda} = \bar{X} = 24.9$. $n = 23$

$s_{\hat{\lambda}} = \sqrt{\bar{X}/n} = 1.04$.

$\hat{\lambda} \pm 1.96 s_{\hat{\lambda}}$ is $(22.9, 26.9)$

② $P\left(\left|\dfrac{\bar{X}-\lambda}{\sqrt{\lambda/n}}\right| < 1.96\right) \approx 0.95$

$= P\left(\dfrac{\bar{X} - 2\bar{X}\lambda + \lambda^2}{\lambda/n} < 1.96^2\right)$

$= P\left(\lambda^2 - (2\bar{X} + 1.96^2/n)\lambda + \bar{X} < 0\right)$

solve for L, U.

③ $\lambda > 0$, but previous c.i.'s may contain values < 0

Try to get a c.i. for $\log \lambda$:

$$P(L_1 < \log\lambda < U_1) = 0.95$$

$$= P(e^{L_1} < \lambda < e^{U_1})$$

Delta method

It can be shown that $\log\bar{X} \overset{\text{approx.}}{\sim} N(\log\lambda, \frac{1}{n\lambda})$

then $\dfrac{\log\bar{X} - \log\lambda}{\sqrt{1/n\lambda}} \overset{\text{approx.}}{\sim} N(0,1)$

λ est'd by $\bar{X}$

so 95% c.i for $\log\lambda$: $\log\bar{X} \pm 1.96\sqrt{\frac{1}{n\bar{X}}}$

" " " λ : $\bar{X}e^{\pm \frac{1.96}{\sqrt{n\bar{X}}}}$

④ Try to find a transformation of $\bar{X}$, $g(\bar{X})$, so that Var $g(\bar{X})$ does not depend on λ —— variance stabilizing transformation

Fact: for large n, $\sqrt{\bar{X}} \overset{\text{approx}}{\sim} N(\sqrt{\lambda}, \frac{1}{4n})$.

then another c.i. for λ is

$$\left(\sqrt{\bar{X}} \pm \frac{1.96}{\sqrt{4n}}\right)^2 = \left(\bar{X} + \frac{1.96^2}{4n}\right) \pm 1.96\sqrt{\frac{\bar{X}}{n}}$$

Suppose that we want to study the coverage property of 95% c.i $\bar{X} \pm 1.96\sqrt{\bar{X}/n}$ for λ in the Poisson distribution. (for fixed n, λ)

1) Generate a sample $X_1, \cdots X_n$ from Poisson(λ) distn.

2) Calculate $\bar{X} \pm 1.96\sqrt{\bar{X}/n}$, see if it contains λ

3) Repeat 1) & 2) N times, get the percentage p to the c.i's that contain λ

4) p should be close to 95% (nomial level) in order to conclude that the coverage is good.

Usually we need to do such studies for a variety of values of n, λ & α (significance level)

GENERAL LOGISTIC REGRESSION

So far, we've discussed logistic regression for 2×2 tables as a special case, i.e. the response is binary, and the predictor (regressor) is also binary.

What are possible extensions of the model?

- Continuous covariates as predictors, binary response
  $\Longrightarrow$ multiple logistic regression modeling

- more than 2 response levels ($R \times C$ tables for example)
  - nominal responses (multinomial logistic regression)
  - ordinal responses (ordinal logistic regression)

General Logistic Regression Modeling

We will now extend our logistic regression models to allow multiple covariates, of any type (nominal, ordinal, or continuous)

- In general, we consider a binary response Y_i for the i^{th} individual, and a general vector of covariates:

$$\mathbf{x}_i = [x_{i1}, ..., x_{iK}]'$$

  where x_{ik} is the k^{th} covariate for individual i.

- In the most general case, the x_{ik}'s can represent a combination of both continuous or categorical covariates.

- **ICU Mortality study** (Hosmer & Lemeshow, *Applied Logistic Regression*)

Y	Mortality (died: $Y = 1$, lived: $Y = 0$)
X_1	sex
X_2	age
X_3	level of consciousness (1=normal, 2=stupor, 3=coma)
X_4	race (1=white, 2=black, 3=other)

Another Example – Arthritis Clinical Trial

- This example is from an arthritis clinical trial comparing the drug auranofin to placebo for treatment of rheumatoid arthritis (Bombardier et al., 1986).

- The response of interest is the self-assessment of arthritis, classified as (0) poor or (1) good.

- Individuals were also given a self-assessment at baseline (before treatment), which was also classified as (0) poor or (1) good.

- To make sure that the treatment groups were balanced with respect to baseline status, the randomization occurred after the baseline measurement was taken

- The dataset contains 293 patients who were observed at both baseline and 13 weeks. The data from 25 cases are shown below:

Subset of cases from the arthritis clinical trial

				Self assessment[b]	
CASE	SEX	AGE	TREATMENT[a]	BASELINE	13 WK.
1	M	54	A	0	0
2	M	64	P	0	0
3	M	48	A	1	1
4	F	41	A	1	1
5	M	55	P	1	1
6	M	64	A	1	1
7	M	64	P	1	0
8	F	55	P	1	1
9	M	39	P	1	0
10	F	60	A	0	1
11	M	49	A	0	1
12	M	32	A	0	1
13	F	62	P	0	0
14	M	50	A	0	1
15	M	54	A	0	0
16	M	36	P	1	1
17	M	63	A	1	1
18	F	63	P	0	0
19	M	65	A	1	0
20	M	60	P	1	1
21	F	59	P	1	1
22	M	57	P	1	1
23	M	58	A	0	1
24	F	35	P	1	1
25	F	31	P	0	1

[a] A = Auranofin, P = Placebo

[b] 0=poor, 1=good.

- We are interested in seeing how the binary response

$$Y_i = \begin{cases} 1 \text{ if good at 13 weeks} \\ 0 \text{ if poor at 13 weeks} \end{cases}$$

is affected by the covariates:

1. BASELINE self-assessment:

$$X_i = \begin{cases} 1 \text{ if good at BASELINE} \\ 0 \text{ if poor at BASELINE} \end{cases}$$

2. GENDER

$$\text{SEX} = \begin{cases} 1 \text{ if male} \\ 0 \text{ if female} \end{cases}$$

3. TREATMENT

$$\text{TRT} = \begin{cases} 1 \text{ if auranofin} \\ 0 \text{ if placebo} \end{cases}$$

4. AGE IN YEARS

- The main question is whether the treatment increases the probability of a more favorable response, after controlling for baseline response, age and sex. Secondary questions might be how age and sex affect the probability of response.

Distribution of Response Outcomes

- Since each individual may represent a unique combination of covariates, we no longer count up all those responding within a stratum defined by covariates. Instead, we focus on the distribution of the response for the i^{th} subject:

$$Y_i \sim Bernoulli(p_i)$$

where $p_i = \text{pr}[Y_i = 1|x_{i1}, ..., x_{iK}]$ follows the logistic regression model

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

Q: what does i.i.d. refer to?

- We will show in the following that the parameter β_k has the interpretation of a **log-odds ratio** between the response and a one unit increase in the covariate x_{ik}, **conditional** on the other covariates.

- To simplify the interpretation of model parameters, we will temporarily drop the subscript i:

$$\text{logit}(\text{pr}[Y = 1|x_1, ..., x_K]) = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K$$

Interpretation of β_k

- Consider the two logits, where we hold all but x_k constant:

  for $x_k = c$:

  $$\text{logit}(P[Y = 1|x_k = c]) \;=\; \beta_0 + ... + \beta_k c ... + \beta_K x_K$$

  for $x_k = c + 1$:

  $$\text{logit}(P[Y = 1|x_k = c + 1]) \;=\; \beta_0 + ... + \beta_k (c + 1) ... + \beta_K x_K$$

- The log-odds ratio for the two groups is the difference in the logits:

  $$\text{logit}(p|x_k = c + 1) - \text{logit}(p|x_k = c) \;=\; \beta_k$$

- Thus, β_k is the log-odds ratio for a one-unit increase in covariate x_k, given all the other covariates are the same.

- For example, if x_k is a dichotomous covariate which equals 1 for the new treatment and 0 for placebo, then β_k is the log-odds ratio for success for new treatment versus placebo, conditional on the other covariates being the same.

Interaction terms

- Suppose there is an interaction between x_{K-1} and x_K :

$$\text{logit}(p) \;=\; \beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K x_K + \gamma x_{K-1} x_K$$

- Now, if we compare the same two logits as before:

$$\begin{aligned}
&\text{logit}(p|x_K = c+1) - \text{logit}(p|x_K = c) \\
&\quad = \beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K (c+1) + \gamma x_{K-1}(c+1) \\
&\qquad - \beta_0 + \beta_1 x_1 + ... + \beta_{K-1} x_{K-1} + \beta_K c + \gamma x_{K-1} c \\
&\quad = \beta_K + \gamma x_{K-1}
\end{aligned}$$

- Thus, conditional on the first $(K-1)$ covariates, the log-odds ratio for a one unit increase in the K^{th} covariate is

$$\beta_K + \gamma x_{K-1}$$

  and **depends on the level of** x_{K-1}

- We could include both two-way and three-way interactions, but interpretation of interactions terms becomes complicated even with just two-way interactions.

Main effects model results

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	0.3327	0.8409	0.1566	0.6923	.	1.395
SEX	1	0.2168	0.3389	0.4095	0.5222	0.053354	1.242
AGE	1	-0.00530	0.0144	0.1361	0.7122	-0.032426	0.995
TRT	1	0.7005	0.3136	4.9897	0.0255	0.193432	2.015
X	1	1.4231	0.3102	21.0539	0.0001	0.365832	4.150

Q: what is the meaning of intercept?

Maximum Likelihood Estimation (MLE) for Logistic Regression

- Consider the general logistic regression model

$$\text{logit}(p_i) \;=\; \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

$$\text{with} \quad Y_i \;\sim\; Bernoulli(p_i) \qquad i = 1, ..., n$$

- The **likelihood**, i.e. probability of observed data, is

$$L(\beta_0, \beta_1, ..., \beta_K) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{(1-y_i)}$$

where

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK})}.$$

- The **log-likelihood** is

$$\log[L(\beta_0, \beta_1, ..., \beta_K)]$$

$$= \beta_0 \left(\Sigma_{i=1}^{n}\, y_i\right) + \beta_1 \left(\Sigma_{i=1}^{n}\, x_{i1} y_i\right) + ... + \beta_K \left(\Sigma_{i=1}^{n}\, x_{iK} y_i\right)$$

$$- \Sigma_{i=1}^{n} \log(1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}})$$

- The **MLE** is the value of $\beta_0, \beta_1, ..., \beta_K$ that maximizes the (log-)likelihood function.

- It is possible that the data are so sparse (most people respond 0 or most people respond 1) that there is no solution to the maximization problem. (This is different from linear regression.)

- But if there is a solution, it can be shown to be unique.

- In practice, if there is no solution, your logistic regression software will say something like 'Convergence not reached after xx iterations'.

- The smaller of the number of 0's or number of 1's, call it the number of 'events", is the **'effective' sample size** for logistic regression. (*Rule of thumb* says that you need at least 10 events for each parameter to be estimated.)

- There are other methods that may be more appropriate with sparse data (conditional logistic regression, exact methods).

Confidence Intervals

95% (asymptotic) confidence interval for β_k can be obtained via

$$\widehat{\beta}_k \pm 1.96 \sqrt{\widehat{\text{Var}}(\widehat{\beta}_k)}$$

- In software outputs, you can look under the column labeled "standard error" to get the square root of the variance for a particular parameter estimate.

- In fact the estimated variance-covariance matrix for the entire vector of estimates $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_K)$ is also given by the statistical software packages.

Confidence interval for a linear combination

- Suppose we want a 95% confidence intervals for a linear combination (sometimes called 'contrast') of $\boldsymbol{\beta}$ of the form

$$\mathbf{c}\boldsymbol{\beta} = c_0\beta_0 + c_1\beta_1 + ... + c_K\beta_K$$

for some set of constants $\mathbf{c} = [c_0, c_1, ..., c_K]$

- For example, consider a model with interaction

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2}$$

- The log-odds ratio for a one unit increase in x_{i2}, given a value of $x_{i1} = x_1$ is

$$\beta_2 + \beta_{12} x_1$$

- For this model, the parameter vector and contrast of interest are:

$$\boldsymbol{\beta}' = [\beta_0, \beta_1, \beta_2, \beta_{12}]$$

$$\mathbf{c} = [0 \;\; 0 \;\; 1 \;\; x_1]$$

- For a 95% confidence intervals for $\mathbf{c}\beta$, we would use

$$\mathbf{c}\widehat{\boldsymbol{\beta}} \pm 1.96 \sqrt{\mathbf{c}\widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}]\mathbf{c}'}$$

$$\text{since} \quad \text{Var}[\mathbf{c}\widehat{\boldsymbol{\beta}}] = \mathbf{c}\text{Var}[\widehat{\boldsymbol{\beta}}]\mathbf{c}'$$

Test Statistics for parameters β_k

- For the logistic model, a test of $H_0 : \beta_k = 0$ represents a test of whether the k^{th} covariate (x_{ik}) affects the probability of success, with the null hypothesis that the probability of success is independent of x_{ik}.

- **Wald Statistic:**

$$Z = \frac{\widehat{\beta}_k}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})_{k+1,k+1}}}$$

This test statistic is asymptotically $N(0,1)$ under the null in large samples.

- **Likelihood Ratio Statistic:**

$$\begin{aligned} T &= 2\{\log L(\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_K) \\ &\qquad - \log L(\tilde{\beta}_0, \tilde{\beta}_1, ..., \tilde{\beta}_k = 0, ..., \tilde{\beta}_K | H_0)\} \\ &= 2 \sum_{j=1}^{n} \left[y_i \log \left(\frac{\widehat{p}_i}{\tilde{p}_i} \right) + (1 - y_i) \log \left(\frac{1 - \widehat{p}_i}{1 - \tilde{p}_i} \right) \right] \end{aligned}$$

where $\widehat{p}_j$ is the MLE, and $\tilde{p}_j$ is the estimate under the null (remember p is a function of the β's). This test statistic follows a χ^2_1 distribution under the null in large samples.

Score test statistic (sometimes called "**Rao's**")

- The score test statistic is:

$$X^2 = \frac{[\Sigma_{i=1}^n x_{ik}(y_i - \tilde{p}_i)]^2}{\widehat{\text{Var}}[\Sigma_{i=1}^n x_{ik}(y_i - \tilde{p}_i)]}$$

- It is computed under the null, and hence $\tilde{p}_i$ in the expression.

Likelihood Ratio Test for Nested Models

- Sometimes you have nested models resulting from, for example, putting additional interaction terms and/or square terms in the model and testing their significance.

- For example, suppose you have Model 1,

  Model 1:

$$p_i = \frac{e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i}}$$

- This model is nested in Model 2:

  Model 2:

$$p_i = \frac{e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i + \boldsymbol{\beta}_2' \mathbf{z}_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i + \boldsymbol{\beta}_2' \mathbf{z}_i}}$$

- We want to test

$$H_0 : \boldsymbol{\beta}_2 = 0$$

- The model with more parameters will always have a larger value for the maximized likelihood, since it is maximized over a larger parameter space.

- The difference between the maximized log likelihoods (i.e. likelihood ratio statistic) can be used to test for significance of the extra parameters in model 2 versus model 1:

$$\Delta = 2\{\log L(\hat{\boldsymbol{\beta}}|\mathrm{M}_2) - \log L(\tilde{\boldsymbol{\beta}}|\mathrm{M}_1)\}$$

- If the smaller model fits, i.e. under the null, Δ follows a χ^2_m distribution in large samples, where m parameters are set to 0 in the smaller model.

Fitted example: ICU data

This data set is available in R package 'aplore3'.

```
> summary(icu$sta)
Lived  Died
  160    40

icu.fit <- glm(sta ~ gender + age + race + loc, family = binomial(),
data = icu)

summary(icu.fit)

Call:
glm(formula = sta ~ gender + age + race + loc, family = binomial(),
    data = icu)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.03430  -0.63669  -0.51957  -0.00017   2.35145

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.19021    0.82094  -3.886 0.000102 ***
genderFemale   -0.16447    0.42932  -0.383 0.701652
age             0.02582    0.01232   2.096 0.036048 *
raceBlack     -16.26983 1464.46752  -0.011 0.991136
raceOther      -0.13027    1.10471  -0.118 0.906131
locStupor      34.91061 2822.61314   0.012 0.990132
locComa         2.99064    0.82556   3.623 0.000292 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 153.91  on 193  degrees of freedom
AIC: 167.91

Number of Fisher Scoring iterations: 17
```

Note that residual deviance here is -2 times the log likelihood evaluated at the MLE.

Q: is this a good model to fit?

```
> summary(icu$race)
White Black Other
  175    15    10
> summary(icu$loc)
Nothing  Stupor    Coma
    185       5      10
```

- Testing nested models:

```
> anova(icu.fit, test = "LRT")
Analysis of Deviance Table

Model: binomial, link: logit

Response: sta

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     199     200.16
gender  1    0.084       198     200.08  0.771314
age     1    7.771       197     192.31  0.005309 **
race    2    1.256       195     191.05  0.533636
loc     2   37.140       193     153.91 8.614e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Q: what are the values of the likelihood ratio test statistic?

More on Logistic Regression

Under the model for subject i

$$p_i = \frac{e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i}}.$$

- The estimated 'risk score' is $\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1' \mathbf{x}_i$, which is the log-odds, also referred to as the linear predictors.

- The estimated or predicted probabilities of response is

$$\hat{p}_i = \frac{e^{\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1' \mathbf{x}_i}}{1 + e^{\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1' \mathbf{x}_i}}.$$

```
> summary(predict(icu.fit))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-19.005  -2.160  -1.625  -1.971  -1.357  33.657

> summary(predict(icu.fit, type = "response"))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.1034  0.1646  0.2000  0.2048  1.0000
```

Earlier there was warning:

```
> icu.fit <- glm(sta ~ gender + age + race + loc,
family = binomial(), data = icu)

Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

How to Build a Model

- Knowledge of the subject matter area, obtaining data.

- We then typically start by looking at the data, literally, as we discussed before.

- Descriptive summary statistics such as those in 'Table 1' are often calculated.

- When we have an outcome variable of interest, in this case binary, we first examine the relationship between it and *each* of the other variables that are potentially predictors.
  - This is to **screen** out the variables that might be considered 'noise';
  - Noise variables can often impact the final performance of the model in a negative way.

  Eg. in the SEER-Medicare data example we showed at the beginning of this course (Hou *et al.*, 2018), from near 9,000 insurance claims codes, we screened down to 2,188 codes for non-cancer mortality, and 1,079 codes for cancer mortality. Screening is also commonly done for -omics studies.

  Q: how do you examine this relationship?

- We then consider multivariate (some call 'multivariable') logistic regression models in this case.

  Q: what is the purpose of building a model?

  [Hint] consider the homework assignment about prediction and causal inference.

Tools for Model Selection

There are various approaches to model selection. In practice, model selection can be done through a combination of them.

- Stepwise procedures
  - forward selection
  - backward selection
  - stepwise selection

- Measures of explained variation, eg. R^2

- Information criteria, eg. AIC, BIC, etc.

- Other dimension reduction methods for high-dimensional data.

Stepwise Procedures

Stepwise procedures have been criticized for being *ad hoc* etc, but continue to be very widely used in practice.

Caution: 'stepAIC' in R gives strange results, and is not recommended. (They are available as automated procedures in Stata and SAS.)
R package 'SignifReg' has p-value (see below) as criterion but only for linear regression.

We briefly describe the stepwise (back-n-forth) procedure there:

(1) Fit a univariate model for each covariate, and identify the predictors significant at some level p_1, say 0.20. (This is the screening step.)

(2) Fit a multivariate model with all significant univariate predictors, and use *backward* selection to eliminate non-significant variables at some level p_2, say 0.10.

(3) Starting with final step (2) model, consider each of the non-significant variables from step (1) using *forward* selection, with significance level p_3, say 0.10.

(4) Do final pruning of main-effects model (omit variables that are non-significant, add any that are significant), using *stepwise* regression with significance level p_4. At this stage, you may also consider adding interactions between any of the main effects currently in the model.

An illustration example:

Survival of Atlantic Halibut (Smith *et al.*)

Obs #	*Surv*ival *Time* (min)	Survival Indicator	*Tow Dur*ation (min.)	Diff in *Depth*	*Length* of Fish (cm)	*Handling* Time (min.)	Total *log(catch)* ln(weight)
100	353.0	1	30	15	39	5	5.685
109	111.0	1	100	5	44	29	8.690
113	64.0	0	100	10	53	4	5.323
116	500.0	1	100	10	44	4	5.323
⋮							

The following is results of Forward Selection in Stata, using p-value < 0.05 as entry criterion.

```
                    begin with empty model

p = 0.0000 <  0.0500  adding   handling
p = 0.0000 <  0.0500  adding   logcatch
p = 0.0010 <  0.0500  adding   towdur
p = 0.0003 <  0.0500  adding   length
```

```
------------------------------------------------------------------------------
survtime |
  censor |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
handling |   .0548994   .0098804      5.556   0.000     .0355341    .0742647
logcatch |  -.1846548    .051015     -3.620   0.000     .2846423   -.0846674
  towdur |   .5417745   .1414018      3.831   0.000     .2646321     .818917
  length |  -.0366503   .0100321     -3.653   0.000    -.0563129   -.0169877
------------------------------------------------------------------------------
```

The following is results of Backward Selection in Stata, using p-value ≥ 0.05 as removal criterion.

```
                    begin with full model

p = 0.1991 >= 0.0500  removing depth


------------------------------------------------------------------------------
survtime |
  censor |      Coef.   Std. Err.       z     P>|z|   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  towdur |   .5417745   .1414018      3.831   0.000   .2646321     .818917
logcatch |  -.1846548    .051015     -3.620   0.000  -.2846423   -.0846674
  length |  -.0366503   .0100321     -3.653   0.000  -.0563129   -.0169877
handling |   .0548994   .0098804      5.556   0.000   .0355341    .0742647
------------------------------------------------------------------------------
```

The following is results of Stepwise Selection in Stata, using p-value < 0.05 as entry criterion, and p-value ≥ 0.10 as removal criterion.

```
                    begin with full model

p = 0.1991 >= 0.1000  removing depth

------------------------------------------------------------------------------
survtime |
  censor |      Coef.   Std. Err.       z     P>|z|  [95% Conf. Interval]
---------+--------------------------------------------------------------------
  towdur |   .5417745   .1414018      3.831   0.000  .2646321     .818917
handling |   .0548994   .0098804      5.556   0.000  .0355341    .0742647
  length |  -.0366503   .0100321     -3.653   0.000 -.0563129   -.0169877
logcatch |  -.1846548    .051015     -3.620   0.000 -.2846423   -.0846674
------------------------------------------------------------------------------
```

Notes:

- When the halibut data was analyzed with the forward, backward and stepwise options, the same final model was reached. However, this will not always be the case.

- Sometimes we want to force certain variables in the models during the whole selection process, even if they may not be significant.

- Depending on the software, different tests (Wald, score, or likelihood ratio) may be used to decide what variables to add and what variables to remove.

Interactions

It is always a good idea to check for interactions:

In this example, there are several important interactions. Here backward selection was used, while forcing all main effects to be included, and considering all pairwise interactions. Here are the results:

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Exp of Estimate
TOWDUR	1	-0.075452	0.01740	18.79679	0.0001	0.927
DEPTH	1	0.123293	0.06400	3.71107	0.0541	1.131
LENGTH	1	-0.077300	0.02551	9.18225	0.0024	0.926
HANDLING	1	0.004798	0.03221	0.02219	0.8816	1.005
LOGCATCH	1	-0.225158	0.07156	9.89924	0.0017	0.798
TOWDEPTH	1	0.002931	0.0004996	34.40781	0.0001	1.003
TOWLNGTH	1	0.001180	0.0003541	11.10036	0.0009	1.001
TOWHAND	1	0.001107	0.0003558	9.67706	0.0019	1.001
DEPLNGTH	1	-0.006034	0.00136	19.77360	0.0001	0.994
DEPHAND	1	-0.004104	0.00118	12.00517	0.0005	0.996

Interpretation:

Handling alone doesn't seem to affect survival, unless it is combined with a longer towing duration or shallower trawling depths.

Hierarchical principle: if an interaction is included in a model, then the main effects are included.

An alternative modeling strategy when we have fewer covariates

With a dataset with only 5 main effects, you might be able to consider interactions from the start. How many would there be?

- Fit model with all main effects and pairwise interactions
- Then use backward selection to eliminate non-significant pairwise interactions (remember to force the main effects into the model at this stage, according to the 'hierarchical principle')
- Once non-significant pairwise interactions have been eliminated, you could consider backwards selection to eliminate any non-significant main effects that are not involved in remaining interaction terms

R^2-type Measures

R^2-type measures have always been considered useful in practice, to quantify how much variation in the outcome is explained by the regressors or predictors.

More recently:

- It has also been used to quantify genetic heritability, including the *polygenic risk scores*.
- For predictability, *out of sample* R^2 has been used in machine learning approaches.

Eg. In a prognostic study in gastric cancer, we wanted to investigate the prognostic effects of blood-based acute phase reactant proteins (i.e. *biomarkers*) and stage on survival. Note that stage is only available after surgery. The types of questions we were interested in:

1. How much of the variability in survival is explained, by the biomarkers and/or stage?

2. How strong are the effects of certain prognostic variables once others have been accounted for?

3. How much predictability is lost, if at all, when replacing a continuously measured covariate by a binary coding?

4. In some other disease areas, eg. CD4 counts in AIDS patients, how much is the effect on survival "captured" by such a *surrogate*?

Note that the R^2 measure concerns explained variation, or predictive capability, but **not** the goodness-of-fit of a model (which is a common misunderstanding).

Counter example: in linear regression, if the regression line is flat i.e. slope is close to zero, then R^2 is close to zero. But the fit can be good. The regressors just have little predictive power.

R^2 measure for logistic regression

For linear regression the R^2 measure, also called the coefficient of determination, is well-known. It is the proportion of the variance in the dependent variable that is explained by the independent variable(s).

For logistic regression (or binary outcome in general), a generalized R^2 (Cox and Snell) can be easily calculated using the likelihood ratio statistic:

- The measure can be defined as

$$R^2 = 1 - e^{-\Gamma},$$

  where

$$\Gamma \;=\; 2\{\log L(\hat{\boldsymbol{\beta}}) - \log L(\mathbf{0})\}/n,$$

  which is the likelihood ratio test statistic divided by n the sample size.

- It is
  - between 0 and 1 (why);
  - if $\hat{\boldsymbol{\beta}} = \mathbf{0}$, then $R^2 = 0$, therefore no regression effect translates to R^2 that is very close to zero;
  - increasing R^2 values generally indicate increasing predictability of the model (i.e. the regressors);

- for nested models, the R^2 value is non-decreasing with the larger model(s) (why);
- the use of R^2 can often be framed as: does the inclusion of additional predictors lead to substantial increase in R^2?

- R^2 can be seen as the **proportion of the explained randomness** (Kent, 1983), which is related to the Kullback-Leibler information:

$$R^2 = 1 - \frac{D(\hat{\boldsymbol{\beta}})}{D(\mathbf{0})},$$

where $D(\mathbf{0}) = \exp\{-2\log L(\mathbf{0})/n\}$ is the *randomness* in Y, and $D(\hat{\boldsymbol{\beta}}) = \exp\{-2\log L(\hat{\boldsymbol{\beta}})/n\}$ is the *residual randomness* of Y explained by $\mathbf{X}$.

- This mimics explained variation (under linear regression), which would be

$$1 - \frac{E\{\mathrm{Var}(Y|\mathbf{X})\}}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}\{E(Y|\mathbf{X})\}}{\mathrm{Var}(Y)}.$$

- Γ estimates twice the Kullback-Leibler information gain, between the fitted model and the null model.

Information Criteria

Information criteria have been used for model selection.

Akaike Information (AI):

- *Risk functions* are often used to evaluate a model, or a statistical procedure in general. Typically the smaller the risk the better.

- Risks are often defined as the expected value of a *loss function*.
  - Eg. squared error loss gives rise to mean squared error (MSE) as a risk function:
  - for estimating a population parameter θ and any estimator $\hat{\theta}$, the squared error is $(\hat{\theta} - \theta)^2$, so

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \text{Var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2,$$

    where $\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$. [Ex.]

- For AI we consider the *deviance loss* function $l(y, \theta) = -2 \log g_\theta(y)$, where g_θ is the density (or probability function) of a family of distributions indexed by parameter θ, that is used to model the observed data y.
  - Eg. the family might be Bernoulli, with a logistic regression model $\theta = (\beta_0, \beta_1, ..., \beta_K)'$.

- The corresponding risk function is $E\{-2\log g_\theta(y)\}$, closely related to the *Kullback-Leibler (KL) information* $E\{\log g_\theta(y)\}$.

- We want to choose the model that minimizes the above risk.

- Note that the KL information gain gives a kind of 'distance' from the true distribution f that generates the data y, to g_θ:
$$KL(f, g_\theta) = E_f\{\log f(y) - \log g_\theta(y)\}.$$

- For a given family g_θ, minimum KL is attained at θ_0 such that $KL(f, g_{\theta_0}) = \min_\theta KL(f, g_\theta)$ or, equivalently,
$$E\{\log g_{\theta_0}(y)\} = \max_\theta E\{\log g_\theta(y)\}.$$

- When the model is *correct*, we have $f = g_{\theta_0}$.

- In practice θ_0 is often estimated by the MLE $\hat{\theta}(y)$.

- Then the risk $-2E\{\log g_{\theta_0}(y)\}$ is 'estimated' by
$$-2E_{y^*}\{\log g(y^*|\hat{\theta}(y))\}.$$
Note that we use y^* to denote the r.v. that the expectation is w.r.t., in order to distinguish from y the observed data that's used to estimate θ_0.

- The **expected risk** in this case is the Akaike Information:

$$AI = -2E_y E_{y^*}\{\log g(y^*|\hat{\theta}(y))\}. \quad (1)$$

  It is also referred to as the *predictive* log-likelihood, or the expected KL. Note that y^* is an independent replicate of y, i.e. from the same distribution as y.

- The model should be chosen to minimize the AI, which itself needs to be estimated.

- **Q:** how would you estimate AI?

- It has been known that the 'apparent' estimate $-2 \log g(y|\hat{\theta}(y))$ under-estimates AI. (why)

- Instead Akaike (1973) showed that

$$AIC = -2 \log g(y|\hat{\theta}(y)) + 2p \qquad (2)$$

  is an approximately unbiased estimator of AI, where p is the dimension of θ.

- Therefore the model is chosen to minimize the AIC.

See ICU example for the computed AIC value.

Bayesian information criterion (BIC)

If p is the number of parameter in a model, the Bayesian information criterion is

$$BIC = -2 \log g(y|\hat{\theta}(y)) + p \cdot \log(n), \tag{3}$$

where n is the sample size.

Penalized log-likelihood

Almost all the methods for model selection we discuss here can be written as choosing β to maximize a penalized log-likelihood:

$$\log g(y|\beta) - P_\lambda(\beta),$$

where $\lambda \geq 0$ is the penalty parameter, and often we can use the penalty $P_\lambda(\beta) = \lambda \Sigma_{j=1}^{p} |\beta_j|^m$.

1. $m = 0$, L_0 penalty: best subset (AIC), stepwise (might require orthonormal design under linear regression), adjusted R^2, generalized cross-validation (GCV).

2. $m = 1$, L_1 penalty: least absolute shrinkage and selection operator (LASSO).

3. $m = 2$, L_2 penalty: ridge regression.

4. Other penalties: elastic net (combined L_1 and L_2 penalties), smoothly clipped absolute deviation (SCAD) etc.

See Harezlak et al. Chapter in "*High-Dimensional Data Analysis in Cancer Research*", Springer 2009.

Other Regression Models for Binary Outcome

- Although logistic regression is by far the most popular way to model Bernoulli data, we can also use other link functions.

- Since a probability must be between 0 and 1, we would like to model

$$p_i = \text{pr}[Y_i = 1 | x_{i1}, ..., x_{iK}]$$

as a function of covariates and parameters that will always be between 0 and 1.

- In **logistic regression**:

$$p_i = F(\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}) = \frac{e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}}}$$

where

$$F(u) = \text{pr}[U \leq u] = \frac{e^u}{1 + e^u}$$

is the cumulative distribution function (CDF) of the logistic distribution.

- In general, we can use any CDF to model p_i, since

$$F(u) = \text{pr}[U \leq u] \in [0, 1]$$

- So we can model

$$p_i = F(\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK})$$

where $F(u)$ is any CDF.

- You can think of $F(u)$ as a function that maps a number

$$-\infty < u < \infty \quad \Longrightarrow \quad 0 < F(u) < 1$$

- The nice thing about this structure is that it allows us to model $F^{-1}(p_i)$ as a **linear** function of the covariates:

$$F^{-1}(p_i) = \mathbf{x}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

- For example, for the logistic link: $F^{-1}(p_i) = \text{logit}(p_i)$

Some examples of link functions:

(1) The logistic link

(2) The Complementary log-log link

(3) The Probit link

Complementary log-log Link

- The CDF from the extreme value distribution is

$$F(u) = \exp[-\exp(-u)]$$

- If we substitute this CDF in $F(\mathbf{x}'\boldsymbol{\beta})$, we get:

$$p_i = \exp[-\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK})]$$

- This is equivalent to modeling the transformation $\log(-\log(p_i))$ as:

$$\log[-\log(p_i)] = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

  which is why it is called the complementary log-log link.

- It is often used for discrete survival data (when thought of as discretizing an underlying latent random variable that follows a proportional hazards model).

The Probit Link

- The **Probit** link corresponds to the standard normal $N(0,1)$ CDF

$$F(u) = \int_{-\infty}^{u} e^{-\frac{u^2}{2}} du = \Phi(u)$$

- There is no 'closed form expression' for $\Phi(u)$ as there is in the logistic, so we usually just denote it by $\Phi(u)$. For a given value of u, you use the computer to find $\Phi(u)$.

- The probit model can therefore be expressed as:

$$p_i = \Phi(\beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK})$$

$$\text{or as } \Phi^{-1}(p_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK}$$

- The fit of models using a probit link is typically very similar to the logistic link.

- This model has a biological interpretation related to tolerance distributions, and is therefore used for modeling data from dose-response studies. Some dose-response studies involve the identification of a "threshold" dose, above which you get the response ($Y = 1$) and below which you do not ($Y = 0$).

Maximum Likelihood Methods for Other Links

- Just as we showed for logistic regression, maximum likelihood methods can be used to estimate the parameters of these models, i.e., maximize

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

- The MLE can be obtained by setting the first derivative vector of the log likelihood to $\mathbf{0}$ and solving for $\widehat{\boldsymbol{\beta}}$, and the negative inverse of the estimated second derivative matrix can be used to estimate the variance.

- The solution is usually obtained by the Newton-Raphson algorithm, just as in logistic regression. You can use SAS Proc Logistic or Proc Probit.

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

How Biased is the Apparent Error Rate of a Prediction Rule?

Bradley Efron

To cite this article: Bradley Efron (1986) How Biased is the Apparent Error Rate of a Prediction Rule?, Journal of the American Statistical Association, 81:394, 461-470

To link to this article: https://doi.org/10.1080/01621459.1986.10478291

How Biased Is the Apparent Error Rate of a Prediction Rule?

BRADLEY EFRON*

A regression model is fitted to an observed set of data. How accurate is the model for predicting future observations? The apparent error rate tends to underestimate the true error rate because the data have been used twice, both to fit the model and to check its accuracy. We provide simple estimates for the downward bias of the apparent error rate. The theory applies to general exponential family linear models and general measures of prediction error. Special attention is given to the case of logistic regression on binary data, with error rates measured by the proportion of misclassified cases. Several connected ideas are compared: Mallows's C_p, cross-validation, generalized cross-validation, the bootstrap, and Akaike's information criterion.

KEY WORDS: Mallows's C_p; Cross-validation; AIC; Bootstrap methods; Logistic regression; Generalized linear models.

1. INTRODUCTION

Suppose the statistician fits a regression model to an observed set of data. How accurate, or inaccurate, is the model for predicting future observations? An obvious first guess is the *apparent error rate*, which is the observed inaccuracy of the fitted model applied to the original data points. However, the apparent error rate usually underestimates the true error rate. The reason is simple: the model is selected to lie near the observed points, which is what *fitting* means, so these points give a falsely optimistic picture of the model's true accuracy.

This article concerns estimating the bias of the apparent error rate. Here is a simple example of our results. A professional football player had the following field-goal kicking record over the 1969–1972 seasons:

Yards:	55	45	35	25	12	[Total]
Successes/Attempts:	$\frac{1}{4}$	$\frac{8}{27}$	$\frac{15}{32}$	$\frac{22}{25}$	$\frac{10}{12}$	$[\frac{56}{100}]$

(1.1)

Yards indicates the distance of the kick, discretized into the five categories shown. A standard linear logistic regression,

$$\Pr\{\text{success}\} = 1/[1 + \exp - \{\alpha_0 + \alpha_1 \cdot \text{Yards}\}], \quad (1.2)$$

was fitted to data (1.1) by maximum likelihood [see Efron (1982) for more details].

The fitted logistic regression estimates the probability of a successful kick as a function of the distance, $\hat{\Pr}\{\text{success}\} = 1/[1 + \exp - \{\hat{\alpha}_0 + \hat{\alpha}_1 \cdot \text{Yards}\}]$. We can consider this to be a prediction rule for future kicks, predicting success or failure as $\hat{\Pr}\{\text{success}\}$ is greater or less than .5. This rule has apparent error rate .310; that is, it mispredicts 31 of the 100 original data points (1.1). How optimistic is the value .310?

The theory that follows, in particular formula (2.4), estimates the downward bias of the apparent error rate to be only .012, indicating that bias is not a serious problem in this case. The reason for the small bias is the large ratio of data points to fitted parameters, 100 to 2. A random subset of 20 data points was selected from the 100 shown in (1.1). The logistic regression (1.2) fitted to just these 20 points had apparent error rate .400; that is, it misclassified 8 of the 20 points. Bias estimate (2.4) was now much larger, equaling .066.

Sections 2–5 concentrate on the special but important case of binary data and logistic regression. Error rates are usually measured as in the football example, by counting mispredictions. However, the theory allows more general measures of prediction error, for instance the Deviance (twice the Kullback–Leibler distance). Considering the prediction error of the Deviance leads to a nice corroboration of Akaike's information criterion, as shown in Section 6.

Section 6 extends the theory to linear models for general exponential families, as discussed for instance in McCullagh and Nelder (1983). Theorem 2 of Section 6 states the general result. This includes the most famous special case of all, ordinary linear regression with prediction error measured by squared Euclidean distance, where our bias estimate is equivalent to Mallows's C_p statistic (1973).

This relationship is examined in Section 7. It is easier to see the connection of our results with other methods, such as cross-validation, in the ordinary linear regression setting. Section 7 gives a comparative discussion of several closely related ideas: cross-validation, generalized cross-validation, bootstrap estimates of prediction error, Mallows's C_p, and Akaike's information criterion.

2. LOGISTIC REGRESSION

Logistic regression fits a model of the form

$$\pi_i = \frac{1}{1 + \exp(-t_i'\alpha)}, \qquad i = 1, 2, \ldots, n, \quad (2.1)$$

to an observed vector of binary data $y = (y_1, y_2, \ldots, y_n)$. Here the y_i independently equal 1 or 0 with probabilities π_i or $(1 - \pi_i)$; the t_i are observed p-dimensional covariate vectors; and α is an unknown p-dimensional vector of parameters.

The maximum likelihood estimate (MLE) of α, say $\hat{\alpha}$, gives estimates $\hat{\pi}_i$ by substitution in (2.1). We can think of the $\hat{\pi}_i$ as predicting whether a future observation with covariate vector t_i will be a 1 or a 0. For example, the predictions $\hat{\eta}_i$ might be given by the rule

$$\begin{aligned} \hat{\eta}_i &= 1 \quad \text{if } \hat{\pi}_i > C_0 \\ &= 0 \quad \text{if } \hat{\pi}_i \le C_0, \end{aligned} \quad (2.2)$$

for some cutoff point C_0. The choice $C_0 = .5$ is common.

* Bradley Efron is Professor of Statistics and Biostatistics, Stanford University, Sequoia Hall, Stanford, CA 94305. The author is grateful to Robert Tibshirani for suggesting the close connection of Section 5 to Akaike's information criterion.

How accurate is prediction rule (2.2)? The apparent error rate

$$\overline{\text{err}} = \#\{y_i \neq \hat{\eta}_i\}/n \tag{2.3}$$

is the proportion of cases in the original data set y incorrectly predicted by $\hat{\eta}$. However, since y was used to construct $\hat{\eta}$, $\overline{\text{err}}$ will usually be biased downward: a new data vector generated according to (2.1) might not be predicted nearly as accurately by the old $\hat{\eta}$.

We will derive estimates for $\omega(\pi)$, the expected downward bias of the apparent error rate as an estimator of the true error rate. The bias estimate for the logistic regression situation (2.1)–(2.3) is

$$\omega(\hat{\pi}) = \frac{2}{n}\sum_{i=1}^{n} \hat{\pi}_i(1 - \hat{\pi}_i)\phi\left(\frac{\hat{c}_i}{\sqrt{\hat{d}_i}}\right)\sqrt{\hat{d}_i}. \tag{2.4}$$

Here $\phi(z) = (2\pi)^{-1/2}\exp(-\frac{1}{2}z^2)$,

$$\hat{c}_i = \log\left(\frac{C_0}{1 - C_0}\right) - t_i'\hat{\alpha}, \tag{2.5}$$

and

$$\hat{d}_i = t_i'\hat{\Sigma}^{-1}t_i, \qquad \hat{\Sigma} \equiv \sum_{j=1}^{n} \hat{\pi}_j(1 - \hat{\pi}_j)t_jt_j'. \tag{2.6}$$

The matrix $\hat{\Sigma}^{-1}$ is the usual estimate for the covariance matrix of $\hat{\alpha}$, so $\hat{d}_i = \hat{\text{var}}(t_i'\hat{\alpha})$ is a quantity available in the output of most logistic regression programs. Formula (2.4) gave the estimates of bias for the football data quoted in the Introduction.

3. OPTIMISM OF THE APPARENT ERROR RATE

This section considers estimating the *expected optimism* of the apparent error rate, in other words the downward bias of $\overline{\text{err}}$ as an estimate of the true error rate. A simple bias formula is derived applying to general prediction rules and general measures of prediction error. Section 4 specializes this formula to the case of logistic regression and counting error, (2.1)–(2.3), obtaining (2.4). The results here, applying to binary data, are extended to general exponential families in Section 6.

Suppose then that $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ are probabilities, giving the data vector $y = (y_1, y_2, \ldots, y_n)$ by independent binary sampling,

$$\begin{aligned} y_i &= 1 \quad \text{with probability} \quad \pi_i \\ &= 0 \quad \text{with probability} \quad 1 - \pi_i, \end{aligned} \tag{3.1}$$

abbreviated $y \sim B(\pi)$. From y we form a vector of predictions $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_n)$, each $\hat{\eta}_i$ in the range $[0, 1]$. For now we do not have to specify the rule $y \to \hat{\eta}$.

Given y_i and $\hat{\eta}_i$, we have some measure $Q[y_i, \hat{\eta}_i]$ of prediction error, for instance the "counting error" of Section 1 (where the $\hat{\eta}_i$ equalled 0 or 1)

$$\begin{aligned} Q[y_i, \hat{\eta}_i] &= 1 \quad \text{if} \quad y_i \neq \hat{\eta}_i \\ &= 0 \quad \text{if} \quad y_i = \hat{\eta}_i. \end{aligned} \tag{3.2}$$

The average prediction error for the vector $\hat{\eta}$ is defined to be

$$Q[y, \hat{\eta}] \equiv \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \hat{\eta}_i] \equiv \overline{\text{err}}. \tag{3.3}$$

For counting error, $\overline{\text{err}}$ in (3.3) is the apparent error rate $\overline{\text{err}}$ defined at (2.3).

A wide class of error measures Q can be generated in the following way: let $q(\hat{\eta}_i)$ be a concave function of $\hat{\eta}_i \in [0, 1]$ satisfying $q(0) = q(1) = 0$. (It is convenient but not necessary to have $q(0)$ and $q(1)$ equal 0, as the more general development in Section 6 shows.) Then define the prediction error to be

$$Q[y_i, \hat{\eta}_i] = q(\hat{\eta}_i) + \dot{q}(\hat{\eta}_i)(y_i - \hat{\eta}_i). \tag{3.4}$$

Here y_i equals 0 or 1, and $\dot{q}(\hat{\eta}_i)$ is the derivative of $q(\hat{\eta}_i)$, uniquely defined by left continuity at sharp corners of the concave function q. In other words $Q[y_i, \hat{\eta}_i]$ is the height at y_i of the tangent line to q through the point $(\hat{\eta}_i, q(\hat{\eta}_i))$. See Efron (1978b) for an extensive discussion of such functions.

Three examples of error measures $Q[y_i, \hat{\eta}_i]$ are shown in Table 1. Example 1, counting error, is (3.2) extended in the obvious way for predictions $\hat{\eta}_i$ possibly intermediate between 0 and 1. Notice that Example 2, squared error, agrees with counting error when $\hat{\eta}_i$ equals 0 or 1, but is different for intermediate values $\hat{\eta}_i \in (0, 1)$. In Example 3, the average prediction error $Q[y, \hat{\eta}]$, (3.3), can be expressed as

$$Q[y, \hat{\eta}] = -(2/n)\log f_{\hat{\eta}}(y), \tag{3.5}$$

where $\log f_\pi(y)$ is the log-likelihood of $y \sim B(\pi)$; then $n\,Q[y, \hat{\eta}]$ equals the deviance, twice the Kullback–Leibler distance (see Section 6).

The *true error rate* $\text{Err}(y, \pi)$ of a prediction vector $\hat{\eta}$ is defined to be

$$\text{Err}(y, \pi) \equiv E_{\text{NEW}}\{Q[y^{\text{NEW}}, \hat{\eta}]\}. \tag{3.6}$$

Here y^{NEW} is a hypothetical new data vector, with the same distribution but independent of the original data vector y, which gave $\hat{\eta}$. The notation in (3.6) indicates expectation over $y^{\text{NEW}} \sim B(\pi)$, with $\hat{\eta}$ held fixed. In the case of counting error (3.2), Err is the expected proportion of incorrect predictions $y_i^{\text{NEW}} \neq \hat{\eta}_i$.

The difference between Err and $\overline{\text{err}}$ is the *optimism*

$$\text{op}(y, \pi) \equiv \text{Err} - \overline{\text{err}}. \tag{3.7}$$

The expectation of $\text{op}(y, \pi)$ over $y \sim B(\pi)$,

$$\omega(\pi) \equiv E_\pi\{\text{op}(y, \pi)\}, \tag{3.8}$$

is the expected optimism for the rule $y \to \hat{\eta}$, the quantity we wish to estimate.

Theorem 1. Let $\hat{\zeta} = (\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_n)$ be the vector with ith component

$$\hat{\zeta}_i \equiv -\dot{q}(\hat{\eta}_i). \tag{3.9}$$

Table 1. Three Measures of Prediction Error for Binary Data: The Measure $Q[y_i, \hat{\eta}_i]$ Is Derived From the Concave Function $q(\hat{\eta}_i)$ According to (3.4)

Name	$q(\hat{\eta}_i)$	$Q[y_i, \hat{\eta}_i]$
1. Counting Error	$\min(\hat{\eta}_i, 1 - \hat{\eta}_i)$	1 if $y_i = 1$, $\hat{\eta}_i < \frac{1}{2}$ or if $y_i = 0$, $\hat{\eta}_i > \frac{1}{2}$ 0 otherwise
2. Squared Error	$\hat{\eta}_i(1 - \hat{\eta}_i)$	$(y_i - \hat{\eta}_i)^2$
3. Deviance (twice Kullback–Leibler)	$-2[\hat{\eta}_i\log(\hat{\eta}_i) + (1 - \hat{\eta}_i)\log(1 - \hat{\eta}_i)]$	$-2\log \hat{\eta}_i^{y_i}(1 - \hat{\eta}_i)^{1-y_i}$

Then the expected optimism is

$$\omega(\pi) = \frac{1}{n} E_\pi\left\{\sum_{i=1}^{n} \hat{\zeta}_i \cdot (y_i - \pi_i)\right\}. \tag{3.10}$$

Proof. From definition (3.4),

$$Q[y_i^{\text{NEW}}, \hat{\eta}_i] - Q[y_i, \hat{\eta}_i] = \hat{\zeta}_i \cdot (y_i - y_i^{\text{NEW}}), \tag{3.11}$$

so

$$\text{op}(y, \pi) = \frac{1}{n} \sum_{i=1}^{n} \hat{\zeta}_i \cdot (y_i - \pi_i). \tag{3.12}$$

The theorem follows from definition (3.8).

Remark A. For the three error measures in Table 1, $\hat{\zeta}_i$ equals (a) $\text{sign}(2\hat{\eta}_i - 1)$, (b) $2\hat{\eta}_i - 1$, and (c) $2\log[\hat{\eta}_i/(1 - \hat{\eta}_i)]$, respectively.

Remark B. Another expression for $\omega(\pi)$ is

$$\omega(\pi) = \frac{1}{n} \sum_{i=1}^{n} \text{cov}_\pi(y_i, \hat{\zeta}_i). \tag{3.13}$$

In the case of counting error, where $\hat{\zeta}_i = \text{sign}(2\hat{\eta}_i - 1) = 2\hat{\eta}_i - 1$ for $\hat{\eta}_i = 0$ or 1 as in (2.2),

$$\omega(\pi) = \frac{2}{n} \sum_{i=1}^{n} \text{cov}_\pi(y_i, \hat{\eta}_i). \tag{3.14}$$

Equation (3.14) is a quantitative statement of the fact that the expected bias of the apparent error rate depends on how much each y_i affects its own prediction $\hat{\eta}_i$.

Remark C. Formula (3.10) can also be expressed as

$$\omega(\pi) = \frac{1}{n} \sum_{i=1}^{n} \pi_i(1 - \pi_i)\Delta_i, \tag{3.15}$$

where

$$\Delta_i \equiv E_\pi\{\hat{\zeta}_i \mid y_i = 1\} - E_\pi\{\hat{\zeta}_i \mid y_i = 0\}. \tag{3.16}$$

This is another statement showing how $\omega(\pi)$ depends on the effect of y_i on its own prediction. Expressions (3.10), (3.13), and (3.15) are numerically identical of course, but (3.15) is slightly more convenient for the theoretical calculations of Section 4.

Remark D. The optimism $\text{op}(y, \pi)$, (3.7), refers to the error-rate bias for a given vector $\hat{\eta}$. The expected optimism $\omega(\pi)$, (3.8), is the expected bias for the *rule* $y \to \hat{\eta}$. We would like to estimate $\text{op}(y, \pi)$, but must settle for estimating $\omega(\pi)$, as briefly discussed in Section 5.

Remark E. Section 5 also considers the problem of estimating the true error rate $\text{Err} = \overline{\text{err}} + \text{op}(y, \pi)$. Constructing estimates of Err better than $\overline{\text{err}}$ is the most obvious purpose of estimating the bias $\omega(\pi)$.

Remark F. The notation Err and $\overline{\text{err}}$ is taken from Efron (1983). However, the definition of Err has been changed in a way that makes the problem easier. The difference has to do with working in the framework for Mallows's C_p statistic, rather than that appropriate to cross-validation calculations. This distinction, which relates to questions of conditionality, is discussed in Section 7.

Remark G. Often the observations y_i occur in groups, for instance, the five groups in the football example (2.7). If $\hat{\eta}_g$ is the common prediction in group g, then we might wish to measure prediction error according to a grouped measure $Q[p_g, \hat{\eta}_g]$, where p_g is the observed proportion of y_i's equaling 1 in group g. It turns out that formula (3.10), which ignores grouping, still gives reasonable answers in the grouped case. Starting with squared error at (3.3), for example, formula (3.10) is the expected optimism for the grouped measure $Q[p_g, \hat{\eta}_g] = (p_g - \hat{\eta}_g)^2$, as well as for the ungrouped measure $Q[y_i, \hat{\eta}_i] = (y_i - \hat{\eta}_i)^2$. See Efron (1978b), in particular the last column of Table 3.

Remark H. The proof of Theorem 1 uses only $E_{\text{NEW}}\{y^{\text{NEW}} \mid y, \pi\} = \pi$, and not the full distributional assumptions y, y^{NEW} independently $\sim B(\pi)$. In particular, (3.10) applies to the case where the y_i are correlated binary variables.

4. DERIVATION OF FORMULA (2.4)

This section specializes Theorem 1 to the case of logistic regression and counting error (2.1)–(2.3), leading to estimate (2.4) for the expected optimism $\omega(\pi)$. Except for the end remarks, the discussion in this section is mainly technical.

Suppose then that the binary data vector is distributed as $y \sim B(\pi)$ as in (3.1), that $\pi = (\pi_1, \ldots, \pi_n)$ is given by the logistic formula (2.1), and that the prediction rule is (2.2). For convenience define

$$\chi_i = \pi_i(1 - \pi_i). \tag{4.1}$$

Then (3.15), (3.16) can be expressed as

$$\omega(\pi) = \frac{1}{n} \sum_{i=1}^{n} \chi_i \Delta_i, \tag{4.2}$$

where, since $\hat{\zeta}_i = 2\hat{\eta}_i - 1$ in this case,

$$\Delta_i = 2[E_\pi\{\hat{\eta}_i \mid y_i = 1\} - E_\pi\{\hat{\eta}_i \mid y_i = 0\}]. \tag{4.3}$$

The derivation of (2.4) consists of finding a simple approximation for (4.3).

Under model (2.1), the p-dimensional vector

$$z = \sum_{i=1}^{n} t_i y_i \tag{4.4}$$

is sufficient for α, having an exponential family of density functions

$$f_\alpha(z) = \exp(\alpha' z - \phi(\alpha)),$$

$$\phi(\alpha) = \sum_{i=1}^{n} \log(1 + \exp(t_i'\alpha)). \tag{4.5}$$

The vector α is the natural parameter of this family, whereas the expectation parameter is the vector

$$\beta \equiv E_\alpha\{z\} = \sum_{i=1}^{n} t_i \pi_i. \tag{4.6}$$

The MLE $\hat{\beta}$ of β equals the observed vector z, with covariance matrix

$$\Sigma = \sum_{i=1}^{n} \chi_i t_i t_i'. \tag{4.7}$$

Notice that according to (2.1), (2.2), $\hat{\eta}_i$ equals 1 or 0 as

$t_i'(\hat{\alpha} - \alpha)$ exceeds or is less than

$$c_i = \log(C_0/(1 - C_0)) - t_i'\alpha. \tag{4.8}$$

Then (4.3) can be written as

$$\Delta_i = 2[\Pr_\pi\{t_i'(\hat{\alpha} - \alpha) > c_i \mid y_i = 1\} - \Pr_\pi\{t_i'(\hat{\alpha} - \alpha) > c_i \mid y_i = 0\}]. \tag{4.9}$$

Standard exponential family theory gives the approximation

$$t_i'(\hat{\alpha} - \alpha) \doteq t_i'\dot{\Sigma}^{-1}(\hat{\beta} - \beta) \tag{4.10}$$

(see Efron 1978a, eq. 2.4), so

$$\Delta_i \doteq 2[\Pr_\pi\{t_i'\dot{\Sigma}^{-1}(\hat{\beta} - \beta) > c_i \mid y_i = 1\} - \Pr\{t_i'\dot{\Sigma}^{-1}(\hat{\beta} - \beta) > c_i \mid y_i = 0\}]. \tag{4.11}$$

Now let $\beta_{(i)} \equiv \sum_{j\neq i} t_j\pi_j$ and $\hat{\beta}_{(i)} \equiv \sum_{j\neq i} t_j y_j$, so

$$\hat{\beta} - \beta = (\hat{\beta}_{(i)} - \beta_{(i)}) + t_i(y_i - \pi_i). \tag{4.12}$$

Likewise define $\dot{\Sigma}_{(i)} \equiv \sum_{j\neq i} \chi_j t_j t_j'$. A standard matrix identity gives

$$\dot{\Sigma}^{-1} = \left(I - \frac{\chi_i\dot{\Sigma}_{(i)}^{-1}t_i t_i'}{1 + \chi_i t_i'\dot{\Sigma}_{(i)}^{-1}t_i}\right)\dot{\Sigma}_{(i)}^{-1}. \tag{4.13}$$

Finally, letting $d_{(i)i} = t_i'\dot{\Sigma}_{(i)}^{-1}t_i$, we have

$$d_i = \frac{d_{(i)i}}{1 + \chi_i d_{(i)i}} \quad \text{or} \quad d_{(i)i} = \frac{d_i}{1 - \chi_i d_i}, \tag{4.14}$$

where $d_i = t_i'\dot{\Sigma}^{-1}t_i$.

From (4.12), (4.13) we get

$$t_i'\dot{\Sigma}^{-1}(\hat{\beta} - \beta) = (1 - \chi_i d_i)t_i'\dot{\Sigma}_{(i)}^{-1}(\hat{\beta}_{(i)} - \beta_{(i)}) + d_i(y_i - \pi_i). \tag{4.15}$$

This is a convenient formula for use in (4.11), since it separates out the dependence of $t_i'\dot{\Sigma}^{-1}(\hat{\beta} - \beta)$ on y_i.

Standard asymptotic theory gives a limiting normal distribution for $\hat{\beta}_{(i)} - \beta_{(i)}$ as the matrix $\dot{\Sigma}_{(i)}$ grows large, $\hat{\beta}_{(i)} - \beta_{(i)} \to N_p(0, \dot{\Sigma}_{(i)})$, so

$$t_i'\dot{\Sigma}_{(i)}^{-1}(\hat{\beta}_{(i)} - \beta_{(i)}) \to N(0, d_{(i)i}). \tag{4.16}$$

The mean 0 and variance $d_{(i)i}$ are exact in (4.16), only the normality being asymptotic. Notice that $\hat{\beta}_{(i)}$ is independent of y_i. We can now write (4.15) as

$$t_i'\dot{\Sigma}^{-1}(\hat{\beta} - \beta) = (1 - \chi_i d_i)\sqrt{d_{(i)i}}Z + d_i(y_i - \pi_i) = [(1 - \chi_i d_i)d_i]^{1/2}Z + d_i(y_i - \pi_i), \tag{4.17}$$

where $Z \to N(0, 1)$ is independent of y_i.

Using the last expression of (4.17) in (4.11) gives

$$\Delta_i \doteq 2\left\{\Phi\left(\frac{c_i + d_i\pi_i}{[d_i(1 - \chi_i d_i)]^{1/2}}\right) - \Phi\left(\frac{c_i - d_i(1 - \pi_i)}{[d_i(1 - \chi_i d_i)]^{1/2}}\right)\right\}, \tag{4.18}$$

$\Phi(z) \equiv \int_{-\infty}^{z} \phi(x)dx$. The factor $1 - \chi_i d_i$ is asymptotically negligible, leading to the slightly cruder approximations

$$\Delta_i \doteq 2\left\{\Phi\left(\frac{c_i}{\sqrt{d_i}} + \sqrt{d_i}\pi_i\right) - \Phi\left(\frac{c_i}{\sqrt{d_i}} - \sqrt{d_i}(1 - \pi_i)\right)\right\}, \tag{4.19}$$

$$\doteq 2\phi\left(\frac{c_i}{\sqrt{d_i}}\right)\sqrt{d_i}. \tag{4.20}$$

Going back to (4.2),

$$\omega(\pi) \doteq \frac{2}{n}\sum_{i=1}^{n} \chi_i\phi\left(\frac{c_i}{\sqrt{d_i}}\right)\sqrt{d_i}. \tag{4.21}$$

Formula (2.4) is (4.21) with

$$\hat{\pi}_i = 1/(1 + \exp(-t_i'\hat{\alpha})) \tag{4.22}$$

substituted everywhere for π_i. In other words, $\omega(\hat{\pi})$ is the MLE for $\omega(\pi)$, or at least the MLE for approximation (4.21) to $\omega(\pi)$. Formula (2.4) has the usual asymptotic optimality properties of maximum likelihood estimates. The favorable finite-sample properties of maximum likelihood, approximate median unbiasedness and high efficiency among nearly unbiased estimators, should also hold. Section 5 describes the performance of (2.4) in a sampling experiment.

Remark I. Theorem 1 combined with (4.15) makes it easy to derive bias expressions like (2.4) applying to deviance and squared error, rather than to counting error,

$$\text{squared error:} \quad \omega(\pi) \doteq \frac{2}{n}\sum_{i=1}^{n} \chi_i^2 d_i$$

$$\text{deviance:} \quad \omega(\pi) \doteq \frac{2p}{n}. \tag{4.23}$$

This last formula is an expression of Akaike's information criterion (AIC) (see the corollary in Sec. 6).

Remark J. Bootstrap methods can be used to approximate the bias estimate $\omega(\hat{\pi})$ for any prediction rule $y \to \hat{\eta}$, not necessarily involving logistic regression, and for any error measure $Q[y, \hat{\eta}]$. Parametric bootstrap data vectors are generated according to $y^* \sim B(\hat{\pi})$, giving bootstrap prediction vectors $y^* \to \hat{\eta}^*$; B such bootstrap replications give the estimate

$$\omega(\hat{\pi}) = \frac{1}{B}\sum_{b=1}^{B}\left[\frac{\sum_{i=1}^{n} \hat{\zeta}_i^*(b)\{y_i^*(b) - \hat{\pi}_i\}}{n}\right]. \tag{4.24}$$

As $B \to \infty$, (4.24) approaches $\omega(\hat{\pi})$, the MLE of $\omega(\pi)$.

Remark K. For the football example described in the Introduction, the approximate MLE (2.4) agreed well with actual MLE $\omega(\hat{\pi})$ evaluated by Monte Carlo, (4.24). Table 2 shows the comparison. The difference between (4.24) and (2.4) are small compared with the statistical variability in $\omega(\hat{\pi})$ (coming from the variability of $\hat{\pi}$ as an estimate of π).

The statistical variability is indicated by reevaluating the approximation based on (4.18) at vectors π moderately distant from the MLE $\hat{\pi}$. For example, the symbol $+\ -$ refers to (4.18) evaluated at the vector π obtained by substituting $(\hat{\alpha}_1 + \hat{s}d_1,$

Table 2. Estimates of $\omega(\pi)$ for the Football Data, All $n = 100$ Kicks, and for a Randomly Chosen Subset of $n = 20$ Kicks, as Described in the Introduction: The Last Four Rows Indicate the Statistical Variability in the Estimate $\omega(\hat{\pi})$

	$n = 100$	$n = 20$
Bootstrap (4.24):	.0120 (±.0011, B = 4,000)	.059 (±.004, B = 1,600)
Approximation (2.4):	.0119	.066
Approximation (4.18):	.0121	.075
+ +	.0143	.029
+ −	.0155	.108
− +	.0076	.063
− −	.0100	.065

$\hat{\alpha}_2 - \hat{s}d_2)'$ in (2.1), where $\hat{s}d_j$ is the original estimated standard deviation of $\hat{\alpha}_j$.

Large numbers of bootstrap replications B were taken in order to make the comparisons in Table 2 more informative. In fact, $B = 200$ bootstraps gave reasonable estimates of $\omega(\hat{\pi})$ in both cases. An advantage of the bootstrap method is that quantities of interest besides $\omega(\pi)$ can be estimated from the same replications, for example the variability in the prediction vector $\hat{\eta}$.

Remark L. Formula (2.4) makes double use of $\hat{\pi}$; as the vector that defines the predictions $\hat{\eta}$, via (2.2), and as the point in the space of possible π vectors at which $\omega(\pi)$ is evaluated. These two uses can be separated. In some cases the prediction vector $\hat{\eta}$ might not be obtained from $\hat{\pi}$, the MLE of π.

Here is an important example: suppose that in (2.1) α is partitioned into (α_0, α_1), and likewise $t_i' = (t_{0i}', t_{1i}')$; that $\hat{\pi}_i^0$ is the MLE of π_i in the lower-dimensional model where α_1 is assumed to be zero; and that $\hat{\eta}_i$ equals 1 or 0 as $\hat{\pi}_i^0$ is greater or less than C_0. In this case we are using a prediction rule based on a possibly inadequate parametric model.

The bias estimate for this situation turns out to be

$$\omega(\hat{\pi}) = \frac{2}{n}\sum_{i=1}^{n} \hat{\pi}_i(1 - \hat{\pi}_i)\phi\left(\frac{\hat{c}_i}{\sqrt{\hat{D}_i}}\right)\frac{\hat{d}_i^0}{\sqrt{\hat{D}_i}}, \qquad (4.25)$$

where $\hat{\pi}_i$ is obtained from (2.1) by substitution of $\hat{\alpha}$, the MLE of α in the full model; $\hat{c}_i$ is as given in (2.5); $\hat{d}_i^0 = t_{0i}'\hat{\Sigma}^{0-1}t_{0i}$, where $\hat{\Sigma}^0 \equiv \sum_{j=1}^n \hat{\pi}_j^0(1 - \hat{\pi}_j^0)t_{0j}t_{0j}'$; and $\hat{D}_i = t_{0i}'\hat{\Sigma}^{0-1}\hat{\Sigma}\hat{\Sigma}^{0-1}t_{0i}$, where $\hat{\Sigma} \equiv \sum_{j=1}^n \hat{\pi}_i(1 - \hat{\pi}_i)t_{0j}t_{0j}'$.

5. A SAMPLING EXPERIMENT

Table 3 reports the results of a sampling experiment on the performance of estimate (2.4) in a small-sample situation. The data for each trial of the sampling experiment consist of 20 independent vectors (y_i, s_i), where

$$y_i = 1, \quad \text{probability } \tfrac{1}{2}$$
$$= 0, \quad \text{probability } \tfrac{1}{2} \qquad (5.1)$$

and

$$s_i \mid y_i \sim N_2((y_i - \tfrac{1}{2}, 0), I) \qquad (5.2)$$

for $i = 1, 2, \ldots, 20$. Conditioning on s_i, model (2.1) applies to $\pi_i = P\{y_i = 1 \mid s_i\}$, with $t_i' = (1, s_i)$, $\alpha = (0, 1, 0)'$ [see Sec. 1 of Efron (1975)]. The sampling experiment comprised 100 trials, with 20 observations (y_i, s_i) for each trial.

For each trial, prediction rule (2.2) based on the logistic regression maximum likelihood estimates $\hat{\pi}_i$ was calculated from the data $\{(y_i, s_i), i = 1, 2, \ldots, 20\}$, $C_0 = .5$. The first two columns of Table 3 show the true error rate Err, (3.6), and the apparent error rate $\overline{\text{err}}$, (2.3). We see that the expected optimism in this situation is substantial, $\omega(\pi) = .342 - .254 = .088$. Four hundred more trials verified this value to within .001.

Column 4 shows the approximate MLE of the bias $\omega(\hat{\pi})$, (2.4); $\omega(\hat{\pi})$ is nearly unbiased for $\omega(\pi)$, with quite small standard deviation. For comparison, column 6 shows the cross-validation estimate of bias, $\hat{\omega}^{\text{CV}} \equiv \widehat{\text{Err}}^{\text{CV}} - \overline{\text{err}}$, where

$$\widehat{\text{Err}}^{\text{CV}} = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \hat{\eta}_{(i)i}], \qquad (5.3)$$

$\hat{\eta}_{(i)i}$ indicating the prediction (2.2) for case i based on the 19 observations (y_j, s_j), $j \neq i$. The estimate $\hat{\omega}^{\text{CV}}$ is also nearly unbiased, but has standard deviation more than four times larger than that of the MLE $\omega(\hat{\pi})$. This comparison of $\omega(\hat{\pi})$ with $\hat{\omega}^{\text{CV}}$ is somewhat unfair, as discussed in Section 7. See also Remark T.

Table 3. First 10 Trials of the Sampling Experiment, and Summary Statistics for 100 Trials

			Maximum Likelihood		Cross-Validation	
Trial	Err	$\overline{\text{err}}$	$\widehat{\text{Err}}$	$\omega(\hat{\pi})$	$\widehat{\text{Err}}^{\text{CV}}$	$\hat{\omega}^{\text{CV}}$
1	.364	.300	.409	.109	.500	.200
2	.302	.300	.405	.105	.400	.100
3	.378	.250	.324	.074	.300	.050
4	.276	.200	.289	.089	.300	.100
5	.320	.250	.335	.085	.300	.050
6	.369	.200	.284	.084	.250	.050
7	.296	.200	.278	.078	.300	.100
8	.437	.100	.177	.077	.100	.000
9	.336	.350	.450	.100	.450	.100
10	.354	.150	.234	.084	.250	.100
100 trials						
Mean	.342	.254	.346	.093	.349	.096
(Sd)	(.055)	(.094)	(.105)	(.015)	(.117)	(.065)
Coeff. of Variation	.16	.37	.30	.16	.34	.68

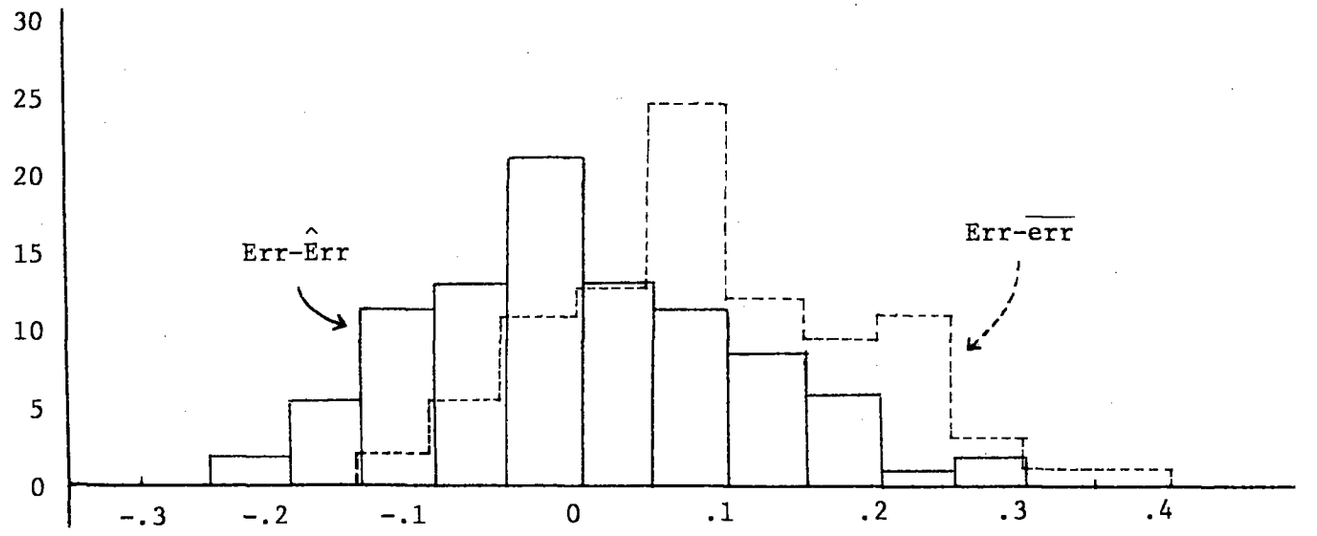


Figure 1. Errors in Estimating Err. Err − Êrr, solid line, compared with Err − ērr, dashed line.

The most obvious use of the bias estimate $\omega(\hat{\pi})$ is to correct $\bar{\text{err}}$ as an estimate of Err, say to

$$\hat{\text{Err}} \equiv \bar{\text{err}} + \omega(\hat{\pi}). \tag{5.4}$$

$\hat{\text{Err}}$ performs well in Table 3, removing the bias in $\bar{\text{err}}$ as an estimate of Err, while decreasing the coefficient of variation of the estimate from .37 to .30. Figure 1 compares the distributions of Err − $\hat{\text{Err}}$ and Err − $\bar{\text{err}}$ for the 100 trials. Notice that Err − $\bar{\text{err}}$ is positive 80% of the time, while Err − $\hat{\text{Err}}$ exceeds zero only 54% of the time.

The cross-validation estimate $\hat{\text{Err}}^{\text{CV}}$ is also nearly unbiased for Err, also with smaller coefficient of variation than $\bar{\text{err}}$. Let MSE($\tilde{\text{Err}}$) indicate the mean squared error of an estimate $\tilde{\text{Err}}$ for Err,

$$\text{MSE}(\tilde{\text{Err}}) \equiv E[\text{Err} - \tilde{\text{Err}}]^2. \tag{5.5}$$

In the sampling experiment,

$$\text{MSE}(\bar{\text{err}}) = .0174, \quad \text{MSE}(\hat{\text{Err}}) = .0115,$$
$$\text{MSE}(\hat{\text{Err}}^{\text{CV}}) = .0135. \tag{5.6}$$

These values can be compared with the MSE for the *ideal constant* estimator $\hat{\text{Err}}^{\text{IC}} = \bar{\text{err}} + \omega(\pi) = \bar{\text{err}} + .088$: MSE($\hat{\text{Err}}^{\text{IC}}$) = .0096. ($\hat{\text{Err}}^{\text{IC}}$ is the preferred estimate of Err if $\omega(\pi)$ is known, which of course is not so in most real problems.) The relative inefficiency of $\tilde{\text{Err}}$ is defined as in Efron (1983) to be

$$\text{REL}(\tilde{\text{Err}}) \equiv [\text{MSE}(\tilde{\text{Err}}) - \text{MSE}(\hat{\text{Err}}^{\text{IC}})] \div [\text{MSE}(\bar{\text{err}}) - \text{MSE}(\hat{\text{Err}}^{\text{IC}})]. \tag{5.7}$$

For our sampling experiment,

$$\text{REL}(\hat{\text{Err}}) = .24, \quad \text{REL}(\hat{\text{Err}}^{\text{CV}}) = .50. \tag{5.8}$$

To summarize the results of the experiment, $\hat{\text{Err}} = \bar{\text{err}} + \omega(\hat{\pi})$ quite effectively improves $\bar{\text{err}}$ as an estimate of Err, and $\omega(\hat{\pi})$ is an excellent estimator of $\omega(\pi)$. Cross-validation is hopelessly inefficient for estimating $\omega(\pi)$, but less bad for estimating Err.

Why would we want to estimate $\omega(\pi)$? In the author's opinion, $\omega(\pi)$ is an interesting measure of how vulnerable a prediction rule is to overfitting. A large value of $\omega(\hat{\pi})$, or perhaps of $\omega(\hat{\pi})/\bar{\text{err}}$, suggests retreating to a more parsimonious prediction rule. However, no quantitative guidelines have been investigated.

Remark M. In the sampling experiment, the correlation between $\omega(\hat{\pi})$ and op(y, π) was cor($\omega(\hat{\pi})$, op) = −.84. This confirms Remark D, that $\omega(\hat{\pi})$ is not estimating the random variable op(y, π), but rather its expectation $\omega(\pi)$.

Remark N. For any estimator $\tilde{\text{Err}} = \bar{\text{err}} + \tilde{\omega}$, the MSE (5.5) is

$$\text{MSE}(\tilde{\text{Err}}) = E[(\bar{\text{err}} + \text{op}) - (\bar{\text{err}} + \tilde{\omega})]^2 = E[\text{op} - \tilde{\omega}]^2. \tag{5.9}$$

In this context $\tilde{\omega}$ is judged by how well it estimates op, no matter what it is supposed to be estimating. In the sampling experiment cor($\hat{\omega}^{\text{CV}}$, op) = .03. This makes $\hat{\omega}^{\text{CV}}$ a relatively less bad estimate of op than of ω, compared with the MLE $\omega(\hat{\pi})$, which has much smaller variance but a substantial negative correlation [see (3.1) of Efron (1983)].

Remark O. In the setting of Efron (1983) it was possible to find a compromise between cross-validation and maximum likelihood that had small variance *and* nonnegative correlation with op. This compromise, the ".632 estimator," was the clear winner in the sampling experiments of the 1983 paper. It is plausible, but so far unverified, that a similar compromise is possible here.

Remark P. Our sampling experiment differs from experiment (2, 20) of Efron (1983) in the choice of prediction rule, logistic regression rather than linear discrimination, and in the definition of Err, as discussed in Section 7. That is why the numbers in Table 3 differ from those in Table 2 of Efron (1983).

6. EXPONENTIAL FAMILIES AND GENERAL LINEAR MODELS

All of our calculations so far have concerned binary data. Similar results hold when the y_i come from a general linear model, as described in McCullagh and Nelder (1983). This section gives a brief discussion of the theory, mostly without proofs.

We suppose that the independent observations y_i are members of a one-parameter exponential family with density functions

$$f_{\mu_i}(y_i) = \exp(\lambda_i y_i - \psi(\lambda_i)), \tag{6.1}$$

where $\mu_i = E\{y_i\}$ is the expectation parameter of the family, λ_i is the natural (or canonical) parameter, and $\psi(\lambda_i)$ is the normalizing function. The two parameters are related by the differential formula $\mu_i = d\psi(\lambda_i)/d\lambda_i$. In the binary case, μ_i equals π_i, λ_i equals the logit $\log\{\pi_i/(1 - \pi_i)\}$, and $\psi(\lambda_i) = \log(1 + e^{\lambda_i})$.

In a general linear model the natural parameters λ_i are expressed as linear combinations of known p-dimensional covariate vectors t_i and an unknown p-dimensional parameter vector α,

$$\lambda_i = t_i'\alpha, \qquad i = 1, 2, \ldots, n. \tag{6.2}$$

Model (2.1) is a special case of (6.2).

Having observed $y = (y_1, y_2, \ldots, y_n)$, we compute the MLE $\hat{\alpha}$, then $\hat{\lambda}_i = t_i'\hat{\alpha}$, and finally

$$\hat{\mu}_i = \frac{d\psi(\lambda_i)}{d\lambda_i}\bigg|_{\hat{\lambda}_i}.$$

Think of $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n)$ as a prediction vector for y, with apparent prediction error

$$Q[y, \hat{\mu}] = \frac{1}{n}\sum Q[y_i, \hat{\mu}_i] \equiv \bar{\text{err}}, \tag{6.3}$$

for some error measure $Q[y_i, \hat{\mu}_i]$. (We could let $\hat{\mu}$ further determine a prediction vector $\hat{\eta}$, as in (2.2), but for the discussion here it is sufficient to take $\hat{\eta} = \hat{\mu}$.) Following definitions (3.6)–(3.8), let $\text{Err}(y, \pi) \equiv E_{\text{NEW}}\{Q[y^{\text{NEW}}, \hat{\mu}]\}$ be the true error rate of $\hat{\mu}$, where y^{NEW} is an independent replication of y; and $\omega(\mu) \equiv E_\mu\{\text{Err} - \bar{\text{err}}\}$, the expected optimism of the rule $y \to \hat{\mu}$. We wish to estimate $\omega(\mu)$.

As in (3.4), we consider error measures $Q[y_i, \hat{\mu}_i]$, which are the difference between a concave function $q(y_i)$ and its tangent through a point $(\hat{\mu}_i, q(\hat{\mu}_i))$,

$$Q[y_i, \hat{\mu}_i] = q(\hat{\mu}_i) + \dot{q}(\hat{\mu}_i)(y_i - \hat{\mu}_i) - q(y_i). \tag{6.4}$$

Example. The concave function

$$q(y_i) = 2\{\psi(\lambda(y_i)) - y_i\lambda(y_i)\}, \tag{6.5}$$

where $\lambda(y_i)$ is the value of λ_i corresponding to $\mu_i = y_i$, makes $Q[y_i, \hat{\mu}_i]$ equal the *deviance*, that is, twice the Kullback–Leibler distance $I(y_i, \hat{\mu}_i)$. In this case

$$\begin{aligned} Q[y, \hat{\mu}] &= \frac{1}{n}\sum_{i=1}^{n} Q[y_i, \hat{\mu}_i] \\ &= \frac{2}{n}[\log f_y(y) - \log f_{\hat{\mu}}(y)], \end{aligned} \tag{6.6}$$

where $f_\mu(y) \equiv \prod_{i=1}^n f_{\mu_i}(y_i)$ [see Efron (1978a)].

A generalized version of (3.10) follows easily from (6.4) and the definition of $\omega(\mu)$:

Theorem 2. Let $\hat{\zeta}$ be the vector with ith component $\hat{\zeta}_i \equiv -\dot{q}(\hat{\mu}_i)$. Then

$$\omega(\mu) = \frac{1}{n}E_\mu\left\{\sum_{i=1}^{n}\hat{\zeta}_i \cdot (y_i - \mu_i)\right\}. \tag{6.7}$$

Theorem 2 applies to any prediction rule $y \to \hat{\mu}$, not necessarily one based on maximum likelihood or general linear models, and to any measure of prediction error of form (6.4).

Corollary. In the special case where $\hat{\mu}$ is the MLE of μ in a general linear model (6.2), and prediction error is based on the deviance (6.6), then

$$\omega(\mu) \doteq 2p/n. \tag{6.8}$$

Proof. Letting $T = (t_1, t_2, \ldots, t_n)$, (6.7) becomes

$$\omega(\mu) = (2/n)E_\mu\{(\hat{\alpha} - \alpha)'T(y - \mu)\}, \tag{6.9}$$

where we have used $\hat{\zeta}_i = 2t_i'\hat{\alpha}$. The exponential family approximation (4.10) then gives

$$\omega(\mu) \doteq (2/n)E_\mu\{(y - \mu)'T'\dot{\Sigma}^{-1}T(y - \mu)\}, \tag{6.10}$$

where

$$\dot{\Sigma} \equiv \sum_{i=1}^{n}\chi_i t_i t_i', \qquad \chi_i \equiv \text{var}_{\mu_i}(y_i). \tag{6.11}$$

However, the last expression in (6.10) equals

$$\begin{aligned} \frac{2}{n}\sum_{i=1}^{n}\chi_i(t_i'\dot{\Sigma}^{-1}t_i) &= \frac{2}{n}\,\text{tr}\,\dot{\Sigma}^{-1}\left\{\sum_{i=1}^{n} t_i t_i'\chi_i\right\} \\ &= \frac{2}{n}\,\text{tr}\,\dot{\Sigma}^{-1}\dot{\Sigma} = \frac{2}{n}\,\text{tr}\,I_p = \frac{2p}{n}. \end{aligned} \tag{6.12}$$

Remark Q. The corollary extends to the case where $\hat{\mu}$ is the MLE of μ in the p_0-dimensional subfamily of (6.2) obtained as in Remark L: by assuming, possibly incorrectly, that the last $p - p_0$ coordinates of α are zero. Then (6.8) becomes $\omega(\mu) \doteq 2p_0/n$.

Remark R. Akaike's information criterion (AIC) suggests penalizing the maximized log-likelihood $\log f_{\hat{\mu}}(y)$ by p in assessing the goodness of fit of a p-parameter model; that is, a choice between competing models containing different numbers of parameters is made by maximizing $\log f_{\hat{\mu}}(y) - p$ [see Atkinson (1980) and Stone (1977)]. Our corollary supports the AIC. Using formula (6.6) for $\bar{\text{err}}$, the corrected estimate $\hat{\text{Err}} = \bar{\text{err}} + \omega(\mu)$ is

$$\hat{\text{Err}} \doteq \frac{2}{n}\log f_y(y) - \frac{2}{n}\{\log f_{\hat{\mu}}(y) - p\}, \tag{6.13}$$

so AIC amounts to selecting the model with the minimum estimate of Err.

The only approximation in the proof of the corollary comes from (4.10). In the case of normally distributed observations, $y_i \sim N(\mu_i, \sigma^2)$, (4.10) is exact and so is (6.8); $\hat{\text{Err}} = \bar{\text{err}} + \omega(\mu)$ equals $\{\|y - \hat{\mu}\|^2/\sigma^2 + 2p\}/n$, Mallows's C_p statistic (see Sec. 7).

Remark S. The proof of Theorem 2 does not use the full distributional assumptions that y_i^{NEW} and y_i are independent observations from f_{μ_i}, independently for $i = 1, 2, \ldots, n$. All we need is $E_{\text{NEW}}\{y_i^{\text{NEW}} \mid y, \mu\} = \mu_i$ and $E_{\text{NEW}}\{q(y_i^{\text{NEW}})\} = E_\mu\{q(y_i)\}$ for $i = 1, 2, \ldots, n$. This last condition was automatically fulfilled for binary data because there we took $q(y_i) = 0$ for $y_i = 0$ or 1. In particular, (6.6) applies to situations where the y_i are correlated.

7. MALLOWS'S C_p AND CROSS-VALIDATION

In the ordinary least squares (OLS) situation, with normally distributed observations, linear models, and squared error prediction assessment, our theory coincides with Mallows's C_p approach (1973). It is easier to pinpoint the differences between Mallows's approach, which is the framework for our results, and cross-validation methods in the OLS context of this section.

The data vector y is now assumed to have an n-dimensional normal distribution with mean vector μ and covariance matrix $\sigma^2 I$, where μ is known to lie in a p-dimensional linear subspace $\mathcal{L}$,

$$y \sim N_n(\mu, \sigma^2 I), \qquad \mu \in \mathcal{L}. \tag{7.1}$$

Contained in $\mathcal{L}$ is a p_0-dimensional subspace $\mathcal{L}_0$, $p_0 \le p$. Let $\hat{\mu}$ and $\hat{\mu}_0$ denote the projections of y into $\mathcal{L}$ and $\mathcal{L}_0$, respectively, with corresponding estimates of σ^2,

$$\hat{\sigma}^2 = \|y - \hat{\mu}\|^2/(n - p), \quad \hat{\sigma}_0^2 = \|y - \hat{\mu}_0\|^2/(n - p_0). \tag{7.2}$$

The statistician is interested in the estimator $\hat{\mu}_0$, perhaps for prediction purposes, despite the possibility that $\mu \notin \mathcal{L}_0$. (Normality is not actually needed here; it is enough for the mean vector and covariance matrix to be as described in (7.1).)

Mallows's C_p is an estimate of prediction error for $\hat{\mu}_0$. Using error measure $Q[y, \hat{\mu}_0] = \|y - \hat{\mu}_0\|^2/n$, the true error rate of $\hat{\mu}_0$ is

$$\mathrm{Err} \equiv E_{\mathrm{NEW}} Q[y^{\mathrm{NEW}}, \hat{\mu}_0] = \frac{1}{n}\{\|\mu - \hat{\mu}_0\|^2 + n\sigma^2\}, \tag{7.3}$$

where $y^{\mathrm{NEW}} \sim N_n(\mu, \sigma^2 I)$ is independent of y. The statistic

$$C_p(\mathcal{L}_0, \mathcal{L}) \equiv \frac{1}{n}\{\|y - \hat{\mu}_0\|^2 + 2p_0\hat{\sigma}^2\} \tag{7.4}$$

is an unbiased estimator of Err, in the sense that both have the same expectation under model (7.1),

$$E_{\mu,\sigma^2}\{\mathrm{Err}\} = E_{\mu,\sigma^2}\{C_p(\mathcal{L}_0, \mathcal{L})\} = \frac{1}{n}\{\|\mu - \mu_0\|^2 + (n + p_0)\sigma^2\}, \tag{7.5}$$

μ_0 being the projection of μ into $\mathcal{L}_0$. Our statistic $C_p(\mathcal{L}_0, \mathcal{L})$ differs slightly from the usual definition of C_p (see Remark Q).

For the OLS situation, Theorem 2 gives

$$\omega(\mu, \sigma^2) = (2p_0/n)\sigma^2. \tag{7.6}$$

In this case, $\overline{\mathrm{err}} = Q[y, \hat{\mu}_0] = \|y - \hat{\mu}_0\|^2/n$. The obvious estimate of Err, as in (5.4), is

$$\hat{\mathrm{Err}} = \overline{\mathrm{err}} + \omega(\hat{\mu}, \hat{\sigma}^2) = \frac{1}{n}\{\|y - \hat{\mu}_0\|^2 + 2p_0\hat{\sigma}^2\}. \tag{7.7}$$

We see that the approach of the earlier sections results in $C_p(\mathcal{L}_0, \mathcal{L})$ when applied to the situation considered by Mallows.

Table 4 summarizes four different estimates of prediction error for the OLS situation. The naive C_p estimate $C_p(\mathcal{L}_0, \mathcal{L}_0)$ is just $C_p(\mathcal{L}_0, \mathcal{L})$ with $\mathcal{L} = \mathcal{L}_0$. In other words, we use $\mathcal{L}_0$ twice, both to define the prediction vector $\hat{\mu}_0$ and to estimate the true error rate Err for the rule $y \to \hat{\mu}_0$. In this sense $C_p(\mathcal{L}_0, \mathcal{L}_0)$ is similar to formula (2.4), or more exactly to (2.4) plus $\overline{\mathrm{err}}$, while $C_p(\mathcal{L}_0, \mathcal{L})$ is similar to (4.25) plus $\overline{\mathrm{err}}$ (see Remark L).

It is interesting that

$$E_{\mu,\sigma^2}\{C_p(\mathcal{L}_0, \mathcal{L}_0)\} \ge E_{\mu,\sigma^2}\{\mathrm{Err}\}, \tag{7.8}$$

with equality if and only if $\mu \in \mathcal{L}_0$. The naive estimator $C_p(\mathcal{L}_0, \mathcal{L}_0)$ tends to overestimate Err when the assumption $\mu \in \mathcal{L}_0$ is false. The *generalized cross-validation* estimate

$$\mathrm{GCV}(\mathcal{L}_0) \equiv \frac{1}{n}\left\{\frac{\|y - \hat{\mu}_0\|^2}{(1 - p_0/n)^2}\right\} = \frac{1}{1 - (p_0/n)^2} C_p(\mathcal{L}_0, \mathcal{L}_0) \tag{7.9}$$

introduced by Craven and Wahba (1979), overestimates Err slightly more.

Generalized cross-validation is a rotationally invariant form of the *cross-validation* estimate

$$\mathrm{CV}(\mathcal{L}_0) \equiv \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{\mu}_{0i}^{(i)}\}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}_{0i}}{1 - P_{ii}^0}\right)^2, \tag{7.10}$$

as motivated in Golub, Heath, and Wahba (1979). Here $\hat{\mu}_{0i}^{(i)}$ is the prediction for y_i calculated from the reduced data set, which omits y_i and its corresponding covariate vector, described more carefully below; and P_{ii}^0 is the iith diagonal element of the projection matrix p^0 into $\mathcal{L}_0$. The term "hat matrix" is often given to P_0.

The average value of P_{ii}^0 is

$$1/n \operatorname{tr} P^0 = p_0/n, \tag{7.11}$$

so $\mathrm{GCV}(\mathcal{L}_0)$ is just $\mathrm{CV}(\mathcal{L}_0)$ with the denominator $1 - P_{ii}^0$ replaced by its average value $1 - p_0/n$. This results in $E\{\mathrm{CV}(\mathcal{L}_0)\}$ usually exceeding $E\{\mathrm{GCV}(\mathcal{L}_0)\}$, because of Jenson's inequality. The inequality $E\{\mathrm{CV}\} \ge E\{\mathrm{GCV}\}$ is always true when $\mu \in \mathcal{L}_0$, and true in an average sense, averaging over spheres of constant $\|\mu - \mu_0\|$ value, when $\mu \notin \mathcal{L}_0$.

Table 4. Four Different Estimates of Prediction Error for the Ordinary Least Squares Situation

Name	Notation	Formula	Expectation (7.1)
1. C_p	$C_p(\mathcal{L}_0, \mathcal{L})$	$\frac{1}{n}\{\|y - \hat{\mu}_0\|^2 + 2p_0\hat{\sigma}^2\}$	$E\{\mathrm{Err}\} = \frac{1}{n}\{\|\mu - \mu_0\|^2 + (n + p_0)\sigma^2\}$
2. Naive C_p	$C_p(\mathcal{L}_0, \mathcal{L}_0)$	$\frac{1}{n}\{\|y - \hat{\mu}_0\|^2 + 2p_0\hat{\sigma}_0^2\} = \frac{n + p_0}{n}\hat{\sigma}_0^2$	$E\{\mathrm{Err}\} + \frac{2p_0}{n - p_0}\frac{\|\mu - \mu_0\|^2}{n}$
3. Generalized cross-validation	$\mathrm{GCV}(\mathcal{L}_0)$	$\frac{1}{n}\left\{\frac{\|y - \hat{\mu}_0\|^2}{(1 - p_0/n)^2}\right\} = \frac{n}{n - p_0}\hat{\sigma}_0^2$	$\frac{E\{C_p(\mathcal{L}_0, \mathcal{L}_0)\}}{1 - p_0^2/n^2}$
4. Cross-validation	$\mathrm{CV}(\mathcal{L}_0)$	$\frac{1}{n}\sum (y_i - \hat{\mu}_{0i}^{(i)})^2 = \frac{1}{n}\sum\left(\frac{y_i - \hat{\mu}_{0i}}{1 - P_{ii}^0}\right)^2$	$\doteq E\{\mathrm{Err}_+\}$ (6.13)

Table 5. First 10 Trials of a Sampling Experiment Comparing 6 Different Estimators of Err and Err_+ in an OLS Situation, and Summary Statistics for 20 Trials: The Bootstrap Estimate of Err_+ is Defined in Section 2 of Efron (1983); $C_p(\mathcal{L}_0, \mathcal{L}_5)$ is (7.4) With $\mathcal{L}_5$ the Space of Fifth Degree Polynomials in x

Trial	Err (7.3)	$\overline{\text{err}}$ (7.7)	$C_p(\mathcal{L}_0, \mathcal{L}_0) \doteq GCV$ (7.9)	$C_p(\mathcal{L}_0, \mathcal{L})$ (7.4)	$C_p(\mathcal{L}_0, \mathcal{L}_5)$ (Quintic)	CV (7.10)	$\hat{Err}_+^{(BOOT)}$ (B = 400)	Err_+ (7.14)
1	2.66	2.41	2.95	2.54	2.49	3.36	3.19	3.42
2	1.93	1.98	2.43	2.18	2.13	3.84	2.97	3.21
3	3.70	3.27	3.99	3.50	3.52	4.73	4.48	2.79
4	2.37	2.73	3.34	2.91	2.91	4.09	3.66	3.42
5	1.73	2.01	2.45	2.20	2.24	2.84	2.67	3.11
6	2.46	3.03	3.71	3.36	3.26	4.80	4.22	2.86
7	2.35	1.65	2.01	1.79	1.83	2.31	2.17	3.46
8	2.02	1.86	2.27	1.99	2.00	3.01	2.54	2.72
9	3.39	2.21	2.70	2.46	2.49	3.67	3.16	4.91
10	2.81	2.85	3.48	3.02	2.95	4.13	3.84	3.72
20 Trials AVE:	2.32	2.21	2.71	2.42	2.40	3.26	2.97	3.39
20 Trials (SD):	(.64)	(.67)	(.81)	(.69)	(.67)	(1.15)	(.98)	(.54)

It may seem strange that $CV(\mathcal{L}_0)$ is biased upward for $E\{\text{Err}\}$, given how plausible $CV(\mathcal{L}_0) = (1/n)\sum_{i=1}^n \{y_i - \hat{\mu}_{0i}^{(i)}\}^2$ looks as an estimator of Err. In fact $CV(\mathcal{L}_0)$ is estimating a somewhat different quantity, which we now describe.

Suppose as in (6.2) that $\mu_i = t_i'\alpha$ for $i = 1, 2, \ldots, n$. The $p \times n$ matrix $T = (t_1, t_2, \ldots, t_n)$ must have row space $\mathcal{L}_{\text{row}}(T) = \mathcal{L}$, in accordance with (7.1). Suppose also that we can partition t_i and α into $t_i' = (t_{0i}', t_{1i}')$ and $\alpha' = (\alpha_0', \alpha_1')$ as in Remark L, where t_{0i} and α_0 are of dimension p_0, and that the $p_0 \times n$ matrix $T_0 = (t_{01}, t_{02}, \ldots, t_{0n})$ has $\mathcal{L}_{\text{row}}(T_0) = \mathcal{L}_0$. Then the projection $\hat{\mu}_0$ of y into $\mathcal{L}_0$ is given by

$$\hat{\mu}_0 = P^0 y, \qquad P^0 = T_0'(T_0T_0')^{-1}T_0. \tag{7.12}$$

Equivalently we can describe the prediction rule $\hat{\mu}_0$ by

$$\hat{\mu}_{0i} = t_{0i}'\hat{\alpha}_0, \qquad \hat{\alpha}_0 = (T_0T_0')^{-1}T_0 y. \tag{7.13}$$

The appropriate context for cross-validation is that where the pairs (t_i, y_i), $i = 1, 2, \ldots, n$, are independently selected according to some joint probability distribution F on $(p + 1)$-dimensional space. Now suppose that one more independent pair is obtained from F, say (t_+, y_+). The predicted value for y_+ based on the original rule (7.13) is $\hat{\mu}_0(t_+) = t_{0+}'\hat{\alpha}_0$. The expected squared error of prediction is

$$\text{Err}_+ \equiv E_+\{y_{0+} - t_{0+}'\hat{\alpha}_0\}^2, \tag{7.14}$$

E_+ indicating expectation over (t_+, y_+), with $\hat{\alpha}_0$ fixed.

The expected value of $CV(\mathcal{L}_0)$ tends toward $E_{\mu,\sigma^2}\{\text{Err}_+\}$ rather than $E_{\mu,\sigma^2}\{\text{Err}\}$. The predictor $\hat{\mu}_{0i}^{(i)}$ appearing in (7.10) equals $t_{0i}'\hat{\alpha}_0^{(i)}$, where $\hat{\alpha}_0^{(i)} = (T_{0(i)}T_{0(i)}')^{-1}T_{0(i)}y_{(i)}$, $T_{0(i)} = (t_{01}, \ldots, t_{0,i-1}, t_{0,i+1}, \ldots, t_{0,n})$, and $y_{(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$. We see that $E_{\mu,\sigma^2}\{CV(\mathcal{L}_0)\}$ equals the expected value of Err_+, for sample size $n - 1$ rather than n.

The results of a small sampling experiment are reported in Tables 5 and 6. The data for each trial of the experiment comprised 20 pairs (x_i, y_i), generated as follows:

$$x_i \sim N(0, 10^2), \quad \mu_i = x_i + .01x_i^2, \quad y_i \sim N(\mu_i, 1), \qquad i = 1, 2, \ldots, 20. \tag{7.15}$$

The spaces $\mathcal{L}$ and $\mathcal{L}_0$ were taken to be those associated with quadratic and linear regression, respectively. In other words, $\hat{\mu}_{0i} = \hat{\alpha}_{00} + \hat{\alpha}_{01}x_{i1}$, the simple linear regression of y_i on x_i based on the data (x_i, y_i), $i = 1, \ldots, 20$; whereas $\hat{\sigma}^2$ in (7.2) was based on a quadratic regression for $\hat{\mu}$, dimension $p = 3$.

Twenty trials of (7.15) were run. Notice in Table 5 that Err_+ is usually larger than Err. This is no surprise; Err_+ is the prediction error for a completely new pair (x_+, y_+), whereas Err is the average prediction error for a new pair having x^{NEW} equal to one of the 20 original x_i values. It is easier to predict in the latter case because the (x, y) pairs are nearer the training set $\{(x_i, y_i), i = 1, \ldots, 20\}$. See Section 6 of Efron (1983).

Table 6 shows how well the six estimators performed in the sampling experiment. Two mean squared errors are shown, MSE (5.5) and MSE_+, which is (5.5) with Err_+ replacing Err.

Several facts are worth mentioning: it is much easier to estimate Err than Err_+; cross-validation is a better estimator of Err_+ than Err, though not wonderful in either case; the bootstrap estimation for Err_+ described in Efron (1983) does somewhat better in both cases; $C_p(\mathcal{L}_0, \mathcal{L})$ does very well in estimating Err, considering it is an unbiased estimator; $\overline{\text{err}}$ does even better in this case, but the MSE criterion favors estimators that are biased downwards; $C_p(\mathcal{L}_0, \mathcal{L}_5)$, based on an overly large choice of $\mathcal{L}$ in (7.4), performs just as well as $C_p(\mathcal{L}_0, \mathcal{L})$ using the correct choice of $\mathcal{L}$.

Table 6. How Well the Six Estimators in Table 5 Estimated Err and Err_+ in the 20 Trials of the Sampling Experiment: MSE is Defined at (5.5); MSE_+ Is the Corresponding Mean Squared Error for Err_+; It Is Much Easier to Estimate Err

	$\overline{\text{err}}$ (7.7)	$C_p(\mathcal{L}_0, \mathcal{L}_0) \doteq GCV$ (7.9)	$C_p(\mathcal{L}_0, \mathcal{L})$ (7.4)	$C_p(\mathcal{L}_0, \mathcal{L}_5)$ (Quintic)	CV (7.10)	$\hat{Err}_+^{(BOOT)}$ (B = 400)
MSE:	.31	.54	.35	.33	1.60	.92
MSE_+:	2.09	1.38	1.71	1.70	1.47	1.33

The main point is that it is easier to estimate Err than Err_+, and that $C_p(\hat{\mu}_0, \hat{\mu})$ is the estimator of choice for Err. The logistic regression experiment in Section 5 reached a similar conclusion, with $C_p(\hat{\mu}_0, \hat{\mu})$ replaced by its binary data analogue (5.4).

Remark T. The difference $\text{Err}_+ - \text{Err}$ is only about .005 in the experiment of Section 5, too small to be apparent in Table 3.

Remark U. Unlike $C_p(\hat{\mu}_0, \hat{\mu})$, neither cross-validation nor $\hat{\text{Err}}_+^{(\text{BOOT})}$ require the statistician to name a space $\hat{\mu}$ guaranteed to contain the mean vector μ. However, they estimate a different quantity from $C_p(\hat{\mu}_0, \hat{\mu})$, Err_+ rather than Err, and with less efficiency.

Remark V. Which quantity is more relevant, Err or Err_+? Arguments can be made both ways, depending on the context, but for comparing different possible models $\hat{\mu}_0$, efficiency of the error estimation is the primary consideration. This offers some pragmatic ground for preferring C_p to cross-validation, though the evidence so far is by no means overwhelming.

Remark W. Despite its name, $\text{GCV}(\hat{\mu}_0)$ is (nearly) a member of the C_p family of estimates.

Remark X. The cross-validation estimate $\text{CV}(\hat{\mu}_0)$ depends on the coordinate system in which $\hat{\mu}_0$ and y are expressed. We can get an invariant version of $\text{CV}(\hat{\mu}_0)$ by averaging (7.10) over a uniform choice among all possible orthogonal coordinate systems. The invariant version of $\text{CV}(\hat{\mu}_0)$ turns out to equal

$$((n - p_0)/(n - p_0 - 2))\text{GCV}(\hat{\mu}_0). \qquad (7.16)$$

This calculation is close to the one in Golub, Heath, and Wahba (1979), which gives exactly $\text{GCV}(\hat{\mu}_0)$, except that they average over a group that includes complex-valued rotations.

Remark Y. The bootstrap estimate $\hat{\text{Err}}_+^{(\text{BOOT})}$ in Tables 5 and 6 is not the analogue of the bootstrap method for binary data described in Remark J. The bootstrap argument of Remark J, applied to the OLS situation of this section, gives exactly $C_p(\hat{\mu}_0, \hat{\mu})$.

[Received April 1985. Revised July 1985.]

REFERENCES

Atkinson, A. C. (1980), "A Note on the Generalized Information Criterion for Choice of a Model," *Biometrika*, 67, 413–418.

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerische Mathematik*, 31, 377–403.

Efron, B. (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of the American Statistical Association*, 70, 892–898.

——— (1978a), "The Geometry of Exponential Families," *Annals of Statistics*, 6, 362–376.

——— (1978b), "Regression and ANOVA With Zero–One Data: Measures of Residual Variation," *Journal of the American Statistical Association*, 73, 113–121.

——— (1982), "Maximum Likelihood and Decision Theory," *Annals of Statistics*, 10, 340–356.

——— (1983), "Estimating the Error Rate of a Prediction Rule: Improvements on Cross-Validation," *Journal of the American Statistical Association*, 78, 316–331.

Golub, G., Heath, M., and Wahba, G. (1979), "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21, 215–223.

Mallows, C. L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675.

McCullagh, P., and Nelder, J. (1983), *Generalized Linear Models*, New York: Chapman & Hall.

Stone, M. (1977), "An Asymptotic Equivalence of Choice of Model by Cross-Validation of Akaike's Criterion," *Journal of the Royal Statistical Society*, Ser. B, 39, 44–47.

Prediction using Logistic Regression

Under the logistic model for subject i,

$$P(Y_i = 1|\mathbf{x}_i) = p_i = \frac{e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i}}{1 + e^{\beta_0 + \boldsymbol{\beta}_1' \mathbf{x}_i}}.$$

- The estimated 'risk score' is $\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1' \mathbf{x}_i$, which is the log-odds, also referred to as the linear predictor.

- The estimated or predicted probabilities of response is

$$\widehat{p_i} = \frac{e^{\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1' \mathbf{x}_i}}{1 + e^{\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1' \mathbf{x}_i}}.$$

```
> summary(predict(icu.fit))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-19.005  -2.160  -1.625  -1.971  -1.357  33.657

> summary(predict(icu.fit, type = "response"))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.1034  0.1646  0.2000  0.2048  1.0000
```

Earlier there was warning:

```
> icu.fit <- glm(sta ~ gender + age + race + loc,
family = binomial(), data = icu)

Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Classification problem

Beyond the predicted probabilities of response, sometimes we would also like to predict whether the response is actually 0 or 1.

- This is also referred to as *classification*.

  Q: how do we do this?

- Based on $\widehat{p}_i$, we can choose a cufoff, eg. $C = 0.5$, and let

$$\widehat{Y}_i = \begin{cases} 1, \text{ if } \widehat{p}_i > C, \\ 0, \text{ if } \widehat{p}_i \leq C. \end{cases}$$

- Then the *prediction error*, in this case *counting error*, is

$$Q(Y_i, \widehat{Y}_i) = \begin{cases} 1, \text{ if } Y_i \neq \widehat{Y}_i, \\ 0, \text{ if } Y_i = \widehat{Y}_i. \end{cases}$$

- The errors can be divided into
  - *false positive*: $Y_i = 0$ but $\widehat{Y}_i = 1$;
  - *false negative*: $Y_i = 1$ but $\widehat{Y}_i = 0$.

- There are also:
  - *true positive*: $Y_i = \widehat{Y}_i = 1$;
  - *true negative*: $Y_i = \widehat{Y}_i = 0$.

- The true positive rate, $\sum_{i=1}^{n} I(Y_i = \widehat{Y}_i = 1)/n$, where $I(\cdot)$ is the indicator function, is called *sensitivity*;

- and the true negative rate, $\sum_{i=1}^{n} I(Y_i = \widehat{Y}_i = 0)/n$, is called *specificity*.

- The terms sensitivity and specificity are often used in medical diagnostics, where positive means there is a disease and negative otherwise; but these terms are also used more broadly.

- A good classification procedure should have both high sensitivity and high specificity.

- The cutoff C used for classification under logistic regression does not have to be 0.5.

- As C varies from 0 to 1, fewer cases are predicted or classified to be from the class '1'.

- Therefore the true positive rate or sensitivity goes ... up or down?

- Meanwhile as C varies from 0 to 1, more cases are predicted or classified to be from the class '0'.

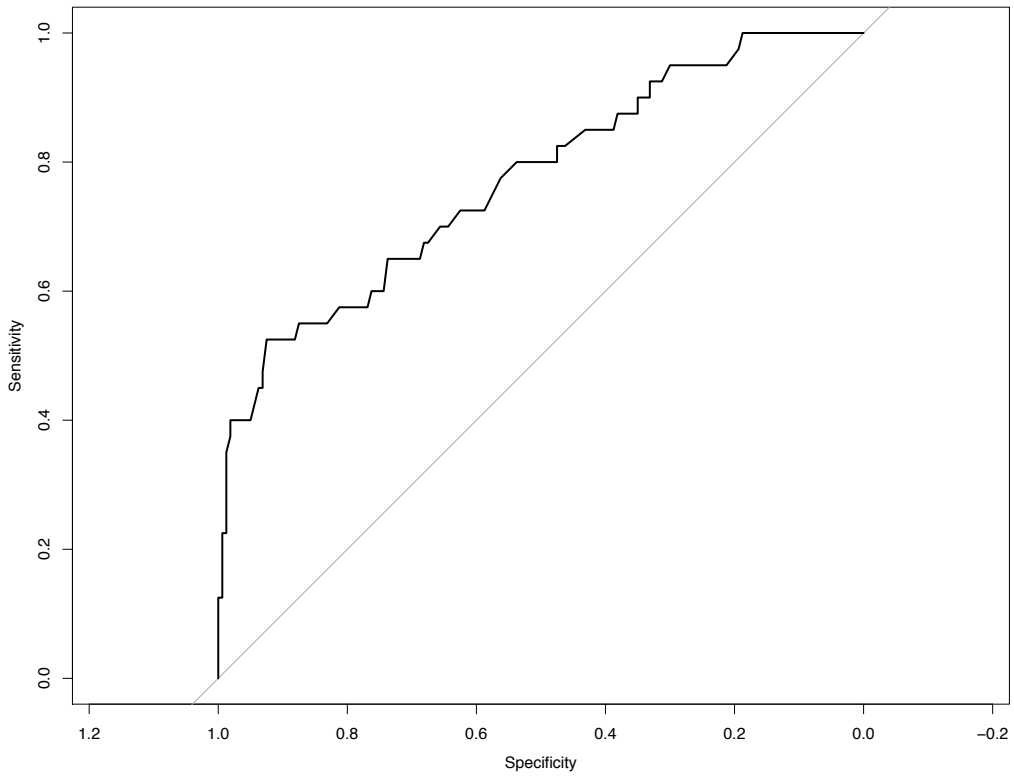
- Therefore the true negative rate or specificity goes which way?

- So there is typically trade-off between sensitivity and specificity.

ROC Curve

As the cutoff C varies from 0 to 1, one can plot the sensitivity against 1− specificity. This is called the receiver operating characteristic (ROC) curve.

```
> install.packages("pROC")
> library("pROC", lib.loc="~/Library/R/3.5/library")
> plot(roc(icu$sta, predict(icu.fit, type = "response")))
## or
> plot.roc(icu$sta, predict(icu.fit, type = "response"))
```



- The ROC curve goes through (0, 0) and (1, 1) (why).

- The diagonal line from (0, 0) to (1, 1) is often plotted to represent 'random guess'.

  Read https://www.r-bloggers.com/what-it-the-interpretation-of-the-diagonal-for-a-roc-curve/ for details on the diagonal ROC curve.

- Any classification procedure with some predictive power should have its ROC curve above the diagonal line.

- As a summary measure of performance, area under the curve (AUC) is computed.

```
> auc(icu$sta, predict(icu.fit, type = "response"))
Setting levels: control = Lived, case = Died
Setting direction: controls < cases
Area under the curve: 0.7718
```

  Q: what is a reasonable range of AUC?

Prediction of a future observation

Having fitted the logistic regression and obtained the parameter estimates, we can also predicted the probability of response for a future observation with $\mathbf{x}$:

$$\widehat{P}(Y = 1|\mathbf{x}) = \widehat{p}(\mathbf{x}) = \frac{e^{\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1' \mathbf{x}}}{1 + e^{\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_1' \mathbf{x}}}.$$

(R function 'predict()' can take 'newdata' that you provide.)

- We can similarly classify the future Y using a cufoff C, and let

$$\widehat{Y}(\mathbf{x}) = \begin{cases} 1, \text{ if } \widehat{p}(\mathbf{x}) > C, \\ 0, \text{ if } \widehat{p}(\mathbf{x}) \leq C. \end{cases}$$

- The prediction error, i.e. counting error, is:

$$Q(Y, \widehat{Y}(\mathbf{x})) = \begin{cases} 1, \text{ if } Y \neq \widehat{Y}(\mathbf{x}), \\ 0, \text{ if } Y = \widehat{Y}(\mathbf{x}). \end{cases}$$

True and Apparent error rate

- To make very clear the distinction between the future and the original data, from here on we will denote the future data as Y^* and write $Q(Y^*, \widehat{Y}(\mathbf{x}^*))$ which is the error or *loss*.

- The error rate, or *risk*, is then $E^*\{Q(Y^*, \widehat{Y}(\mathbf{x}))\}$, where E^* is with respect to Y^*. This is also what B. Efron in a series of papers called *true error rate*.

  Q: how do we estimate the true error rate?

- If we have a future random sample of Y_j^*, $j = 1, ..., m$, we can estimate the risk by $\sum_{j=1}^{m} Q(Y_j^*, \widehat{Y}(\mathbf{x}_j^*))/m$.

- But what we have immediately at hand is the *'apparent' error rate*: $\sum_{i=1}^{n} Q(Y_i, \widehat{Y}_i)/n$.

- Efron showed in a series of papers that the apparent error rate '$\bar{e}rr$' **underestimates** the true error rate 'Err'; see Efron (1986) Table 3.

- The difference between the two he calls *optimism*: $Err - \bar{e}rr$.

Test-training samples

Q: how do we correct for the optimism, i.e. **bias**?

- In some cases, analytic expressions may be derived. (AIC is an example, with deviance loss, and the correction asymptotically is $2p$.)

- But more generally, *test-training samples* may be used:
  - Use the training sample to build the model, and estimate the parameters;
  - use the test sample as future data to estimate the error rate.

- We can randomly split the original data into test and training samples; eg. 75% to train, and rest 25% to test.

- But this split reduces the sample size.

Cross-validation

- Instead of split into test-training samples, *resampling* methods might be used, i.e. sampling from the original data.

- A *K-fold cross-validation* (CV) randomly divides the original data into K approximately equal-sized parts;

- the procedure then takes turns to use:
  - each of these parts as a test sample;
  - the rest as the training sample.

- The K-fold CV estimate of the error rate is

$$\hat{Err}^{cv} = \frac{1}{n} \sum_{i=1}^{n} Q(Y_i, \widehat{Y}^{-k(i)}(\mathbf{x}_i)),$$

  where $\widehat{Y}^{-k(i)}$ denotes the prediction using the training sample with the part containing subject i removed.

- When $K = n$, this is also called the leave-one-out CV.

- Leave-one-out CV gives approximately unbiased estimate of the error rate, but it has large variance because the training samples are largely the same.

- K-fold CV with smaller K tends to have smaller variance, at the expense of larger training-set-size bias.

- $K = 10$ is often used; and multiple (eg. 10) runs of 10-fold CV can further improve the stability due to random splitting.

- See again Efron (1986) Table 3.

  Note: test-training sample and CV are general procedures, and can be applied to other quantities than counting error; eg. *out-of-sample* AUC, R^2 etc.

From *High-Dimensional Data Analysis in Cancer Research* (Li and Xu eds., center color pages), performance of different methods:

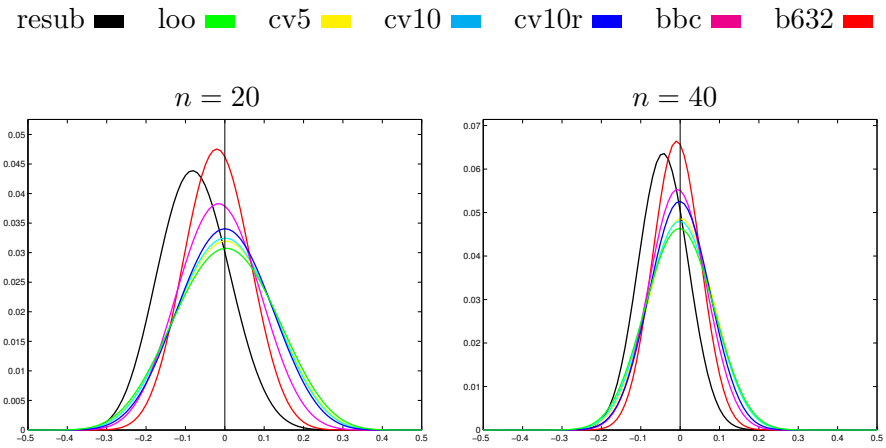


Figure 1: Beta-fits of empirical deviation distribution. resub: "apparent" estimate; loo: leave-one-out; cv10r: repeated CV10; bbc: bias-corrected bootstrap; b632: .632 bootstrap.

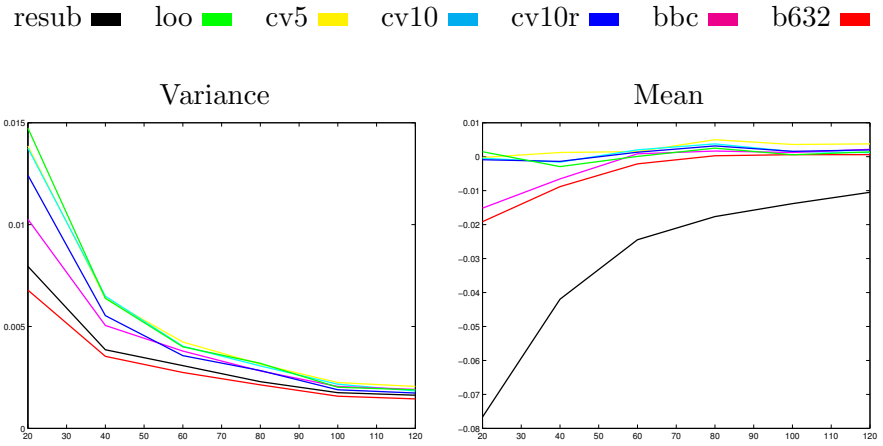


Figure 2: Plots of the empirical deviation distribution.

Loss and Risk

- In general, for $y = 0, 1$,

$$Q(Y, y) = \begin{cases} 1, & \text{if } Y \neq y, \\ 0, & \text{if } Y = y \end{cases}$$

  is a *loss function*.

- The expected loss, $E\{Q(Y, y)\}$, is called the *risk function*.

- The risk is a measurement of the prediction performance.

- Note that the expectation in the risk is with respect to Y, while y is deterministic.

- If we have a random sample of Y_j^*, $j = 1, ..., m$, from the distribution of Y, we can estimate the risk by $\sum_{j=1}^{m} Q(Y_j^*, y)/m$.

- Here we use the '*' to distinguish 'future' $Y_1^*, ..., Y_m^*$ from the original data $Y_1, ..., Y_n$.

Expected Risk

- And now, **instead of y above**, we are using the prediction $\widehat{Y}$ that is a function of the original data $(\mathbf{x}_1, Y_1), ..., (\mathbf{x}_n, Y_n)$ (which have given us the estimate β's)

- We can plug in $\widehat{Y}$ for y in the loss to have $Q(Y, \widehat{Y})$, but how should we understand the risk?

- To make very clear the distinction between the future and the original data, from here on we write the loss as $Q(Y^*, \widehat{Y}(\mathbf{x}))$.

- The risk is then $E^*\{Q(Y^*, \widehat{Y}(\mathbf{x}))\}$, where E^* is with respect to Y^*. This is also what B. Efron in a series of papers called *true error rate*.

- The above risk is a random quantity, as it depends on the original data through $\widehat{Y}$.

- The *expected risk* is defined as $E[E^*\{Q(Y^*, \widehat{Y}(\mathbf{x}))\}]$, where E is with respect to the original data in $\widehat{Y}$.

  Q: how do we estimate the expected risk?

BIOMETRICS 58, 305–315
June 2002

Survival Analysis with Time-Varying Regression Effects Using a Tree-Based Approach

Ronghui Xu

Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, Massachusetts 02115, U.S.A.
email: rxu@jimmy.harvard.edu

and

Sudeshna Adak

Bioinformatics Group, IBM India Research Lab, Indian Institute of Technology, New Delhi, India 110 016
email: asudeshn@in.ibm.com

SUMMARY. Nonproportional hazards often arise in survival analysis, as is evident in the data from the International Non-Hodgkin's Lymphoma Prognostic Factors Project. A tree-based method to handle such survival data is developed for the assessment and estimation of time-dependent regression effects under a Cox-type model. The tree method approximates the time-varying regression effects as piecewise constants and is designed to estimate change points in the regression parameters. A fast algorithm that relies on maximized score statistics is used in recursive segmentation of the time axis. Following the segmentation, a pruning algorithm with optimal properties similar to those of classification and regression trees (CART) is used to determine a sparse segmentation. Bootstrap resampling is used in correcting for overoptimism due to split point optimization. The piecewise constant model is often more suitable for clinical interpretation of the regression parameters than the more flexible spline models. The utility of the algorithm is shown on the lymphoma data, where we further develop the published International Risk Index into a time-varying risk index for non-Hodgkin's lymphoma.

KEY WORDS: Change point; Classification and regression tree; Maximized score test; Nonproportional hazards; Time-varying regression effect.

1. Introduction

Despite the enormous success of the proportional hazards regression model (Cox, 1972, 1975), starting in the early 1980s, many authors have noted violations of the proportional hazards assumption in certain applications (Lancaster and Nickell, 1980; Gail, Wieand, and Piantadosi, 1984; Struthers and Kalbfleisch, 1986; Bretagnolle and Huber-Carol, 1988; Ford, Norrie, and Ahmadi, 1995). This occurs often under clinical settings when important prognostic variables measured at baseline are used to predict survival of the patients. For example, in the International Non-Hodgkin's Lymphoma Prognostic Factors Project (Shipp, Harrington, and Anderson, 1993), clinical features that were considered predictive of overall survival were evaluated for 3273 adult patients with aggressive non-Hodgkin's lymphoma from 16 institutions and cooperative groups in the United States, Europe, and Canada who were treated between 1982 and 1987 with combination-chemotherapy regimens containing doxorubicin. The project was undertaken to develop a model for identifying patients as having high or low risk for survival based on the clinical characteristics prior to treatment. The identification of these groups was of utmost importance to individual patients and their physicians. At the conclusion of the project, five important prognostic variables were identified and an international risk index was formed, which has nowadays been widely adopted in non-Hodgkin's lymphoma. The international index assigned patients to one of four risk groups based on the number of risk factors present at diagnosis (age > 60, Ann Arbor stage III or IV, serum lactate dehydrogenase [LDH] $> 1.5 \times$ normal, performance status of bedridden, and the number of extranodal disease sites >1): 0 or 1 risk factor = low risk, 2 risk factors = low intermediate risk, 3 risk factors = high intermediate risk, and 4 or 5 risk factors = high risk. When patients from the study were assigned to one of the four risk groups, the estimated survival curves (Shipp et al., 1993, Figure 1, left panels) were distinctly different, indicating good predictability of the International Risk Index. The right panels of the same figure from Shipp et al. (1993) showed plots of death rates from each of the four risk groups, which clearly indicated crossing hazards, i.e., time-varying hazard ratios for the four risk groups.

A straightforward generalization of the proportional hazards model to accommodate time-varying regression effects is given by

$$\lambda(t \mid \boldsymbol{Z}) = \lambda_0(t)\exp(\boldsymbol{\beta}(t)'\boldsymbol{Z}), \tag{1}$$

where $\lambda(\cdot \mid \boldsymbol{Z})$ is the conditional hazard function given the covariate vector $\boldsymbol{Z}$, $\lambda_0(\cdot)$ is the baseline hazard, and $\boldsymbol{\beta}(\cdot)$ the vector of time-varying regression coefficients. Note that model (1) is completely general in a two-sample case since $\boldsymbol{\beta}(t)$ would simply be the log hazard ratios between the two groups; this is also true for any $k(>2)$-sample cases.

For model (1), Sleeper and Harrington (1990), Gray (1992), and Hastie and Tibshirani (1993) used splines to estimate the time-varying regression coefficients $\boldsymbol{\beta}(t)$; Pettitt and Bin Daud (1990) and Grambsch and Therneau (1994) considered smoothing the Schoenfeld residuals to estimate time-varying regression effects; Verweij and Van Houwelingen (1995) estimated the values of $\boldsymbol{\beta}(t)$ at each observed failure time using penalized partial likelihood; Sargent (1997) and Gustafson (1998) considered Bayesian approaches for the same problem; and Zucker and Karr (1990) and Murphy and Sen (1991) studied the asymptotic properties of the spline and sieves estimates of $\boldsymbol{\beta}(t)$, respectively. The last two papers, however, did not provide guidelines for estimating $\boldsymbol{\beta}(t)$ in practice. For the non-Hodgkin's lymphoma data, Figure 3 shows the estimated log hazard ratios using piecewise constant splines (Gray, 1992) and loess smoothing of the Schoenfeld residuals (Grambsch and Therneau, 1994). These curves give a general idea of the time-varying effect of these covariates but are difficult to summarize for clinical use and to combine into a time-varying risk index.

Following the original spirit of the project, in this article, we aim to approximate the time-varying regression effects in a somewhat simplistic fashion that is more suitable for clinical applications—the step-function approach. In taking such an approach, we join many other authors in noticing that any model is only an approximation to the much more complex reality. Note also that, among the references mentioned above, Gray (1992) recommended piecewise constant over quadratic or cubic splines for estimating $\boldsymbol{\beta}(t)$ because of its advantages in computational speed, complexity, and stability in the right tail. Verweij and Van Houwelingen (1995) and Murphy and Sen (1991) also assumed piecewise constants but in a much more dense way. The algorithm described in this article is based on tree methods, which were first introduced by Morgan and Sonquist (1963) and then popularized by Breiman et al. (1984). Since then, the tree method has been adopted for right-censored data by many authors, though not (as we are aware) aimed at estimating time-varying regression effects. References on tree-methods for survival data can be found in Zhang and Singer (1999) and LeBlanc (2001).

In the next section, we describe a maximized score statistic that can be used to find a single change point. Section 3 shows how a recursive segmentation of the time axis can be obtained by repeated use of the maximized score statistic. The recursive segmentation is continued until a large number of segments is obtained. Some of the segments are then recombined using a pruning algorithm similar to that of Breiman et al.'s (1984) classification and regression trees (CART). The final selection of the change points is obtained by a bootstrap procedure that corrects for overoptimism due to split point optimization, as described in Section 4. Via simulation experiments in Section 5, we study the performance of this tree-based approach in estimating the time-varying regression effects. We return to the International Non-Hodgkin's Lymphoma Prognostic Factors Project data in Section 6 with the results from this tree-based method, which extends the published International Risk Index to a time-varying risk index for non-Hodgkin's lymphoma. The last section contains some further discussion.

2. Optimal Selection of a Single Change Point Using a Maximized Score Test

O'Quigley and Pessione (1991) proposed a test for the equality of two survival distributions against the alternative of crossing hazards. Here we generalize the test statistic to compare the regression effects on two intervals, $(0, \gamma]$ and (γ, τ), where γ is fixed for the time being and τ is an upper limit of time, which could be either finite or infinite. The approach described next is also similar to Liang, Self, and Liu (1990). Assume that

$$\boldsymbol{\beta}(t) = \begin{cases} \boldsymbol{\beta} + \boldsymbol{\phi}, & t < \gamma \\ \boldsymbol{\beta} - \boldsymbol{\phi}, & t \geq \gamma. \end{cases} \tag{2}$$

Denote T the failure time random variable of primary interest, C a censoring time random variable, and $\boldsymbol{Z}$ a vector of p covariates. Let $X = \min(T, C)$, $\delta = I(T \leq C)$, and $Y(t) = I(X \geq t)$. From i.i.d. observations $(X_i, \delta_i, \boldsymbol{Z}_i)_{i=1}^n$, the partial likelihood for $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ can be written

$$L(\boldsymbol{\beta}, \boldsymbol{\phi}) = \prod_{i|\delta_i=1} \frac{e^{\{\boldsymbol{\beta}+[I(X_i<\gamma)-I(X_i\geq\gamma)]\boldsymbol{\phi}\}'\boldsymbol{Z}_i}}{\sum_{j=1}^{n} Y_j(X_i) e^{\{\boldsymbol{\beta}+[I(X_i<\gamma)-I(X_i\geq\gamma)]\boldsymbol{\phi}\}'\boldsymbol{Z}_j}}, \tag{3}$$

where $I(\cdot)$ is the indicator function.

We make use of the score test, also called the Rao (1973) test, for testing the constancy of $\boldsymbol{\beta}(t)$ over $[0, \tau]$, i.e., H_0: $\boldsymbol{\phi} = \mathbf{0}$. The score test statistic is given by

$$S(\gamma) = \boldsymbol{U}(\hat{\boldsymbol{\beta}}, \mathbf{0})' \boldsymbol{I}(\hat{\boldsymbol{\beta}}, \mathbf{0})^{-1} \boldsymbol{U}(\hat{\boldsymbol{\beta}}, \mathbf{0}), \tag{4}$$

where $\hat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$ obtained by maximizing the partial likelihood under H_0 and $\boldsymbol{U}(\hat{\boldsymbol{\beta}}, \mathbf{0})$, and $\boldsymbol{I}(\hat{\boldsymbol{\beta}}, \mathbf{0})$ are derived from the log partial likelihood and represent the appropriate first derivative and information matrix evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}, \boldsymbol{\phi} = \mathbf{0}$. It can be easily shown that

$$\begin{aligned} \boldsymbol{U}(\hat{\boldsymbol{\beta}}, \mathbf{0}) &= \frac{\partial \log L(\boldsymbol{\beta}, \boldsymbol{\phi})}{\partial \boldsymbol{\phi}} \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \boldsymbol{\phi}=\mathbf{0}} \\ &= \sum_{i|X_i<\gamma, \delta_i=1} \boldsymbol{r}_i - \sum_{i|X_i\geq\gamma, \delta_i=1} \boldsymbol{r}_i, \end{aligned} \tag{5}$$

where

$$\boldsymbol{r}_i = \boldsymbol{Z}_i - \frac{\sum_{j=1}^n Y_j(X_i)\boldsymbol{Z}_j e^{\hat{\boldsymbol{\beta}}'\boldsymbol{Z}_j}}{\sum_{j=1}^n Y_j(X_i) e^{\hat{\boldsymbol{\beta}}'\boldsymbol{Z}_j}}$$

are the Schoenfeld (1982) residuals;

$$\boldsymbol{I}(\hat{\boldsymbol{\beta}}, \mathbf{0}) = (\boldsymbol{I}_{\boldsymbol{\phi}\boldsymbol{\phi}} - \boldsymbol{I}_{\boldsymbol{\phi}\boldsymbol{\beta}} \boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} \boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\phi}}) \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \boldsymbol{\phi}=\mathbf{0}}, \tag{6}$$

$$\boldsymbol{I}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\hat{\boldsymbol{\beta}}, \mathbf{0}) = -\frac{\partial^2 \log L(\boldsymbol{\beta}, \boldsymbol{\phi})}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'} \Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \boldsymbol{\phi}=\mathbf{0}} = \sum_{i|\delta_i=1} \boldsymbol{v}_i, \tag{7}$$

$$\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}},\mathbf{0}) = -\frac{\partial^2 \log L(\boldsymbol{\beta},\boldsymbol{\phi})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\boldsymbol{\phi}=\mathbf{0}} = \sum_{i|\delta_i=1} \boldsymbol{v}_i, \quad (8)$$

$$\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\phi}}(\hat{\boldsymbol{\beta}},\mathbf{0}) = \boldsymbol{I}'_{\boldsymbol{\phi}\boldsymbol{\beta}}(\hat{\boldsymbol{\beta}},\mathbf{0}) = -\frac{\partial^2 \log L(\boldsymbol{\beta},\boldsymbol{\phi})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\phi}'} = \sum_{X_i<\gamma,\delta_i=1} \boldsymbol{v}_i - \sum_{X_i\geq\gamma,\delta_i=1} \boldsymbol{v}_i, \quad (9)$$

where

$$\boldsymbol{v}_i = \frac{\sum_{j=1}^{n} Y_j(X_i)\boldsymbol{Z}_j^{\otimes 2} e^{\hat{\boldsymbol{\beta}}'\boldsymbol{Z}_j}}{\sum_{j=1}^{n} Y_j(X_i) e^{\hat{\boldsymbol{\beta}}'\boldsymbol{Z}_j}} - \left(\frac{\sum_{j=1}^{n} Y_j(X_i)\boldsymbol{Z}_j e^{\hat{\boldsymbol{\beta}}'\boldsymbol{Z}_j}}{\sum_{j=1}^{n} Y_j(X_i) e^{\hat{\boldsymbol{\beta}}'\boldsymbol{Z}_j}}\right)^{\otimes 2},$$

with $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a}\boldsymbol{a}'$ for a vector $\boldsymbol{a}$. The score statistic $S(\gamma)$ given by (4) has an asymptotic χ^2_p distribution under H_0 for fixed γ. To find the best change point γ_1, we choose the γ that maximizes $S(\gamma)$ over $(0,\tau)$. The use of maximally selected chi-square statistics in detecting change points has been seen in the literature and its asymptotic properties studied (cf., Davies, 1977, 1987; Miller and Siegmund, 1982; Csörgő and Horváth, 1997).

In practice, however, we use a modified score statistic that gives better performance when combined with the tree method. It can be easily shown that

$$\boldsymbol{I}(\hat{\boldsymbol{\beta}},\mathbf{0}) = 4\left[\sum_{X_i<\gamma,\delta_i=1} \boldsymbol{v}_i - \left(\sum_{X_i<\gamma,\delta_i=1} \boldsymbol{v}_i\right)\left(\sum_{\delta_i=1} \boldsymbol{v}_i\right)^{-1} \times \left(\sum_{X_i<\gamma,\delta_i=1} \boldsymbol{v}_i\right)\right]. \quad (10)$$

In addition, it is known (Kim and Tsiatis, 1990; Xu and O'Quigley, 2000) that the $\boldsymbol{v}_i$'s are approximately constants, so that $\Sigma_{X_i\geq\gamma,\delta_i=1}\boldsymbol{v}_i \approx [n(\gamma)/N]\,\Sigma_{\delta_i=1}\boldsymbol{v}_i$, where $n(\gamma) = \Sigma_1^n \delta_i Y_i(\gamma)$ is the number of events to the right of γ and $N = \Sigma_1^n \delta_i$ is the total number of events. This leads to the approximation

$$\boldsymbol{I}(\hat{\boldsymbol{\beta}},\mathbf{0}) \approx 4\left[\frac{n(\gamma)}{N}\cdot\frac{N-n(\gamma)}{N}\left(\sum_{\delta_i=1} \boldsymbol{v}_i\right)\right]; \quad (11)$$

therefore, $\boldsymbol{I}(\hat{\boldsymbol{\beta}},\mathbf{0})$ tends to be small near zero or τ. This causes the well-known end-cut preference phenomenon for tree-based procedures, where a split tends to occur at one of the two ends of the data. One way to reduce such tendency is by multiplying $S(\gamma)$, or equivalently, dividing $\boldsymbol{I}(\hat{\boldsymbol{\beta}},\mathbf{0})$, by $p_{\text{left}}(\gamma) \times p_{\text{right}}(\gamma)$ (see Breiman et al., 1984, p. 316), where

$$p_{\text{right}}(\gamma) = 1 - p_{\text{left}}(\gamma) = \frac{n(\gamma)}{N}. \quad (12)$$

Combining (11) with the end-cut preference reduction given by (12), we have $\boldsymbol{I}_{\text{mod}}(\hat{\boldsymbol{\beta}},\mathbf{0}) = 4\,\Sigma_{\delta_i=1}\boldsymbol{v}_i$, four times the partial likelihood information matrix from the proportional hazards model. Thus, we have the modified score statistic

$$S_{\text{mod}}(\gamma) = \boldsymbol{U}(\hat{\boldsymbol{\beta}},\mathbf{0})'\left[\boldsymbol{I}_{\text{mod}}(\hat{\boldsymbol{\beta}},\mathbf{0})\right]^{-1}\boldsymbol{U}(\hat{\boldsymbol{\beta}},\mathbf{0}). \quad (13)$$

Notice that, unlike $\boldsymbol{I}(\hat{\boldsymbol{\beta}},\mathbf{0})$, $\boldsymbol{I}_{\text{mod}}(\hat{\boldsymbol{\beta}},0)$ no longer depends on γ. Therefore, in the search for the optimal split point γ_1, this matrix only needs to be inverted once to obtain the modified score statistic $S_{\text{mod}}(\gamma)$ for all γ. In Section 5, simulation results show that the modified score statistic $S_{\text{mod}}(\gamma)$ performs better than the ordinary score test $S(\gamma)$.

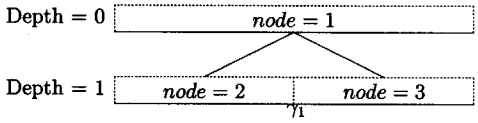
3. Recursive Segmentation and Binary Trees

As described above, the optimal change point γ_1 is determined by maximizing the split statistic. This requires only that the proportional hazards model be fitted to the data. Once the change point has been obtained, it is possible to repeat the above procedure within each interval, resulting in a tree, where each interval represents a node of the tree. The algorithm for growing the binary tree is as follows:

1. Start with the root node, i.e., the entire data and fit the model

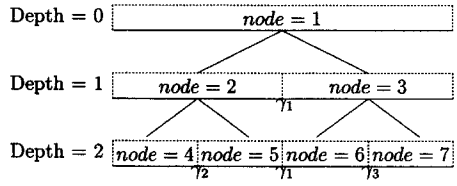
   $$\mathbf{M}_0 : \boldsymbol{\beta}(t) = \boldsymbol{\beta} \quad \text{for all } t.$$

   Determine the point, γ_1, at which the score test is maximized. Define the children of the node as the segments resulting from adding the change point to the model and fit model M_1.

   $$\mathbf{M_1} : \boldsymbol{\beta}(t) = \begin{cases} \boldsymbol{\beta}_1, & \text{for } t < \gamma_1 \\ \boldsymbol{\beta}_2, & \text{for } t \geq \gamma_1 \end{cases}$$

2. Determine the optimal change points, γ_2, by maximizing the score test for node 2, and γ_3, for node 3. Define the children of the node as the segments resulting from adding the change points to the model and fit model M_2.

   $$\mathbf{M_2} : \boldsymbol{\beta}(t) = \begin{cases} \boldsymbol{\beta}_1, & \text{for } 0 \leq t < \gamma_1 \\ \boldsymbol{\beta}_2, & \text{for } \gamma_1 \leq t < \gamma_2 \\ \boldsymbol{\beta}_3, & \text{for } \gamma_2 \leq t < \gamma_3 \\ \boldsymbol{\beta}_4, & \text{for } t \geq \gamma_3 \end{cases}$$

3. The recursive segmentation procedure can be continued. In general, determine the optimal change points at depth d by fitting model M_d. Then obtain model M_{d+1} by determining the optimal change points within each node at depth d.

Remark 3.1. Fitting model M_d at depth d requires fitting Cox proportional hazards models to each interval as follows: on an interval $[\gamma_k, \gamma_{k+1})$, simply exclude all the observations that have occurred before γ_k and treat all those beyond γ_{k+1} as censored.

Remark 3.2. The optimal change point within an interval $[\gamma_k, \gamma_{k+1})$ is obtained from the Schoenfeld residuals and the

information matrix of the Cox proportional hazard model fitted to that interval.

Remark 3.3. Remarks 3.1 and 3.2 result in a fast algorithm for growing the binary tree. Other choices of splitting criterion could also have been used, such as maximized likelihood (Anderson and Senthilselvan, 1982) or residual sum of squares (O'Quigley and Flandre, 1994; Xu, 1996; O'Quigley and Xu, 2001) in analogy to CART. These, however, would require much more intensive computation since there is no fast algorithm for updating when an observation is included or deleted from a candidate interval, and it requires refitting the model at every data point within an interval. Therefore, the criterion of the score statistic is particularly useful.

Remark 3.4. The stopping rule for the splitting of the tree can be determined via specifying (1) a maximum allowable depth, (2) a minimum number of events within each node, or (3) a threshold for the maximized score statistic for each node. We may declare a node terminal if any of the above three occurs. In particular, we recommend that a minimum number of events always be specified since it relates directly to the estimation of the parameters.

4. Pruning and Selection of a Pruned Subtree

4.1 *Pruning*

As in CART, the growing algorithm results in oversegmentation of the data. A pruning algorithm from CART can be used to merge the intervals that do not provide a sufficient increase in the log partial likelihood. Here we use $-\log$(partial likelihood) as the cost. Define a tree T to be a subtree of T_0 if T has the same root as T_0 and every node of T is a node of T_0; denote $T \preceq T_0$. Define the optimally pruned subtree with respect to a penalty parameter α,

$$T^*(\alpha) = \arg\min_{T \preceq T_0} \{C(T) + \alpha|T|\}, \tag{14}$$

where $C(T)$ is the cost of a subtree T and $|T|$ denotes the number of terminal nodes of T. Since we are splitting the time axis simultaneously for all the covariates, the estimates of the parameters and the value of the partial likelihood within one node are unaffected by whether any of the other nodes are combined or not. Hence, the following optimality result can be obtained from the properties of CART (Breiman et al., 1984, pp. 284–293): There exists a finite sequence $\alpha_1 = 0 \leq \alpha_2 \leq \alpha_3 \leq \cdots \leq \alpha_K$ such that, for $k = 1, 2, \ldots, K-1$, $T^*(\alpha) = T^*(\alpha_k) = T_k$ for $\alpha_k \leq \alpha < \alpha_{k+1}$ and $T^*(\alpha) = T_K =$ root node for $\alpha \geq \alpha_K$, where $T^*(\alpha)$ is the optimally pruned subtree determined by the pruning algorithm.

Remark 4.1. It is possible to use the maximized score statistics for pruning by defining a goodness of split for each node as the maximized score statistic of the node. A pruning algorithm is then used to select the subtree with the maximized goodness of split. The required pruning algorithm is described in LeBlanc and Crowley (1993). In simulation studies and in the lymphoma data, it was seen that the cost minimization pruning gives very similar results to pruning by maximizing the goodness of split. This is perhaps not surprising due to the asymptotic equivalence of the score test and the likelihood ratio test.

4.2 *Selection of a Subtree*

Once the sequence $\{T_k\}_{k=1}^K$ of optimally pruned subtrees is obtained, we are left with the problem of selecting a subtree. Because the splitting is adaptively based on the data, the cost $C(T_k)$ is smaller than it would be with previously chosen change points. Following CART, one possible approach to estimating this bias is via learning sample–test sample; i.e., we divide the data into a learning sample and a test sample, use the learning sample to grow the tree and to determine the sequence of optimally pruned subtrees, and then we use the test sample to evaluate the performance and estimate the bias. This approach, however, usually requires a large validation sample and, for our particular application, the size of the tree would depend on the size of the learning sample via the second stopping rule described in Section 3. Another frequently used bias correction method, cross-validation, encounters the same problem in this case. So in the following, we describe a bootstrap procedure that is similar to LeBlanc and Crowley (1993). As discussed in their paper, although it is known that bootstrap estimates themselves can be biased for trees, this may not be as much of a problem here because, as will be seen below, only the partition is obtained from the bootstrap sample.

Let $\mathcal{L}$ denote the data that have been used to grow the full tree and obtain the sequence of pruned subtree T_k's as in Sections 3 and 4.1. Now we draw B bootstrap samples from $\mathcal{L}$ by resampling the triple $(X_i, \delta_i, \boldsymbol{Z}_i)$. Define $\alpha'_k = (\alpha_k \alpha_{k+1})^{1/2}$ for $k = 1, \ldots, K-1$ and $\alpha'_K = \alpha_K + 1$. The reason for taking the geometric means is that α_k was chosen to minimized the cost of the tree in the kth pruning step for the original sample and would give an underestimate of the cost of the tree in a test sample, while using α_{k+1} in the kth step would lead to an overestimate. The geometric means were used in CART (Breiman et al., 1984, Sections 3.4, 11.5) for both learning-test samples and for cross-validation. Note that T_k is optimal for α'_k. For a bootstrap sample $\mathcal{L}_b^*$ $(1 \leq b \leq B)$, we grow a tree and find the optimally pruned subtree $T_{b,k}^*$ for each α'_k, then calculate

$$\omega_{b,k} = C(T_{b,k}^*) - C(\mathcal{L} \downarrow T_{b,k}^*), \tag{15}$$

where $C(\mathcal{L} \downarrow T_{b,k}^*)$ is minus the log partial likelihood of the original data $\mathcal{L}$ fitted to the model given by $T_{b,k}^*$. Then $\omega_{b,k}$ is an estimate of the bias in the cost caused by the adaptive splitting, and we obtain the bias-corrected cost

$$\hat{C}(T_k) = C(T_k) - \frac{1}{B}\sum_{b=1}^{B} \hat{\omega}_{b,k}. \tag{16}$$

In order to choose a pruned subtree, one picks the tree that minimizes $\hat{C}^{\alpha_c}(T_k) = \hat{C}(T_k) + \alpha_c|T_k|$, where α_c is a predetermined penalty parameter. The penalty here is needed since the log partial likelihood tends to be larger over a higher dimensional parameter space, even after the bootstrap bias correction. This is observed both in our simulation studies and in the example of Section 6. The choice of α_c may be based on certain information criteria such as the Akaike information criteria (AIC) (Akaike, 1974), which gives $\alpha_c = p$, or the Schwarz Bayesian information criterion (BIC) (Schwarz, 1978; Volinsky and Raftery, 2000), which gives $\alpha_c = 0.5p\log(N)$, where N is the number of events. LeBlanc and Crowley (1993) also proposed using the 0.95 quantile of the chi-square distribution.

5. Simulation Experiments

Using a step function to approximate $\beta(t)$, it has been shown (Murphy and Sen, 1991) that, with increasing sample size, if the interval lengths go to zero uniformly at a certain rate, the maximum partial likelihood estimate is consistent and asymptotically normal. This is similar to the CART case where the node size is also required to go to zero (Breiman et al., 1984). However, such consistency may be somewhat irrelevant in practical applications since one of the attractive features of the tree-based approach is to provide simple, interpretable estimates. Therefore, we emphasize a sparse segmentation of the time axis and the data adaptive selection of split points in our analysis of the non-Hodgkin's lymphoma data. In the following simulation, we evaluate the performance of the tree-based procedure and show that it provides consistent results. The simulations show the applicability of this method not only to the non-Hodgkin's Lymphoma data but to other case studies as well.

Single covariate. Data were first simulated with a single covariate distributed as uniform$(0, 1)$. We simulated survival data from model (1) using three different cases of $\beta(t)$:

Case 1: $\beta(t) \equiv 1.5$ (no change).
Case 2: $\beta(t) = 2$ for $t \leq 0.2$ and $\beta(t) = 0$ otherwise (strong change).
Case 3: $\beta(t) = 2$ for $t \leq 0.2$ and $\beta(t) = 1$ otherwise (mild change).

The change point at 0.2 corresponds to roughly the 0.43 quantile of the marginal survival distribution in cases 2 and 3. The baseline hazard $\lambda_0(t) = 1$ and the censoring distribution was uniform$(0, \tau)$, where τ was chosen to achieve 25% censoring in each case. For all three cases, the maximum allowable depth in the tree algorithm was four, with a minimum number of 15 events in each interval. We considered the two split statistics $S(\cdot)$ from (4) and $S_{\text{mod}}(\cdot)$ from (13) for sample sizes 100 and 500, with 100 simulations in each case. Penalty parameters $\alpha_c = 2$ and 3 were used for sample sizes 100 and 500, respectively. These are approximately the BIC choice of α_c when there is 25% censoring (Volinsky and Raftery, 2000). Table 1 shows the distributions of the number of terminal nodes, and the distributions of the first split points are plotted in Figure 1, using *density*$(\cdot)$ of Splus. The distribution of the first split points reflects the behavior of the splitting statistics.

- In all three cases, the split statistics $S(\cdot)$ and $S_{\text{mod}}(\cdot)$ both showed improvement in performance when sample size was increased from $n = 100$ to $n = 500$.
- In all three cases, $S_{\text{mod}}(\cdot)$ proved to have higher accuracy than $S(\cdot)$ in predicting the correct number of segments. The only exception to this was with no underlying change points (case 1) and $n = 100$, where $S(\cdot)$ gave more trees with one terminal node as compared with $S_{\text{mod}}(\cdot)$. This was expected, as $S(\cdot)$ generally tends not to split because of the end-cut preference together with the minimal node-size requirement. However, when the sample size was increased to 500, $S_{\text{mod}}(\cdot)$ again proved to be better than $S(\cdot)$.
- With no underlying change points (case 1), the tree method tends to give the correct answer of one terminal node, reaching a high accuracy rate of 96% with $S_{\text{mod}}(\cdot)$ and $n = 500$. The distributions of the split points for case 1 also appear uniformly flat over the time axis in Figure 1. The end-cut preference of $S(\cdot)$ appears near zero (note that, because of the underlying exponential failure time distribution, there is relatively little information in the right tail).
- In case 2, where there was a substantial change at $t = 0.2$ and when $n = 500$, both split statistics were able to detect the change and mostly gave two terminal nodes, although $S_{\text{mod}}(\cdot)$ achieved a higher 96% than the 83% of $S(\cdot)$. For the smaller sample size of 100, $S_{\text{mod}}(\cdot)$ gave two terminal nodes about half of the time and $S(\cdot)$ about a third of the time. In Figure 1, the distribution of the split points using $S_{\text{mod}}(\cdot)$ also peaks more around $t = 0.2$ than using $S(\cdot)$.
- Finally, in case 3, although there was an underlying change point, the change in $\beta(t)$ was rather mild. O'Quigley and Xu (2000) discussed measuring the variation in $\beta(t)$ over time, the variance of $\beta(T)$ being a natural candidate. It is straightforward to verify that var$\{\beta(T)\}$ in case 2 is four times that of case 3. From our simulation results, for the smaller sample size $n = 100$, the procedure tends to give only one terminal node, and even with the larger sample size $n = 500$, 60–70% of the time we have one terminal node as well. We consider this a proper behavior of the tree-based approach since we do not want it to be overly sensitive to mild changes in $\beta(t)$, especial-

Table 1
Distributions of the number of terminal nodes from simulation[a]

	Number of terminal nodes	Case 1		Case 2		Case 3	
		$n = 100$ (%)	$n = 500$ (%)	$n = 100$ (%)	$n = 500$ (%)	$n = 100$ (%)	$n = 500$ (%)
$S(\cdot)$	1	**89**	**91**	66	10	88	62
	2	8	3	**33**	**83**	**12**	**29**
	3	3	2	1	6	—	6
	4	—	4	—	—	—	3
	5	—	—	—	1	—	—
$S_{\text{mod}}(\cdot)$	1	**81**	**96**	51	2	86	67
	2	10	—	**48**	**96**	**14**	**30**
	3	6	—	1	—	—	2
	4	3	2	—	1	—	1
	5	—	2	—	1	—	—

[a] Entries in bold correspond to rows with 'true' number of terminal nodes that generated the data.

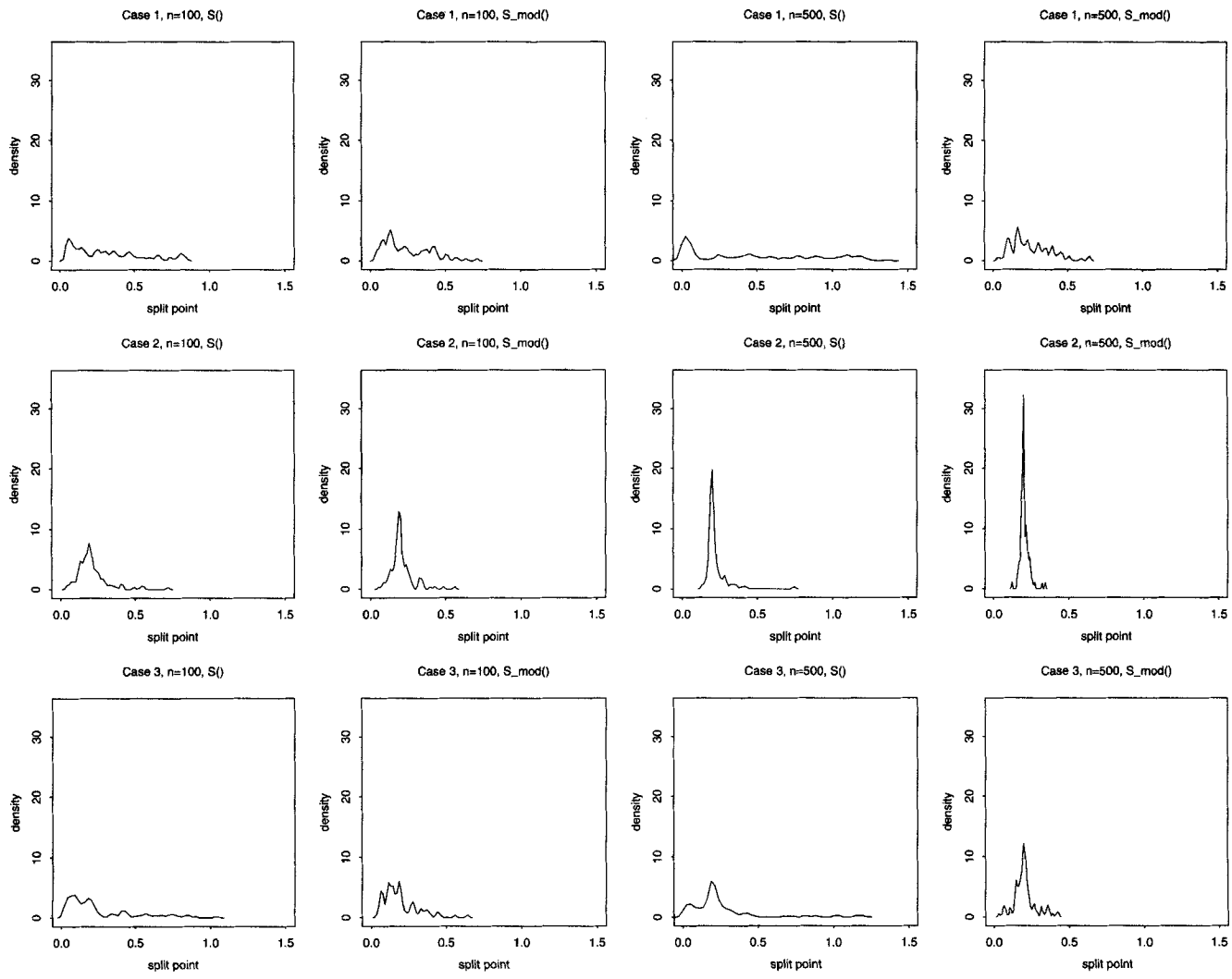


Figure 1. Distributions of the first split points from simulation.

ly if the sample size is small. The distribution of the first split points, as seen from Figure 1, is not quite concentrated for $n = 100$ but is reasonably well centered at $t = 0.2$ for $S_{\text{mod}}(\cdot)$ and $n = 500$. The splits using $S(\cdot)$ have rather wide spreads.

From the above simulation, we see that the number of terminal nodes depends on the sample size as well as the variation in $\beta(t)$ over time. Overall, $S(\cdot)$ does not appear to perform as well as $S_{\text{mod}}(\cdot)$. So, in the following, we will only use $S_{\text{mod}}(\cdot)$ for simulation with two covariates and for the lymphoma data.

Multiple covariates. To study the effect of simultaneous segmentation as described in Sections 2 and 3, when there are multiple covariates and when the regression effects for different covariates might have different underlying change points, we carried out the following experiments. The data were simulated with two independent covariates Z_1 and Z_2, both distributed as uniform$(0, 1)$, with

$$\beta_1(t) = 2 \quad \text{when } t \le 0.05, \qquad 0 \quad \text{when } t > 0.05;$$

$$\beta_2(t) = 2 \quad \text{when } t \le 0.35, \qquad 0 \quad \text{when } t > 0.35.$$

The change points were chosen so that they roughly correspond to 1/3 and 2/3 quantiles of the marginal failure distribution. The censoring distribution was again uniform$(0, \tau)$, with τ chosen to give 25% censoring. We used penalty parameter $\alpha_c = 0$, 2, 4, and 6. Here $\alpha_c = 0$ imposes no penalty for the number of terminal nodes, $\alpha_c = 2$ corresponds to AIC, and $\alpha_c = 6$ is approximately BIC as well as the 0.95 quantile of χ^2_2. The sample size was 500, with 200 simulations.

From the simulation, we found that, with $\alpha_c = 0$, the averages of the estimates of $\beta_1(t)$ and $\beta_2(t)$ looked reasonably good (plots not shown). With increasing penalty, the estimates of $\beta_2(t)$ became less satisfactory. We further examined the results under $\alpha_c = 6$: 83% of the simulations gave two terminal nodes and 13% gave three terminal nodes. It turns out that, for the 25 runs with three terminal nodes, both estimates $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ are fairly good on average. But for the 166 runs with two terminal nodes, the split points are concentrated around $t = 0.05$ and the average of $\hat{\beta}_2(t)$ is about 1.5 for

$t > 0.05$, while the estimates of $\beta_2(t)$ on $(0, 0.05)$ and of $\beta_1(t)$ are quite accurate. An explanation for this is that, if a single change point is to be imposed on both $\beta_1(t)$ and $\beta_2(t)$, since there is more information at the earlier change point (subjects fail or are censored over time), the score statistic tends to pick this earlier change. It then follows (Xu and O'Quigley, 2000) that the estimate of $\beta_2(t)$ on $(0.05, \infty)$ should lie between the true values two and zero, as is the case.

The above simulation suggests that, when there are potentially different underlying change points in the covariates, we should consider using a relatively small penalty α_C. This would allow the discovery of all the underlying change points. In fact, this is mostly achieved in the growing part of the tree procedure since we would always have over-segmentation. Following that, we can recombine different nodes for different covariates as the final step (see also Discussion). We have encountered such examples in our practical data experience, although they are not shown here.

6. Non-Hodgkin's Lymphoma Prognostic Factors Data

For the non-Hodgkin's lymphoma prognostic factors data described in Section 1, there were 1968 patients with complete data on the five binary covariates and 593 events were observed in these patients. Assuming model (1), we use the tree-based approach to estimate time-varying regression effects for these covariates.

The recursive segmentation algorithm of Section 3 (with a minimum node size of 20 events as the stopping rule) led to a full tree, which terminated at depth four with eight terminal nodes. The full tree and the segmentation of the time axis as created by the terminal nodes are shown in Figure 2. We then used the pruning algorithm of Section 4.1 to obtain a sequence of eight subtrees, which are listed in the first five columns of Table 2. Subtree 1 is the full tree with eight terminal nodes and subtree 8 is the root with a single terminal node (i.e., the whole time interval). The fifth column of the table (labeled prune-off) indicates how a subtree is obtained by pruning off the branches in order, starting from the full tree. For example, subtree 2 is obtained by pruning off the branches below node 9, subtree 3 is from further pruning off the branches below node 11, and so on. In Figure 2, the height of a branch is drawn proportional to the magnitude of change in the log partial likelihood corresponding to that split. Note that these heights match the order in which the subtrees are obtained: The first prune-off below node 9 corresponds to the shortest branches of the tree, the second shortest branches are below node 11, which are the second to be pruned off, etc.

One hundred bootstrap samples were used to correct for the overoptimism in the costs. The bootstrap-corrected cost $\hat{C}$ for each subtree is in the last column of Table 2. With no penalty, the subtree with the minimum cost is subtree 2, with seven terminal nodes, which is perhaps too noisy. If we set the penalty parameter according to BIC, $\alpha_C = 0.5 \times 5 \times \log(593) = 15.96$ and the resulting optimal tree is subtree 7, with two terminal nodes and a single change point at 1.07 years. This turns out to be the same subtree as using the 0.95 quantile of chi-square with 5 d.f., i.e., $\alpha_C = 11$. The result also appears consistent with Figure 2, where this split has a significantly longer branch than all the other splits. The final estimates of $\beta(t)$ for each covariate on the two intervals are shown in Table 3 as well as in Figure 3. Note that 9.02 years is the maximum observation time in this data set. In Figure 3, we also plotted the piecewise constants splines (Gray, 1992), the loess estimates (Grambsch and Therneau, 1994), and the proportional hazards estimate of constant β, i.e., tree 8. It is clear that the tree-based method gives a reasonable summary of the changing covariate effects over time, which is suitable for clinical interpretation purposes. Notice also that the age effect is basically constant according to tree 7, but for the purpose of defining a time-varying risk index below, we will leave the two time periods uncombined for the effect of age.

Based on the conditional inference given in Table 3, we see that the regression effects of performance status and the number of extranodal sites as recorded at baseline are no longer significant after about 1 year. This is sensible because, a year after the treatment, if a patient is still alive, most likely his or her performance status and extranodal sites would have had substantial change, so the baseline values have little predictability of future prognosis. Thus, we define a time-varying international risk index based on the number of risk factors as follows:

- At diagnosis, the risk groups are as defined by Shipp et al. (1993)—zero or one risk factor, low risk; two risk factors, low intermediate risk; three risk factors, high intermediate risk; and four or five risk factors, high risk.
- After 1.07 years (or approximately 1 year), if a patient is still alive, his or her risk category will be based only on age, stage, and serum LDH—zero or one risk factor,

Table 2
Pruning and bootstrap results for lymphoma data

Tree number (k)	Number of terminal nodes ($\lvert T_k \rvert$)	Penalty parameter (α_k)	$-\log$(Partial likelihood) ($C(T_k)$)	Prune-off	Bootstrap estimate ($\hat{C}(T_k)$)
1	8	0.00	6001.51	—	6107.38
2	7	3.62	6005.14	9	6028.77
3	6	4.18	6009.31	11	6032.90
4	5	5.63	6014.94	6	6038.01
5	4	5.99	6020.93	5	6037.84
6	3	12.93	6033.86	2	6039.93
7	2	15.21	6049.07	3	6049.46
8	1	34.65	6083.72	1	6079.86

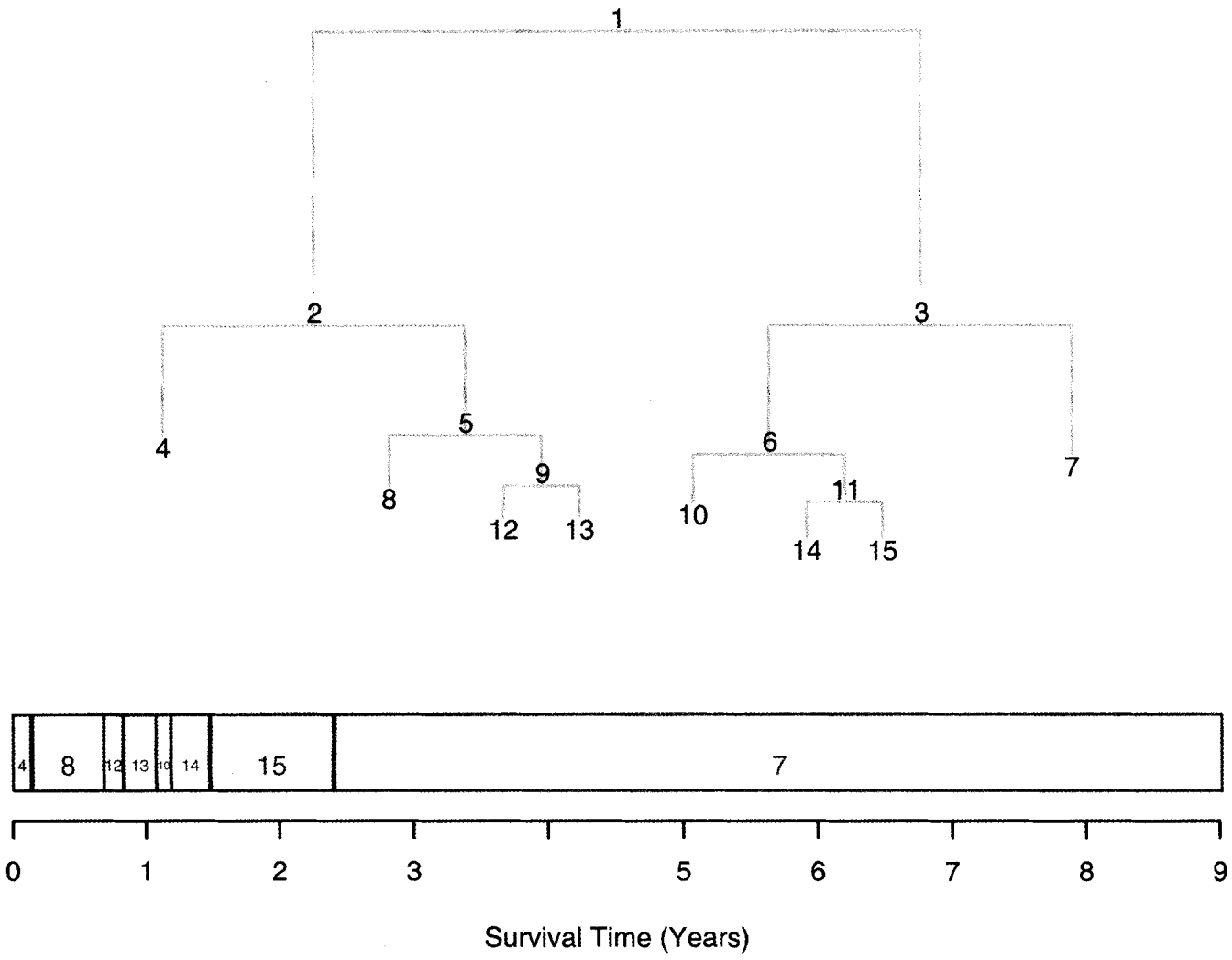


Figure 2. The full tree from the lymphoma data.

low risk; two risk factors, low intermediate risk; and three risk factors, high intermediate risk. Thus, all patients in the high-risk group surviving beyond 1.07 years would be downgraded to a lower risk category.

The time-varying risk index for the patients enrolled in the International Non-Hodgkin's Lymphoma Prognostic Factors Project is shown in Table 4. Of the 1469 patients alive beyond 1.07 years, 472 (= 197+42+136+43+54) patients were downgraded to a lower risk category based on the time-varying risk index. Of the 235 patients who were in the high-risk category at diagnosis, approximately 18% were downgraded to the low intermediate risk category and another 23% to the high intermediate category, having survived beyond 1.07 years. Such

Table 3
Final estimate of $\beta(t)$ (two terminal nodes)

	$t \in [0, 1.07)$				$t \in [1.07, 9.02]$			
Covariate	Coef.	exp(Coef.)	SE(Coef.)	p-Value	Coef.	exp(Coef.)	SE(Coef.)	p-Value
Age	0.522	1.69	0.092	<0.01	0.535	1.71	0.103	<0.01
Stage	0.352	1.42	0.120	<0.01	0.555	1.74	0.118	<0.01
Serum LDH	0.725	2.07	0.095	<0.01	0.383	1.47	0.115	<0.01
Performance status	0.917	2.50	0.095	<0.01	−0.056	0.95	0.135	0.68
Extranodal site	0.501	1.65	0.096	<0.01	−0.048	0.95	0.119	0.69

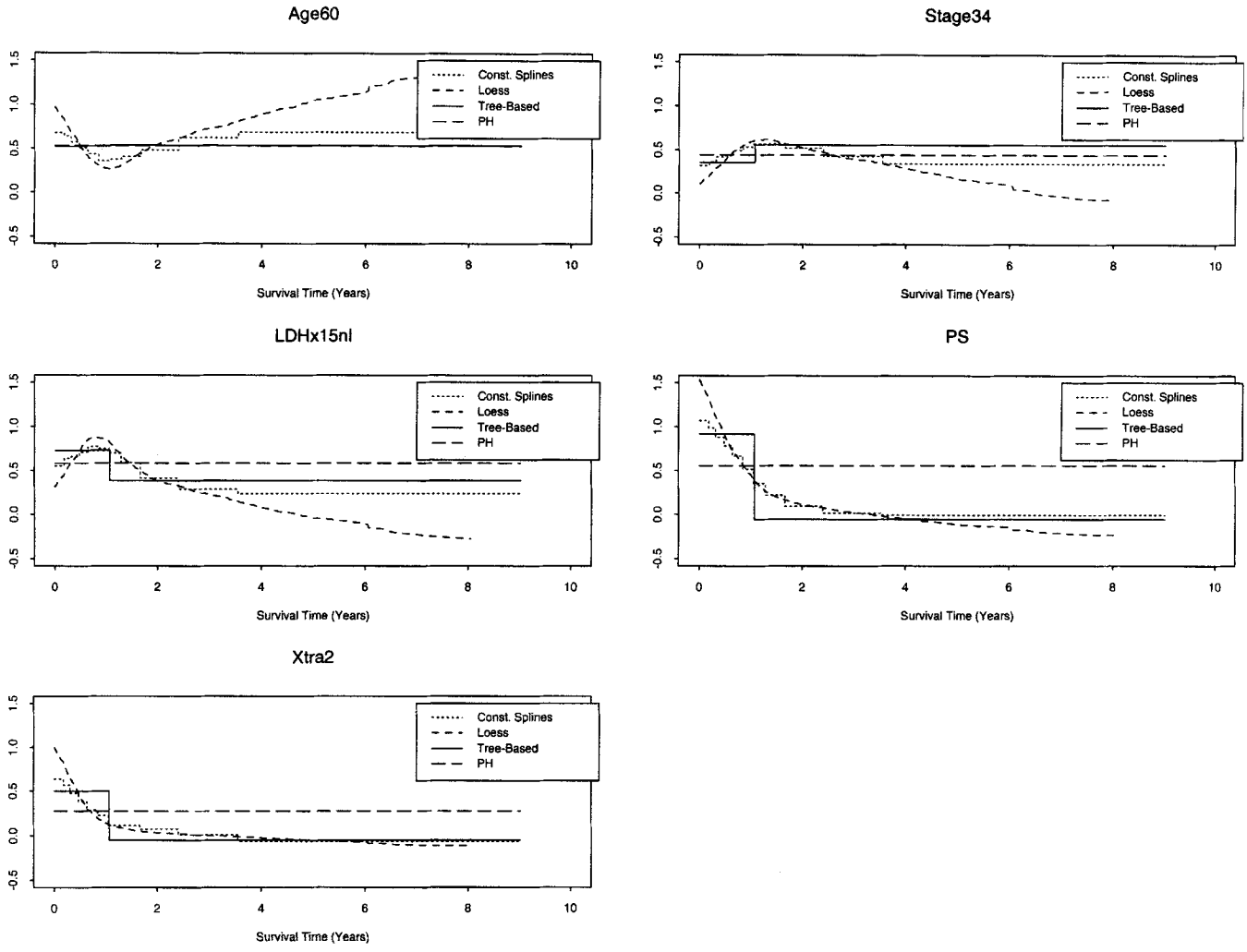


Figure 3. Estimates of $\beta(t)$ from the lymphoma data.

Table 4
Time-varying international risk index for lymphoma patients (LI = low intermediate; HI = high intermediate)

	Risk index at diagnosis (Shipp et al., 1993)				
Status at 1.07 year	Low	LI	HI	High	Total
Died/Lost to Follow-Up Prior to 1.07 Years					
Died	88 (10.4%)	115 (21.6%)	140 (38.9%)	135 (57.4%)	478 (24.3%)
Lost to follow-up	9 (1.1%)	6 (1.1%)	3 (0.8%)	3 (1.3%)	21 (1.1%)
Alive with Follow-Up Past 1.07 Years					
Revised risk index					
Low	745 (88.5%)	197 (37.1%)	42 (11.7%)	0	984 (50%)
LI	0	213 (40.1%)	136 (37.8%)	43 (18.3%)	392 (19.9%)
HI	0	0	39 (10.8%)	54 (23.0%)	93 (4.7%)
Total	842 (100%)	531 (100%)	360 (100%)	235 (100%)	1968 (100%)

downgrading is potentially useful in clinical decisions such as protocol design and target patient population.

Although not shown here, we also used the pruning algorithm with the goodness of split based on the score statistic as discussed in Remark 4.1. The resulting set of nested subtrees was identical to that of Table 2.

7. Discussion

The tree-based approach described here provides an exploratory tool for finding the change points in estimating $\beta(t)$ under model (1). The use of recursive partitioning along the time axis is similar to Adak (1998). As pointed out by one reviewer, tree-based techniques are often very useful when partitioning in a multidimensional covariate space, while our work concerns segmentation along the single time axis. We note that recursive splitting is computationally less expensive than many other data adaptive methods. In fact, as far as we are aware, there has not been a feasible approach developed in the literature for finding multiple change points of $\beta(t)$. In the general change-point literature, Csörgő and Horváth (1997) discussed tests for multiple changes and, interestingly, a binary segmentation method was mentioned. The tree-based approach considered in this article may be viewed as providing a defined algorithm to search for multiple change points.

Under model (1), it is possible to allow for time-dependent covariates because the implementation of the tree algorithm essentially fits proportional hazards models. However, the interpretation of such a model needs care, as time-dependent covariates are sometimes used to count for time-varying regression effects. One practical situation where such models might be applicable would be where the time-dependent covariate represents actual measurements taken over time.

Following the estimation of $\beta(t)$, it is straightforward to estimate the conditional survival probabilities. Although the piecewise constant $\hat{\beta}(t)$ may not estimate $\beta(t)$ precisely at every t, it was observed in Xu and O'Quigley (2000) that the estimate of conditional survival may not be very sensitive to the regression coefficient estimates. This of course is an area that deserves further investigation of its own.

Finally, in practice, for model (1), we often have multiple covariates, and the regression effects of all the covariates may not vary in a synchronized fashion over time (notice though that some of the covariates could be correlated, reflecting certain underlying biological features). This generally should not cause problems in the growing part of the tree algorithm because we usually end up with oversegmentation of the data. Following the pruning, it is possible to choose different final splits for different covariates. This is a rather flexible and perhaps subjective step and, as for model selection in general, a certain degree of subjective judgment is inevitable. For some applications, we may use the union of all the split points for different covariates, i.e., the smallest subtree that includes sufficient splits for all covariates; this would give a time-varying prognostic index, which is often used in clinical research. Another possibility exists in applications where it might be desirable to restrict some components of $\beta(t)$ to be monotone over time. Such requirements may also be incorporated in deciding the final splits for the covariates.

Acknowledgements

The authors would like to thank Dr David P. Harrington for providing the International Non-Hodgkin's Lymphoma Prognostic Factors Project data. We would also like to thank the reviewers and the editor for helpful suggestions.

Résumé

Le problème des risques non proportionnels au cours du temps apparaît souvent dans les analyses de survie, comme c'est le cas dans l'étude internationale des facteurs pronostiques dans le lymphome non hodgkinien. Une méthode fondée sur les arbres de régression pour travailler sur de telles données est développée pour mettre en évidence et estimer l'effet dépendant du temps sous un modèle de Cox. Cette méthode utilise une approximation de l'effet qui dépend du temps par un effet constant par intervalles et permet d'estimer les points de changements. Un algorithme rapide sur la maximisation du test du score est utilisé dans la segmentation récursive de l'axe du temps. Après la segmentation, un algorithme d'élagage avec les mêmes propriétés optimales que la méthode CART est utilisé pour déterminer une segmentation moins dense. La technique du bootstrap est utilisée pour corriger le surajustement lié à l'optimisation du point de coupure. Le modèle constant par intervalle est plus facile à interpréter cliniquement que le modèle utilisant la méthode des splines. L'utilisation de cet algorithme est illustré sur les données du lymphome. Toujours sur ces données, nous proposons de transformer l'Index International de risque en un index dépendant du temps.

References

Adak, S. (1998). Time-dependent spectral analysis of nonstationary time series. *Journal of the American Statistical Association* **93,** 1488–1501.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions in Automatic Control* **19,** 716–723.

Anderson, J. A. and Senthilselvan, A. (1982). A two-step regression model for hazard functions. *Applied Statistics* **31,** 44–51.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees.* Belmont, CA: Wadsworth International Group.

Bretagnolle, J. and Huber-Carol, C. (1988). Effects of omitting covariates in Cox's model for survival data. *Scandinavian Journal of Statistics* **15,** 125–138.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62,** 269–276.

Csörgo, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis.* Chichester, U.K.: Wiley.

Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64,** 247–254.

Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74,** 33–43.

Ford, I., Norrie, J., and Ahmadi, S. (1995). Model inconsistency, illustrated by the Cox proportional hazards model. *Statistics in Medicine* **14,** 735–746.

Gail, M. H., Wieand, S., and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments

with nonlinear regressions and omitted covariates. *Biometrika* **71,** 431–444.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81,** 515–526.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association* **87,** 942–951.

Gustafson, P. (1998). Flexible Bayesian modelling for survival data. *Lifetime Data Analysis* **4,** 281–299.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B* **55,** 757–796.

Kim, K. and Tsiatis, A. A. (1990). Study duration for clinical trials with survival response and early stopping rule. *Biometrics* **46,** 81–92.

Lancaster, T. and Nickell, S. (1980). The analysis of re-employment probabilities for the unemployed. *Journal of the Royal Statistical Society, Series A* **143**(2), 141–165.

LeBlanc, M. (2001). Tree-based methods for prognostic stratification. In *Handbook of Statistics in Clinical Oncology,* J. Crowley (ed), 457–472. New York: Marcel Dekker.

LeBlanc, M. and Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association* **88,** 457–467.

Liang, K. Y., Self, S. G., and Liu, X. (1990). The Cox proportional hazards model with change point: An epidemiologic application. *Biometrics* **46,** 783–793.

Miller, R. and Siegmund, D. (1982). Maximally selected chi square statistics. *Biometrics* **38,** 1011–1016.

Morgan, J. and Sonquist, J. (1963). Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association* **58,** 415–434.

Murphy, S. A. and Sen, P. K. (1991). Time-dependent coefficients in a Cox-type regression model. *Stochastic Processes and Their Applications* **39,** 153–180.

O'Quigley, J. and Flandre, P. (1994). Predictive capability of proportional hazards regression. *Proceedings of the National Academy of Science, USA* **91,** 2310–2314.

O'Quigley, J. and Pessione, F. (1991). The problem of a covariate-time qualitative interaction in a survival study. *Biometrics* **47,** 101–115.

O'Quigley, J. and Xu, R. (2000). Inference for the Cox model under proportional and nonproportional hazards. In *New Approaches in Applied Statistics,* A. Ferligoj and A. Mrvar (eds), 3–19. Ljubljana: FDV.

O'Quigley, J. and Xu, R. (2001). Explained variation in proportional hazards regression. In *Handbook of Statistics in Clinical Oncology,* J. Crowley (ed), 397–409. New York: Marcel Dekker.

Pettitt, A. N. and Bin Daud, I. (1990). Investigating time dependence in Cox's proportional hazards model. *Applied Statistics* **39,** 313–329.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications,* 2nd edition. New York: Wiley.

Sargent, D. J. (1997). A flexible approach to time-varying coefficients in the Cox regression setting. *Lifetime Data Analysis* **3,** 13–25.

Schoenfeld, D. A. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69,** 239–241.

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

Shipp, M. A., Harrington, D. P., and Anderson, J. R. (1993). A predictive model for aggressive non-Hodgkin's lymphoma. *New England Journal of Medicine* **329,** 987–994.

Sleeper, L. A. and Harrington, D. P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association* **85,** 941–949.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73,** 363–369.

Verweij, J. A. and Van Houwelingen, H. A. (1995). Time-dependent effects of fixed covariates in Cox regression. *Biometrics* **51,** 1550–1556.

Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics* **56,** 256–262.

Xu, R. (1996). Inference for the proportional hazards model. Ph.D. thesis, University of California, San Diego.

Xu, R. and O'Quigley, J. (2000). Estimating average regression effect under nonproportional hazards. *Biostatistics* **1,** 423–439.

Zhang, H. and Singer, B. (1999). *Recursive Partitioning in the Health Sciences.* New York: Springer.

Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *Annals of Statistics* **18,** 329–353.

Received June 2001. *Revised November* 2001.
Accepted November 2001.

Tree-based Ensembles

While trees enjoy a number of desirable properties, their performance in terms of accuracy can be improved by *tree-based ensembles*: combining a variety of suitably chosen trees.

- Tree-based ensembles combine the predictions of many different trees to give an aggregated prediction.

- To obtain the different trees, some ensemble methods use randomness in the tree-fitting procedure;

- others fit non-random trees to different versions of the dataset;

- and some employ both of these strategies.

- Methods also differ in how the predictions are aggregated.

- In regression, a simple aggregates prediction is the average of the predictions from the individual trees.

- A simple version for classification is to use the most frequently predicted class, in a procedure known as "voting" the trees.

Bagged Trees

Bagging stands for "bootstrap aggregating" (Breiman, 1996).

- **Bootstrap**, like CV, is a type of resampling method.
- It resamples from the original i.i.d. data set of size n *with replacement*, to give a bootstrapped data set of size n.
- Some of the original data points will not appear in the bootstrap sample;
- others may appear more than once.
- One would draw bootstrapped samples many times, eg. $B = 100$ or more.
- Bagging grows B trees on the B bootstrapped samples.
- For classification, voting might be used: for any given $\mathbf{x}$, each tree gives a 'vote' for which class $\widehat{Y}(\mathbf{x})$ should belong to, and the class with the most vote wins.

Bagging example

Compare this to the regression surface on page 1 of the previous chapter, reproduced on the next page.

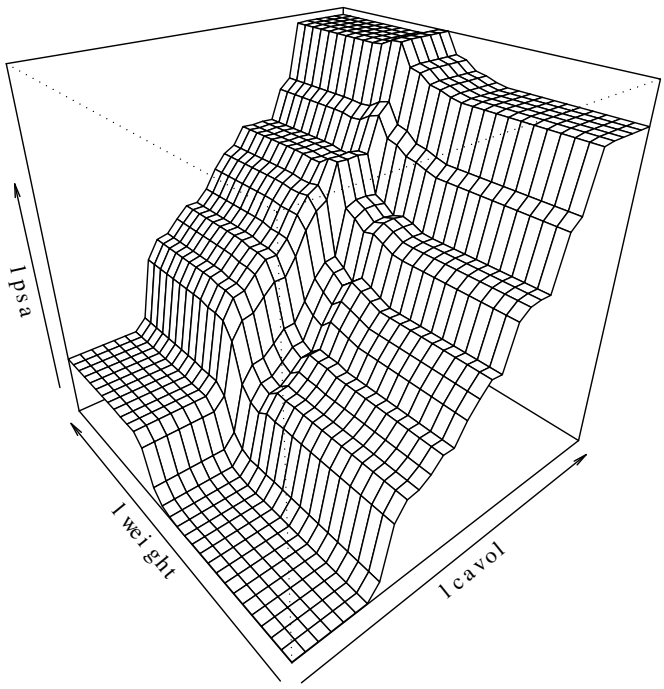
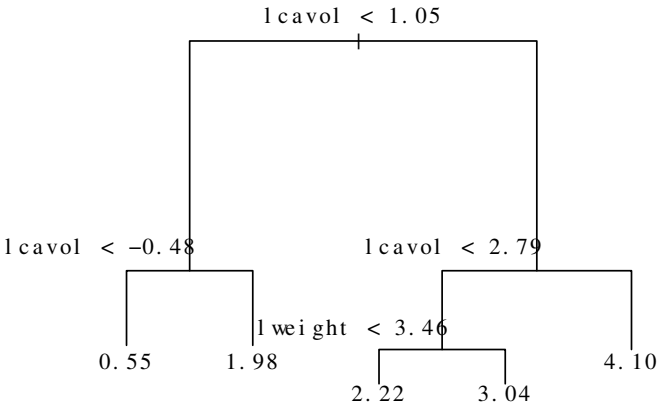
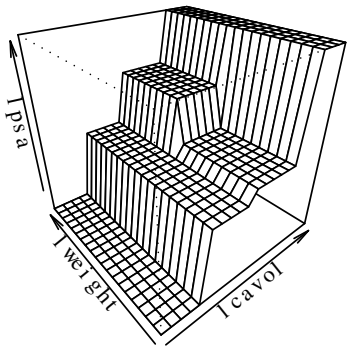
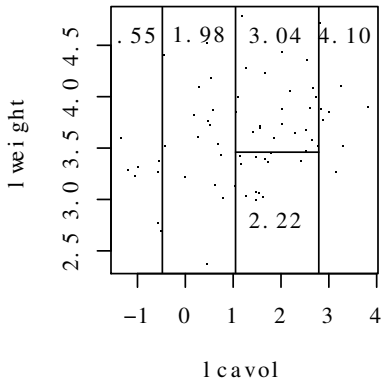


Figure 1: Regression surface for bagging 100 regression trees, 2-dimensional prostate cancer data.

From a single tree:







Properties of Bagged Trees

- Bagged trees retain the positive characteristics of trees listed in the previous chapter.
- Bagged trees seldom do worse than an individual tree.
- But they are computationally slower;
- and more difficult to interpret.
- Bagging tends not to perform as well as the ensemble methods below (Chapter 5 of Li and Xu book).

Random Forests

Random forests (Breiman, 2001) are tree-based ensembles that use *bootstrap* samples and *randomness* in the tree-building procedure.

- Take m to be much smaller than the total number of predictors, p.
- Trees are grown on bootstrap samples, using a random sample of m predictors.
- The m predictors are chosen independently for each node.
- The splitting of the node is done as for a single tree in CART.
- The trees are grown large, and not pruned.
- For classification, the trees are grown until each terminal node contains members of only one class.

Properties of Random Forests

Random forest achieves accurate predictions via *low bias* and *low correlation* among trees.

- Low bias is achieved by growing large trees;

- low correlation results from making the trees as dissimilar as possible.

- In bagging,
  - the trees differ only because they are fit to different bootstrap samples;
  - but every two different **bootstrap samples have a lot of overlap**.
  - This gives trees that are too similar to give low correlation;
  - so they tend to make mistakes in similar places.

- In Random Forests,
  - random sampling of a small number (m) of predictors at each node forces the trees to be quite different;
  - this reduces correlation and improves predictive power.

- Random Forests retain the positive characteristics of trees listed in the previous chapter.

- In addition, Random Forests (RF) have the following features.

  Note: Technically, bagging is a special case of RF, by setting $m = p$ (though in RF $m << p$).

  They can both be computed using R package 'RandomForest', and the quantities below like generalization error and variable importance can be obtained for both.

Built-in generalization error

- When a bootstrap sample is taken from the data, some observations do not make it into the bootstrap sample. These are called "out-of-bag" data.

- Out-of-bag data can be used to give an internal estimate of generalization (i.e. prediction) error.

- For each observation, the generalization error can be estimated by averaging the error rate of the predictions from the trees for which it was out-of-bag, like in CV.

- An overall error rate is computed by averaging over all the observations.

Tuning parameters

There are mainly two tuning parameters for Random Forests: m and the number of trees.

These are like the penalty or complexity parameter in CART.

- Random Forests are known not to be sensitive to the choice of either of these parameters.

- The default choices are 500 trees and $m = \sqrt{p}$, where p is the total number of predictors.

- Breiman (2001) showed that adding more trees to a random forest does not lead to overfitting.

- So the only real concern with the number of trees is that it should be large enough;

- this can be checked using the out-of-bag error rate.

Variable importance

Unlike a single tree, it is not straightforward to see the contribution of a given predictor.

- Random Forests gives a measure of variable importance, by randomly **permute** the values of a variable.

- That is, for n observations, it does a random *permutation* of the values of that variable.

- It then compares the difference between the error rates based on the original data versus the permuted data.

- If RF captures an interaction (trees are good at this), the variables involved are likely to show up as "important".

Boosted trees

Early AdaBoost was developed in Freund and Schapire (1996).

Today commonly used gradient boosting machines (GBM) was developed in Friedman (2001), and implemented in R package 'gbm'.

- Boosting does not use bootstrapped data.
- Instead one can think of it as fitting trees to *weighted* versions of data.
- It does so sequentially, so that it slowly learns.
- The **misclassified observations gets more weights**.
- The trees used in boosting are typically small, sometimes as small as "stumps", which have only one split. Otherwise they have just a few terminal nodes.

Bagging and boosting can both be applied to procedures *that are not trees*.

RF versus GBM

RF and GBM both are somewhat powerful predictors.

They also share many positive characteristics of trees listed before.

Example: Prostate Cancer Microarrays
(From Li and Xu book, chap. 5)

The data set consists of 6033 gene expression values for 102 arrays, namely 50 normal samples and 52 tumor samples.

Of interest was to predict disease class, normal or tumor, based on the gene expression values.

Table 1: CPU time (seconds) for RF and GBM, average of 5 runs on a 2GHz machine

Method	Number of Trees	Number of Predictors 1000	2000	3000	4000	5000	6000
RF	100	1.1	2.3	3.5	4.6	5.7	6.6
GBM	100	4.2	8.3	12.6	16.9	21.1	24.9
RF	500	5.5	11.1	17.0	22.2	28.2	32.2
GBM	500	8.2	16.2	24.2	32.5	40.4	47.8
RF	1000	10.7	22.2	34.1	44.4	56.2	65.2
GBM	1000	13.3	26.1	38.9	52.0	64.8	76.6

At default values for all parameters, the average error rates from 10 runs of 10-fold CV were 9.4% for RF, and 14.2% for GBM.

When the # trees for GBM was increased to 1000 and 1500, its average error rate dropped to 10.2% and 9.6%, respectively.

Table 2: Sensitivity to tuning parameter choice. Table gives number of errors for prostate cancer data, computed by 10-fold CV with the specified choice of tuning parameters and everything else set at default values.

		Number of Trees				
		100	250	500	750	1000
Random	m					
Forests	25	13	13	10	10	12
	50	12	9	8	9	7
	75	9	10	**8**	9	9
	100	7	9	7	7	7
	150	8	8	7	7	7
	200	9	7	7	7	7
	300	7	7	8	7	7
Gradient	*Shrinkage*					
Boosting	.0001	50	43	25	14	12
Machines	.0005	27	13	11	11	10
	.0010	**13**	11	10	9	9
	.0015	14	11	10	10	9
	.0020	11	10	11	9	8
	.0025	9	11	9	7	7
	.005	10	9	7	6	6
	.01	11	7	6	6	6
	.05	7	6	6	6	6
	.1	5	7	6	6	6
	.5	9	7	7	12	15
	1.0	19	9	18	18	20

Default values are shown in bold, with the default value of $m = \sqrt{6033} \approx 75$ for RF.

MATH 189 Final Project

Winter 2020

Description of the Dataset

The dataset "dat.rda" is a selected subdataset from the dataset named NHANES. You can load the dat.Rda to get the data for your final project.

If you would like to see the description of each variables in the dataset, you could take a look at the description of the original NHANES dataset. The following link shows a detailed dexcription of the NHANES package https://cran.r-project.org/web/packages/NHANES/NHANES.pdf. You can also find the description of variables using the help function in R after you install and require the package.

```
install.packages("NHANES")
require(NHANES)
? NHANES
```

The dataset "dat.rda" contains 17 variables and 2783 observations selected from the original NHANES dataset. The observations are from the year 2011 and with the age between 18 and 59. The 17 variables are:

```
##       [,1]
##  [1,] "Gender"
##  [2,] "Age"
##  [3,] "Race1"
##  [4,] "Education"
##  [5,] "MaritalStatus"
##  [6,] "HHIncome"
##  [7,] "BMI_WHO"
##  [8,] "HealthGen"
##  [9,] "Depressed"
## [10,] "SleepTrouble"
## [11,] "PhysActiveDays"
## [12,] "TVHrsDay"
## [13,] "CompHrsDay"
## [14,] "AlcoholYear"
## [15,] "SmokeNow"
## [16,] "RegularMarij"
## [17,] "SexOrientation"
```

If you are interested in how the data dat.Rda is obtained from the original NHANES dataset, see the following codes for the data generating process. For the convenience of your analysis, the categorical variable HealthGen has been grouped into two categories Excellent/Vgood, versus Good/Fair/Poor, and are denoted by 1 and 0 repectively.

```
require(NHANES)
dat <- NHANES
SurveyYr <- NHANES$SurveyYr
summary(SurveyYr)
# select the observations that is in year 2011
dat <- dat[SurveyYr=="2011_12", ]
```

```r
age <- dat$Age
#summary(age)
# select the obervations with age between 18 and 59
dat <- dat[(age >= 18) & (age <= 59), ]

#colnames(dat)
# select the variables
ind <- c("Gender","Age","Race1","Education","MaritalStatus","HHIncome",
         "BMI_WHO","HealthGen","Depressed","SleepTrouble","PhysActiveDays",
         "TVHrsDay","CompHrsDay","AlcoholYear","SmokeNow","RegularMarij",
         "SexOrientation")
dat <- dat[,ind]

# dichotomize into two categories: Excellent/Vgood vs others
dat$HealthGen = as.factor(as.numeric((dat$HealthGen == "Excellent" | dat$HealthGen == "Vgood" )))
```

Final Project

For the final project, look back at all the analysis approaches you have used throughout the quarter. Consider HealthGen as the outcome, grouped into Excellent/Vgood, versus Good/Fair/Poor. (In the dataset dat.Rda, the HealthGen variable is assigned value 1 if its original value is Excellent/Vgood and is assigned value 0 otherwise.) Consider all other variables as potential predictors. Develop a comprehensive and reproducible analysis report, to explore the relationship between these variables and the outcome. Pay attention to (but not limited to) the following:

1) Missing data: do not remove observations with any missing data from the start; after screening you might reduce to a smaller set of variables, therefore remove fewer observations at that point. Also you may consider removing variables with too much missing.

2) Include "Table 1".

3) After univariate screening, building a multiple logistic regression model to predict the general health outcome of very good or excellent versus otherwise. State clearly your criteria at each step in the narrative.

4) Assess the predictability of the model by computing the (generalized) R-squared and the area under the ROC curve (AUC), as well as the cross-validated AUC.

5) Use the variables that have passed the univariate screening, to build a classification tree. Describe clearly how you arrive at the final tree. Compute the error rate of your classification tree.

6) Discuss any limitations in the analysis.

Bonus (2%): Explore random forest on the data above.

The report should be a pdf file generated by R markdown. Be sure to append all your R codes in the back of the report, using comments to mark each section of the codes in terms of what it does.

High Dimensional Data

As in the previous microarray data example, when the number of predictors p, is large compared to the sample size n, we refer to this type of data as high dimensional.

- When $p \geq n$, typical regression models cannot be fit.
- Previously we said there needs to be at least 10 to 15 events or samples per parameter to be estimated.
- We saw that methods like RF or GBM can handle the microarray data.
- If we want to fit regression type of models, methods called regularization can be used for high-d data.

Regularization

- Let β be a p-dimensional vector of regression coefficients, where p is large compared to n.

- We choose β to maximize a penalized log-likelihood:

$$\log L(y|\beta) - P_\lambda(\beta),$$

  where $\lambda \geq 0$ is the penalty parameter.

- Often we can use the penalty $P_\lambda(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|^m$.

1. $m = 0$, L_0 penalty: best subset (AIC), stepwise (might require orthonormal design under linear regression), adjusted R^2, generalized cross-validation (GCV).

2. $m = 1$, L_1 penalty: least absolute shrinkage and selection operator (LASSO).

3. $m = 2$, L_2 penalty: ridge regression.

4. Other penalties: elastic net (combined L_1 and L_2 penalties), smoothly clipped absolute deviation (SCAD) etc.

See Harezlak et al. Chapter in "*High-Dimensional Data Analysis in Cancer Research*", Springer 2009.

LASSO

An outstanding feature of L_1 penalty is that it estimates some coefficients to be exactly zero.

- This way LASSO effectively does variable selection.
- It can handle the $p \geq n$ situation that cannot be handled by AIC, stepwise etc.
- It also shrinks the non-zero coefficients towards zero, so the resulting estimates are biased.
- The penalty parameter λ is selected via methods like cross-validation, for prediction accuracy.

Large Sample Behaviors

Consistency refers to the behavior of a procedure as the sample size $n \to \infty$.
(The concept of limiting behaviors is taught in MATH 140, 142.)

- With proper choice of λ, *consistent prediction* can be achieved by LASSO.

- **Consistent model selection** means the probability of selecting the true model (i.e. the set of true predictors) goes to one as $n \to \infty$.

- LASSO is NOT consistent in model selection (neither are the stepwise procedures).

- *Adaptive LASSO* is consistent in model selection, under *irrepresentability conditions* on the design matrix (i.e. predictors).

- However, irrepresentability conditions are generally considered too strong these days.

- The focus now is on *inference* (i.e. CI's, hypo. testing) when p grows with n.

Received: 11 April 2017 | Revised: 9 April 2018 | Accepted: 26 April 2018

DOI: 10.1002/sim.7822

RESEARCH ARTICLE

WILEY Statistics in Medicine

High-dimensional variable selection and prediction under competing risks with application to SEER-Medicare linked data

Jiayi Hou[1] | Anthony Paravati[2] | Jue Hou[3] | Ronghui Xu[3,4] | James Murphy[2]

[1]Altman Clinical and Translational Research Institute, University of California, San Diego, La Jolla, CA 92093, U.S.A.

[2]Department of Radiation Medicine and Applied Sciences, University of California, San Diego, La Jolla, CA 92093, U.S.A.

[3]Department of Mathematics, University of California, San Diego, La Jolla, CA 92093, U.S.A.

[4]Department of Family Medicine and Public Health, University of California, San Diego, La Jolla, CA 92093, U.S.A.

Correspondence
Ronghui Xu, Family Medicine and Public Health, University of California, San Diego, 9500 Gilman Drive, Mail Code 0112, La Jolla, CA 92093, U.S.A.
Email: rxu@ucsd.edu

Funding information
American Society of Clinical Oncology (ASCO); National Institutes of Health Clinical and Translational Science Award (CTSA), Grant/Award Number: UL1TR001442

Competing risk analysis considers event times due to multiple causes or of more than one event types. Commonly used regression models for such data include (1) cause-specific hazards model, which focuses on modeling one type of event while acknowledging other event types simultaneously, and (2) subdistribution hazards model, which links the covariate effects directly to the cumulative incidence function. Their use in the presence of high-dimensional predictors are largely unexplored. Motivated by an analysis using the linked SEER-Medicare database for the purposes of predicting cancer versus noncancer mortality for patients with prostate cancer, we study the accuracy of prediction and variable selection of existing machine learning methods under both models using extensive simulation experiments, including different approaches to choosing penalty parameters in each method. We then apply the optimal approaches to the analysis of the SEER-Medicare data.

KEYWORDS
boosting, cumulative incidence function, electronic medical record, LASSO, machine learning, precision medicine

1 | INTRODUCTION

As an illustration project of how information contained in patients' electronic medical records can be harvested for the purposes of precision medicine, we consider the large data set linking the Surveillance, Epidemiology and End Results (SEER) Program database of the National Cancer Institute with the federal health insurance program Medicare database for prostate cancer patients of age 65 or older. Each year, 180 000 men are examined with prostate cancer in the United States, and the important clinical decision commonly encountered in this patient population is whether to pursue aggressive cancer-directed therapy in the presence of preexisting comorbidities. Prostate cancer can progress slowly, and a proportion of men will die of competing causes before their prostate cancer becomes symptomatic. Current clinical guidelines for the management of prostate cancer instruct clinicians to make treatment decisions based on 2 factors: (1) an estimation of the aggressiveness of a patient's tumor and (2) estimation of a patient's overall life expectancy.[1] Classical cancer-specific survival prediction relies on 3 main risk factors: tumor stage, Gleason score, and prostate-specific antigen. On the other hand, currently, no tool exists to predict noncancer survival for this patient population. As

noncancer and cancer survival are not independent, ie, so-called competing risks in the statistical literature, an accurate comprehensive survival prediction tool should consider both types of risks simultaneously.

We restrict our analysis to patients examined between 2004 and 2009 in the SEER-Medicare database. After excluding additional patients with missing clinical records, we have a total of 57 011 patients who have information available on 7 relevant clinical variables (age, prostate-specific antigen, Gleason score, American joint committee on cancer (AJCC) stage, and AJCC stage T, N, M, respectively), 5 demographical variables (race, marital status, metro, registry, and year of diagnosis), plus 8971 binary insurance claim codes. We assumed that the survival prediction would occur at the time of diagnosis; therefore, we used clinical and demographic information at the time of diagnosis and insurance claims data during the year prior to diagnosis. Insurance claims capture medical diagnoses and procedures through health care common procedure coding system (HCPCS) codes, international classification of diseases, 9th revision (ICD-9) diagnosis codes, and ICD-9 procedure codes. These claims indirectly describe events that occur in surgical procedures, hospitalization, and outpatient activities. We converted each unique insurance claim code into a binary variable denoted 1 if the claim appeared anytime in the year before diagnosis and 0 if the code was absent. Until December 2013 (end of follow-up for these data), there were a total of 1247 deaths due to cancer and 5221 deaths unrelated to cancer. It is well understood that for time-to-event data, the number of events dictates the effective sample size, so in this case, we have more predictors to consider than the effective sample size.

Classical statistical methods, such as stepwise regression, have been known to suffer from model inconsistency and are computationally infeasible when the number of covariates is equal to or greater than the (effective) sample size. A group of machine learning methods, in particular supervised learning, has shown good performance empirically when the data are of high dimensionality.[2] The goals of these methods are[3] (1) prediction, to find a set of covariates which results in minimal prediction error in independent test data, and (2) variable selection, estimate the true sparsity pattern with low false positive rate for each covariate. In theory, consistent variable selection requires stronger assumptions, known as neighborhood stability or irrepresentable condition, which roughly translates to not too strongly correlated design, and the beta-min condition, which requires all nonzero coefficients to be sufficiently large. Both these conditions might be difficult to achieve for high-dimensional data in practice.[3] Therefore, methods like least absolute shrinkage and selection operator (LASSO) can be consistent for estimating the underlying regression function, for a properly chosen penalty parameter (see below also), but can perform very poorly for variable selection with strongly correlated design. Fortunately in our application, the prediction of mortality rates is of interest, and in this paper, we will focus on the performance of machine learning methods in estimating the true cumulative incidence function (CIF) at any given time. Meanwhile, during the process of simulation experiments, we also obtain results on variable selection as a side product and an added aspect.

Researchers have studied different approaches to analyze survival data with high-dimensional covariates. Notably, Tibshirani[4] proposed the LASSO under the Cox proportional hazards model. Zhang and Lu[5] investigated the statistical properties of adaptive LASSO for the Cox proportional hazards model. Hothorn et al[6] introduced a random forest algorithm and a generic gradient boosting algorithm for right censoring data. When considering theoretical aspects, Bradic et al[7] studied a group of penalty functions and established strong oracle properties of nonconcave penalized methods for ultrahigh-dimensional covariates in the presence of right censoring. In comparison, very few high-dimensional methods have been developed in the presence of competing risks. Binder et al[8] first proposed a boosting approach for fitting the proportional subdistribution hazards (PSDH) model. Ha et al[9] considered variable selection for clustered competing risks data under the subdistribution hazard frailty model. Very recently, Fu et al[10] considered penalized approaches under the same model. Given the high-dimensional nature of our data, in this paper, we will investigate the accuracy of variable selection and prediction using existing computational software under 2 commonly used models: the proportional cause-specific hazards (PCSH) model and the PSDH model. This leads to the approach of Binder et al under the PSDH model, and LASSO and adaptive LASSO approaches under the PCSH model, both being readily implemented and applicable to our large data set. In addition, we have implemented a LASSO algorithm under the PSDH model. All these approaches rely critically on the selection of a "penalty" parameter, and there are different ways to select this parameter. We will empirically evaluate these different methods using Monte Carlo simulations. The ultimate goal is to assess the prediction accuracy of the CIF as a risk assessment tool.

The remainder of this paper is organized as follows: In Section 2, we review the PCSH and the PSDH models. In Section 3, we review the relevant machine learning methods that have been or can be feasibly implemented to analyze competing risks data under each model. In Section 4, we conduct comprehensive simulation studies on these methods with varying numbers of predictors (relative to the sample size) that are continuous or binary (and in case of binary, sparse, or not sparse). In Section 5, we apply the machine learning methods under either model to classify prostate patients from the SEER-Medicare linked data into different risk groups according to their predicted CIFs. Finally, Section 6 contains discussion and directions for future work.

2 | COMPETING RISK MODELS

Competing risks occur when multiple types of failures coexist and the occurrence of one type of failure may prevent the observation of the other types of failure. In addition, the failure times may be subject to right censoring. Let $\epsilon = 1,...,J$ be the cause or type (we use the 2 words interchangeably in the following) of failure. Let $T = \min_{j=1}^{J} \tilde{T}_j$, where $\tilde{T}_j$ is the (possibly) latent failure time due to cause j. Let $X_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, where C_i is the potential censoring time and is assumed noninformative. Denote $S(t) = P(T > t)$ as the survival function of T. The CIF for failure type j is $F_j(t) = P(T \leq t, \epsilon = j)$. Obviously, $S(t) = 1 - \sum_{j=1}^{J} F_j(t)$, and $\sum_{j=1}^{J} F_j(\infty) = 1$. Denote the cause-specific hazard function of type j as $\lambda_j(t) = \lim_{\Delta t \to 0+} P(t \leq T < t + \Delta t, J = j | T \geq t)/\Delta t$. Then one can also show that

$$F_j(t) = \int_0^t \lambda_j(u) S(u) du, \tag{1}$$

leading to a nonparametric estimate of the CIF if we use[11]

$$\hat{\lambda}_j(t_i) = \frac{d_{ji}}{n_i}, \tag{2}$$

where d_{ji} denotes the number of failures from cause j at time t_i and n_i the number of subjects at risk at t_i, and

$$\hat{S}(t) = \prod_{j:t_i \leq t} \left\{ 1 - \sum_{j=1}^{J} \hat{\lambda}_j(t_i) \right\}. \tag{3}$$

Then we have $\hat{F}_j(t) = \sum_{i:t_i \leq t} \hat{p}_j(t_i)$, where $\hat{p}_j(t_i) = \hat{\lambda}_j(t_i) \hat{S}(t_i^-)$. While $\hat{F}_j(t)$ is a complex function of the $\hat{\lambda}_j(t_i)$'s, the $(1-\alpha)100\%$ pointwise (for each t) confidence intervals can calculated using the "Cuminc()" function in the R package "mstate."

2.1 | The PCSH model

In the regression settings, the Cox-type proportional hazards model can be used to model the above cause-specific hazards, and existing software for fitting the Cox model for classical survival data without competing risks can be used to fit the PCSH model.[12] Under this model, however, the dependence of the CIF of a particular failure type on the covariates involves also the effects of the covariates on the cause-specific hazards of all other types of failures.

Given a p-dimensional vector of covariates Z, under the proportional hazards assumption of the cause-specific hazard function, we have

$$\lambda_j(t|Z) = \lambda_{0j}(t) \exp(\beta_j' Z), \tag{4}$$

for $j = 1,...,J$. To estimate β_j, we can use any software for the regular Cox model for one cause at a time, by treating all other types of events as if censored. This is because the (partial) likelihood for all event types factors into a separate likelihood function for each event type, and the likelihood function for each event type treats all other types of events as if censored.

To estimate the CIF given $Z = z_0$, we have similar to the above nonparametric estimation:

$$\begin{aligned} \hat{F}_j(t; z_0) &= \int_0^t \hat{S}(u; z_0) d\hat{\Lambda}_j(u; z_0) \\ &= \sum_{i=1}^{n} \frac{\hat{S}(X_i; z_0) \delta_{ji} I(X_i \leq t) \exp(\hat{\beta}_j' z_0)}{\sum_{i'=1}^{n} I(X_i \leq X_{i'}) \exp(\hat{\beta}_j' Z_{i'})}, \end{aligned} \tag{5}$$

where $\hat{S}(u; z_0) = \exp\{-\sum_{j=1}^{J} \hat{\Lambda}_j(u; z_0)\}$, $\hat{\Lambda}_j(u; z_0) = \hat{\Lambda}_{0j}(u) \exp(\hat{\beta}_j' z_0)$, and $\hat{\Lambda}_{0j}(u)$ is a Breslow-type estimator of the baseline cumulative hazard.[13] Notice that in estimating the overall survival function $\hat{S}$, we need to fit the models for all event types, even if we are only interested in the CIF of type j.

A $(1-\alpha)100\%$ pointwise confidence interval can be computed following Cheng et al.[14] We implemented an R package "CompetingRisk"[15] to compute the above estimator of the CIF with its pointwise confidence intervals.

2.2 | The PSDH model

To link the covariates directly to the CIF, Fine and Gray[16] proposed to model the so-called subdistribution hazards. The proportional hazards modeling of the subdistribution hazards, also known as Fine-Gray model, has gained popularity in recent years.

Gray[17] introduced the subdistribution hazard function as $\tilde{\lambda}_j(t) = -\frac{d}{dt}\log\{1-F_j(t)\}$. Under the proportional hazards assumption of the subdistribution hazard function for cause 1, we have[16]

$$\tilde{\lambda}_1(t|Z) = \tilde{\lambda}_{01}(t)\exp(\beta' Z). \tag{6}$$

It is easy to see that model (6) provides a direct way to estimate the CIF of cause 1, so that there is no need to fit models for the other causes to estimate F_1.

Fine and Gray[16] proposed estimating equations for β. Geskus[18] further showed that these estimating equations can be solved using weighted Cox regression, ie, software for the regular Cox model incorporating weights. The baseline subdistribution hazard is again estimated using a modified version of the Breslow estimator. The $(1-\alpha)100\%$ pointwise confidence intervals can be constructed by sampling standard normal random variables and otherwise closed-form formulas.[16]

The PCSH and the PSDH models are typically not valid at the same time, and limited empirical experiences seem to indicate that in real data applications, the 2 models can lead to similar conclusions.[19] In addition, Lambert et al[20] considered flexible parametric modeling of the CIF given covariates using splines. Koller et al[21] argued favorably for using the PCSH for etiology or efficacy hypotheses and the PSDH for prognosis in clinical settings. Our data are more complex than the more classic clinical data in that they include many treatment procedures captured by the claims codes, as well as prognostic variables. From a modeling point of view, both models might be used to approximate the underlying but unknown data-generating mechanism. In the following, we consider both the PCSH and the PSDH models.

3 | REGULARIZATION

Unlike classical statistical methods, machine learning methods aimed at high-dimensional covariates often involve the selection of a tuning parameter, based on the minimal estimated prediction error. There are 2 ways to estimate this prediction error: cross-validation, which is computationally intensive, or approximation methods such as the C_p-type statistics. When a log-likelihood loss function is used, the latter leads to the well-known Akaike information criterion (AIC). Another commonly used information-based criterion is the Bayesian information criterion (BIC), which imposes a larger penalty than the AIC.

To choose the tuning parameter, denoted by λ below, we consider the following different methods:

- CV10: λ associated with the minimum 10-fold cross-validated (CV) prediction error (referred to as "error" in the following);
- CV + 1SE: λ associated with the minimum 10-fold CV error plus one standard error of the CV estimated errors;
- min AIC/BIC: λ associated with the minimum AIC or BIC criteria;
- elbow AIC / BIC: λ associated with the largest descent in AIC or BIC.

Under the Cox regression model, the AIC is defined as $-2\log(L) + 2s$, where L is the partial likelihood and $s = |S(\hat{\boldsymbol{\beta}})|$ is the number of nonzero estimated regression coefficients, ie, the size of the estimated active set $S(\hat{\boldsymbol{\beta}})$.[22,23] Bayesian information criterion under the Cox model is defined as $-2\log(L) + 2s\log(k)$, where k is the number of observed uncensored events.[24] We apply these definitions to the PCSH and the PSDH models below, where k is the number of observed events from the cause of interest, and the error is the negative log partial likelihood. The "elbow" criteria are described in Tibshirani et al[25] as a way to avoid overselection in practice.

3.1 | Least absolute shrinkage and selection operator

Least absolute shrinkage and selection operator is an L_1 penalization method proposed by Tibshirani[26] for building parsimonious models when the performance of classical methods such as stepwise regression or best subset selection is not satisfactory. For linear regression, LASSO solves a penalized least squares problem along the regularization path, where the regression coefficients associated with unimportant covariates shrink to exactly zero while granting nonzero coefficients for important covariates. The theoretical properties of LASSO have been extensively studied under the linear regression model. Meinshausen and Bühlmann[27] showed consistency of LASSO under the neighborhood stability condition, when the true nonzero coefficients are sufficiently large in absolute value. This condition is equivalent to the irrepresentable condition used by Zhao and Yu.[28] Although some of these theoretical conditions might be difficult to achieve in practice, LASSO has gained numerous attention as a technique to reduce dimensionality and construct predictive models. One of the main reasons for its popularity is its computational simplicity, involving convex optimization only.

Tibshirani[4] extended LASSO to the Cox regression model. Since the Cox regression software is typically used to fit the PCSH model, we may apply the same LASSO algorithm as proposed in Tibshirani.[4] Under the PCSH model, the partial likelihood for event type j is

$$L_j(\beta_j) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta_j' Z_i)}{\sum_{l \in R_i} \exp(\beta_j' Z_l)} \right\}^{\delta_i I(\epsilon_i = j)}, \tag{7}$$

where $R_i = \{l : X_l \geq X_i\}$ is the risk set at time X_i. The LASSO estimator of β_j minimizes the L_1-penalized log partial likelihood:

$$-\frac{1}{n} \log L_j(\beta_j) + \lambda \sum_{k=1}^{p} |\beta_{jk}|, \tag{8}$$

where the penalty parameter λ will be chosen by the methods described above in our simulations. We note that the R package "glmnet" implementation of LASSO has been widely used and is able to fix Cox regression model as well as the PCSH model with large data sets.

Under the PSDH model,[16] the pseudolikelihood used for inference is the same as the partial likelihood for complete (ie, no censoring) data, but otherwise with weights in the risk sets to account for censoring:

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta' Z_i)}{\sum_{l \in R_i} w_l(X_i) \exp(\beta' Z_l)} \right\}^{I(\delta_i \epsilon_i = 1)}, \tag{9}$$

where $R_i = \{l : X_l \geq X_i \text{ or } \delta_l \epsilon_l > 1\}$ is the risk set consisting of individuals who have not had any event or who have had an event of other causes and $w_l(t) = \hat{G}(t) I(t \geq X_l) \delta_l / \hat{G}(X_l) + I(t < X_l)$ where $\hat{G}(t)$ is the Kaplan-Meier estimate of $P(C > t)$. The LASSO estimator is similar to the above under the PSDH model, ie, it minimizes:

$$-\frac{1}{n} \log L(\beta) + \lambda \sum_{k=1}^{p} |\beta_k|, \tag{10}$$

where the penalty parameter λ will again be chosen by the methods described above. We note that while Fu et al[10] considered the LASSO and other penalized approaches under the PSDH model, we were not able to apply the associated R package "crrp" to the linked SEER-Medicare data as it ran out of memory. Instead, we have developed our own R program using gradient descent algorithm written in Fortran, which is in the process of being made into a package.

Tibshirani[26] discussed the standard errors of the LASSO estimator, which was approximated by a sandwich-type covariance estimator based on the penalized log-likelihood. However, recent works on inference under high dimensions by Zhang and Zhang[29] and van de Geer et al[30] have shown that such regularized estimators are generally biased (at $\sqrt{n}$-rate) and proposed bias correction procedures for inference purposes. A related group of authors of this paper is currently developing a similar inference procedure under the PSDH model in high dimensions (see https://arxiv.org/abs/1707.09561).

3.2 | Adaptive LASSO

Adaptive LASSO was initially developed by Zou.[31] As mentioned earlier, Zhang and Lu[5] proposed adaptive LASSO for the Cox proportional hazards model, where adaptively weighted L_1 penalties were used. Instead of the LASSO penalty above, the adaptive LASSO penalty has the form $\lambda\sum_{k=1}^{p}|\beta_k/\tilde{\beta}_k|$, where $\tilde{\beta} = (\tilde{\beta}_1, \ldots \tilde{\beta}_p)$ maximized the (partial) likelihood without any penalty. This way, small weights are given to large coefficients and large weights for small coefficients. The resulting estimator enjoys the oracle properties (not enjoyed by LASSO) and has a convex penalty form that ensures the existence of global optimizers and can be efficiently solved by standard algorithms just like LASSO.

The algorithms described above for LASSO can be readily applied to adaptive LASSO under both the PCSH and the PSDH models.

3.3 | Boosting

Freund and Schapire[32] introduced the AdaBoost algorithm to solve classification problems by combining rough and moderately accurate "rules of thumb" repeatedly. Later, Friedman[33] developed boosting methods for linear regression as a numerical optimization method to minimize the squared error loss function. Boosting can be viewed as a gradient descent optimization algorithm in function space and is essentially the same as the matching pursuit algorithm in signal processing.[34] Bühlmann [35] proved that boosting with the squared error loss is consistent in high-dimensional linear models, where the number of predictors is allowed to grow as fast as exponential to the sample size.

For the PSDH model with high-dimensional data, Binder et al[8] proposed a likelihood-based boosting approach. Binder et al[8] applied the componentwise boosting approach of Bühlmann and Yu[36] to the likelihood in (9): Starting from initial value zero for β, at step k, for component j of β, let $\gamma_{k,j} = \beta_{k,j} - \beta_{k-1,j}$ be the potential update of that component from the previous step $(k-1)$ to the current step k, and find $\gamma_{k,j}$ by minimizing

$$l_{pen}(\gamma_{k,j}) = -\log L_j(\beta_{k-1,1}, \ldots, \beta_{k-1,j-1}, \beta_{k,j}, \beta_{k-1,j+1}, \ldots, \beta_{k-1,p}) + \frac{\lambda}{2}\gamma_{k,j}^2; \tag{11}$$

then among all $j = 1,\ldots,p$, only update the component j that has the largest score statistic $U'_{pen}(0)I^{-1}_{pen}(0)U_{pen}(0)$ where $U_{pen}(\gamma) = \partial l_{pen}(\gamma)/\partial\gamma$ and $I_{pen}(\gamma) = \partial^2 l_{pen}(\gamma)/\partial\gamma^2$. For boosting, the number of steps is the main tuning parameter, while λ in (11) is chosen so that the number steps is greater than 50.[8] The R package "CoxBoost" can be used, and the number of boosting steps will be chosen by the methods described above in our simulations.

4 | SIMULATIONS

4.1 | Setup

To investigate the performance of LASSO, adaptive LASSO, and boosting under the PCSH and the PSDH models, respectively, we conducted comprehensive simulation studies with both continuous and dichotomized covariates in competing risks data. We assumed $J = 2$, and we considered sample size $n = 500$ and number of covariates $p = 20$, 500, and 1000. We repeated each simulation setting 100 times.

For continuous covariates, the covariate vector for each subject was generated for the following correlation structures:

1. Independent: Each covariate was independently generated from $N(0,1)$.
2. Exchangeable: The covariate vector was generated from a multivariate normal distribution with mean zero, marginal variance of one, and a block diagonal covariance matrix—each block of size 10 and within a block the pairwise correlation $\rho(i,i') = 0.5$.
3. AR(1): The covariate vector was generated from a multivariate normal distribution with mean zero, marginal variance of one, and a block diagonal covariance matrix—each block of size 10 and within a block the pairwise correlation $\rho(i,i') = 0.5^{|i-i'|}$.

For binary covariates, the covariate vector was first generated the same as in the above, then dichotomized at threshold a, with $<a$ coded as 1 and 0 otherwise. We considered $a = 0$, to give a balanced binary distribution, and

$a=-1$, to give a relatively sparse 16% of 1's. We set the number of nonzero regression coefficients, ie, the size of the active set, to be $s_1=5$ and $s_2=3$ for causes 1 and 2, respectively. We let $\beta_{1,1\cdots,5}=(1.96,-0.79,-0.5,-1.35,1.29)$, $\beta_{2,11\cdots,13}=(-1.16,-0.86,0.5)$, and the rest of the β_1 and β_2 values were zero. These β values were used under both the PCSH and the PSDH models.

To simulate survival outcomes under the PCSH model, we followed the approach described in Beyersmann et al[37]; that is, we simulated the event time T first, then we simulated the cause ϵ given T. We assumed the baseline hazard functions for types 1 and 2 failures to be $\lambda_{01}(t)=0.15$ and $\lambda_{02}(t)=0.10$, respectively. The overall (not cause-specific) cumulative hazard function for T was then $\Lambda(t|z)=t\{\lambda_{01}\exp(\beta_1{}'z)+\lambda_{02}\exp(\beta_2{}'z)\}$, and T was generated using the fact that $U=\exp(-\Lambda(T))\sim U(0,1)$ given z. The cause ϵ was generated proportional to the cause-specific hazard function, ie, $P(\epsilon=1|z)=\lambda_{01}\exp(\beta_1{}'z)/\{\lambda_{01}\exp(\beta_1{}'z)+\lambda_{02}\exp(\beta_2{}'z)\}$. Under this model, the true CIF for cause j was

$$F_j(t|z)=\int_0^t S(u|z)\lambda_{0j}\exp(\beta_j{}'z)du=\lambda_{0j}\exp(\beta_j{}'z)\frac{e^{tM}}{M}, \tag{12}$$

where $M=-\{\lambda_{01}\exp(\beta_1{}'z)+\lambda_{02}\exp(\beta_2{}'z)\}$. The censoring times were generated from $U(0,20)$, which resulted in an average event rate of 45.8% for cause 1 and 33.6% for cause 2 with continuous covariates, an average event rate of 51.8% for cause 1 and 27.2% for cause 2 with balanced binary covariates, and an average event rate of 59.8% for cause 1 and 17.8% for cause 2 with sparse binary covariates..

To simulate under the PSDH model, we followed the approach described in Fine and Gray.[16] The CIF for failure from cause 1 was given by

$$F_1(t|z)=P(T\le t,\epsilon=1|z)=1-\{1-p(1-e^{-t})\}^{\exp(\beta_1'z)}, \tag{13}$$

where we used $p=0.6$. As this was a subdistribution function, with a point mass $1-F_1(\infty|z)$ at infinity, the proper distribution function that was used to generate T was $F(t|z)=F_1(t|z)/F_1(\infty|z)$, so that $F(T)\sim U(0,1)$ given z. Note that $P(\epsilon=1|z)=F_1(\infty|z)$ and $P(\epsilon=2|z)=1-P(\epsilon=1|z)$. Finally, the event times for failure from cause 2 were generated according to an exponential distribution with rate $\exp(\beta_2'z)$. The censoring times were generated from $U(0,20)$, resulting in an average event rate of 53.5% for cause 1 and 35.1% for cause 2 with continuous covariates, an average event rate of 55.8% for cause 1 and 33.4% for cause 2 with balanced binary covariates, and an average event rate of 55.5% for cause 1 and 33.5% for cause 2 with sparse binary covariates.

We also considered higher cause 2 event rates at the suggestion of a reviewer, and more details are provided in the Supporting Information.

4.2 | Results

We evaluate the performance of prediction at a given covariate vector value z_0. We set $z_0=(0.5,\cdots,0.5)_{1\times p}$ for the continuous case; and for all the binary cases, each element of z_0 was independently drawn with a fixed seed from Bernoulli distribution with $p=0.5$.

Figures 1 to 3 show the empirical distributions of the estimated $F_1(2|z_0)$ over the 100 simulation runs, where the vertical line marks the true $F_1(2|z_0)$; the empirical distributions were plotted using the R function "density()." The results under the PCSH model using adaptive LASSO and under the PSDH model using adaptive LASSO and boosting are given in these figures for sparse binary covariates, while results for continuous and balanced binary covariate, as well as additional results under both the PCSH and the PSDH models using LASSO, are given in the Supporting Information due to limitation of space.

In the figures, the blue dashed lines are for the oracle estimator, which fits the exact true active set $S(\beta)$. The oracle estimator varied extremely slightly with the 3 correlation structures for Z, any variation appearing due to Monte Carlo, and the one under the AR(1) structure is plotted here. It is seen that the distribution of the oracle estimator is more concentrated for the continuous covariates than for the sparse binary covariates, which reflects the "effective sample size" that is reduced with the sparse binary covariates. The solid lines are the estimated $F_1(2|z_0)$ under each model after regularization, with different colors representing different correlation structures of Z.

Under the PCSH model using adaptive LASSO to regularize (Figure 1 and Supporting Information), the performances were generally not satisfactory as compared with the oracle estimator for $p=500$ and 1000. The worst performances were seen when using minimum AIC and BIC to choose the penalty parameter; some of these results

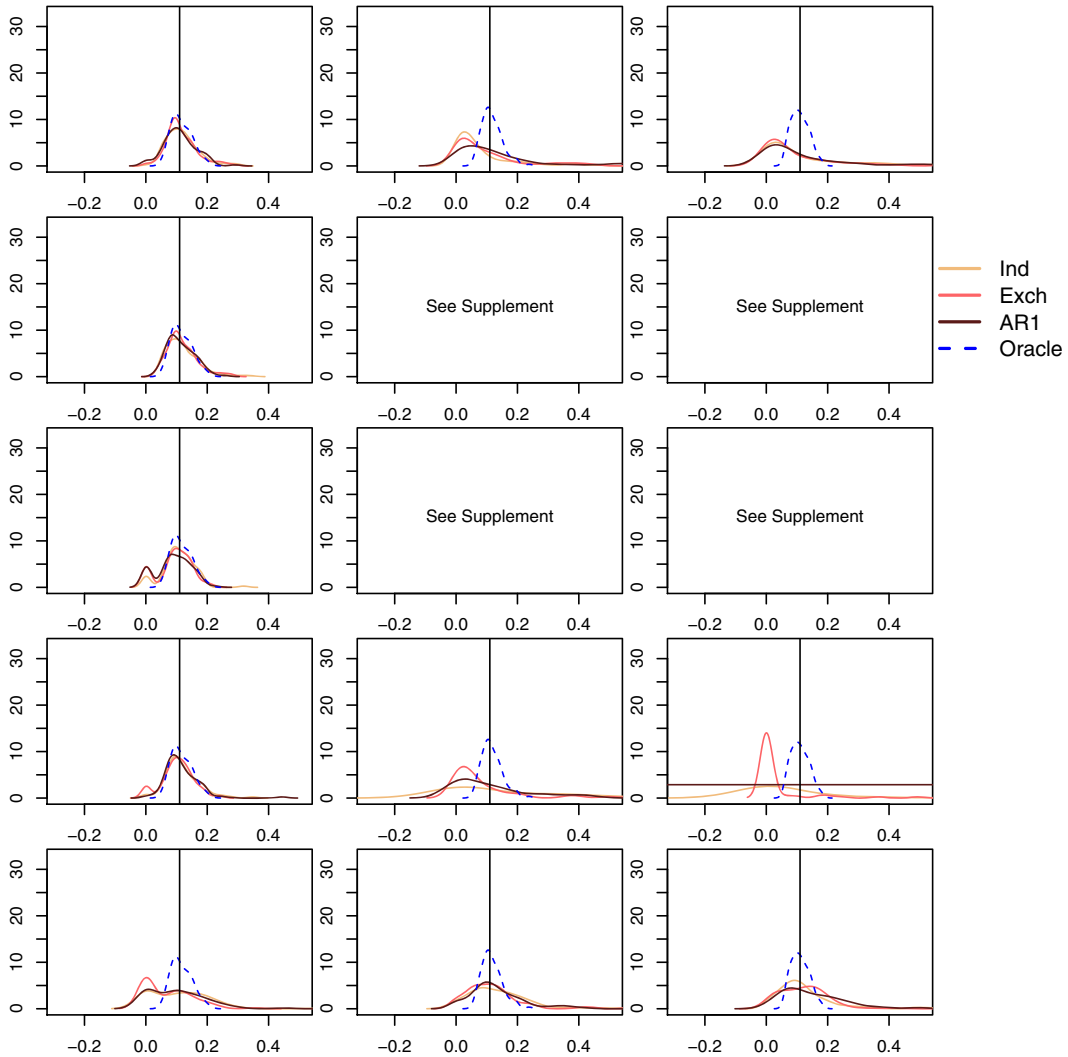


FIGURE 1 The (smoothed) empirical distribution of $\hat{F}_1(2|z_0)$ under the proportional cause-specific hazards model with adaptive least absolute shrinkage and selection operator, for sparse binary covariates. The 3 columns correspond to $p = 20$, 500, and 1000. The rows correspond to different ways of selecting penalty parameter λ, from top to bottom: (1) CV10, (2) minimum Akaike information criterion (AIC), (3) minimum Bayesian information criterion (BIC), (4) elbow AIC, and (5) elbow BIC. The true $CIF_1(t = 2|z_0) = 0.11$ [Colour figure can be viewed at wileyonlinelibrary.com]

were so extreme that "density()" failed to work and these were instead shown in the Supporting Information using boxplots. Elbow BIC appeared to perform the best for continuous covariates, but not so for binary covariates even when $p = 20$. CV10 had the best performance for binary covariates in general, but it too deteriorated for $p =$ 500 and 1000.

Under the PSDH model using adaptive LASSO (Figure 2 and Supporting Information), the results appeared to be similar to those under the PCSH model using adaptive LASSO for continuous covariates. For the sparse binary covariates, even the oracle estimator had a very wide spread, and the results were generally not satisfactory for $p = 500$ and 1000.

Under the PSDH model using boosting (Figure 3 and Supporting Information), we see that for continuous covariates, the estimators performed reasonably well when CV10 or min AIC/BIC was used to choose the number of boosting steps; with CV10, the estimation was perhaps the best. The performance deteriorated with binary covariates for $p = 500$ and 1000, with the performance under the independent structure the worst of all. We note that in Bühlmann[35] simulation studies (their table 1), the mean squared error for boosting with correlated design was also smaller than that with uncorrelated design, and their Figure 1 showed that boosting tended to overselect covariates in the uncorrelated design than the correlated design.

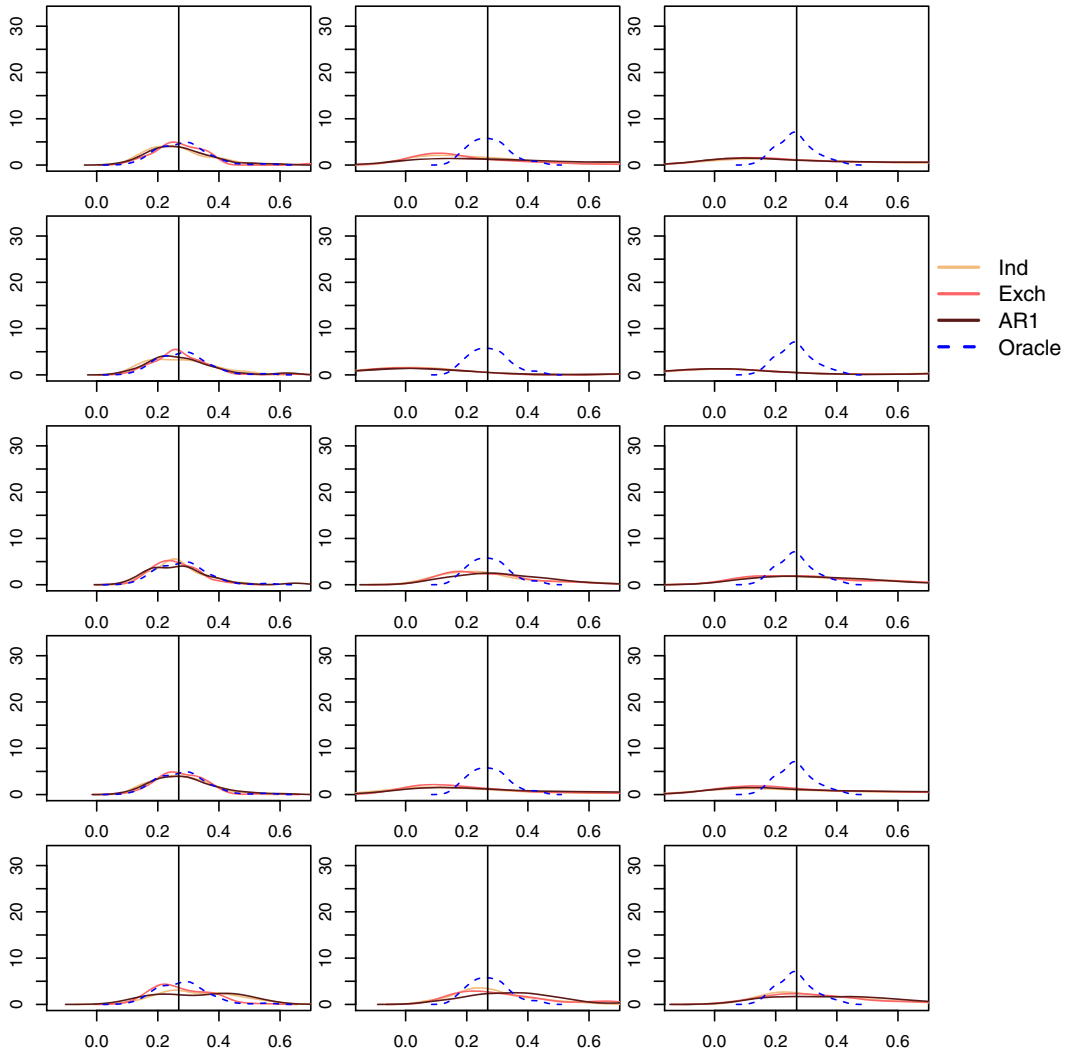


FIGURE 2 The (smoothed) empirical distribution of $\hat{F}_1(2|z_0)$ under the proportional subdistribution hazards model with adaptive least absolute shrinkage and selection operator, for sparse binary covariates. The 3 columns correspond to $p = 20$, 500, and 1000. The rows correspond to different ways of selecting penalty parameter λ, from top to bottom: (1) CV10, (2) minimum Akaike information criterion (AIC), (3) minimum Bayesian information criterion (BIC), (4) elbow AIC, and (5) elbow BIC. The true $CIF_1(t = 2|z_0) = 0.27$ [Colour figure can be viewed at wileyonlinelibrary.com]

Table 1 gives the mean integrated squared errors (MISEs) of $\hat{F}_1(t|z_0)$ from t= 0 to 20 (using numerical integration with increments of 0.1), for all 5 combinations of models and regularization methods when the penalty parameter was chosen using CV10. We note that $t = 20$ is the maximum follow-up time. In the Supporting Information, we also provide the MISEs for other methods of selecting the penalty parameter (min AIC, min BIC, elbow AIC, and elbow BIC), and in general, CV10 outperformed these other selection methods. It is seen from Table 1 that LASSO and adaptive LASSO were generally comparable under the PCSH model, with LASSO having slightly smaller MISE for sparse binary covariates and when p was large. Adaptive LASSO had smaller MISE than LASSO under the PSDH model for continuous covariates, but not always so for binary covariates, especially when p was large. Boosting under the PSDH model had much smaller MISE compared with LASSO or adaptive LASSO.

The Supporting Information provide additional simulation results including, among other things, variable selection results. Variable selection was not the main goal for our application, but it was a necessary step before prediction and the results helped to explain the prediction accuracy. Often when the prediction results were poor (as discussed above), it was because variable selection was poor; for example, when there were a couple of hundred false positives, the corresponding prediction results were extremely poor. Boosting had no more than 5 false positives in all cases. We also note that elbow BIC had a tendency to underselect, ie, with relatively more false negatives.

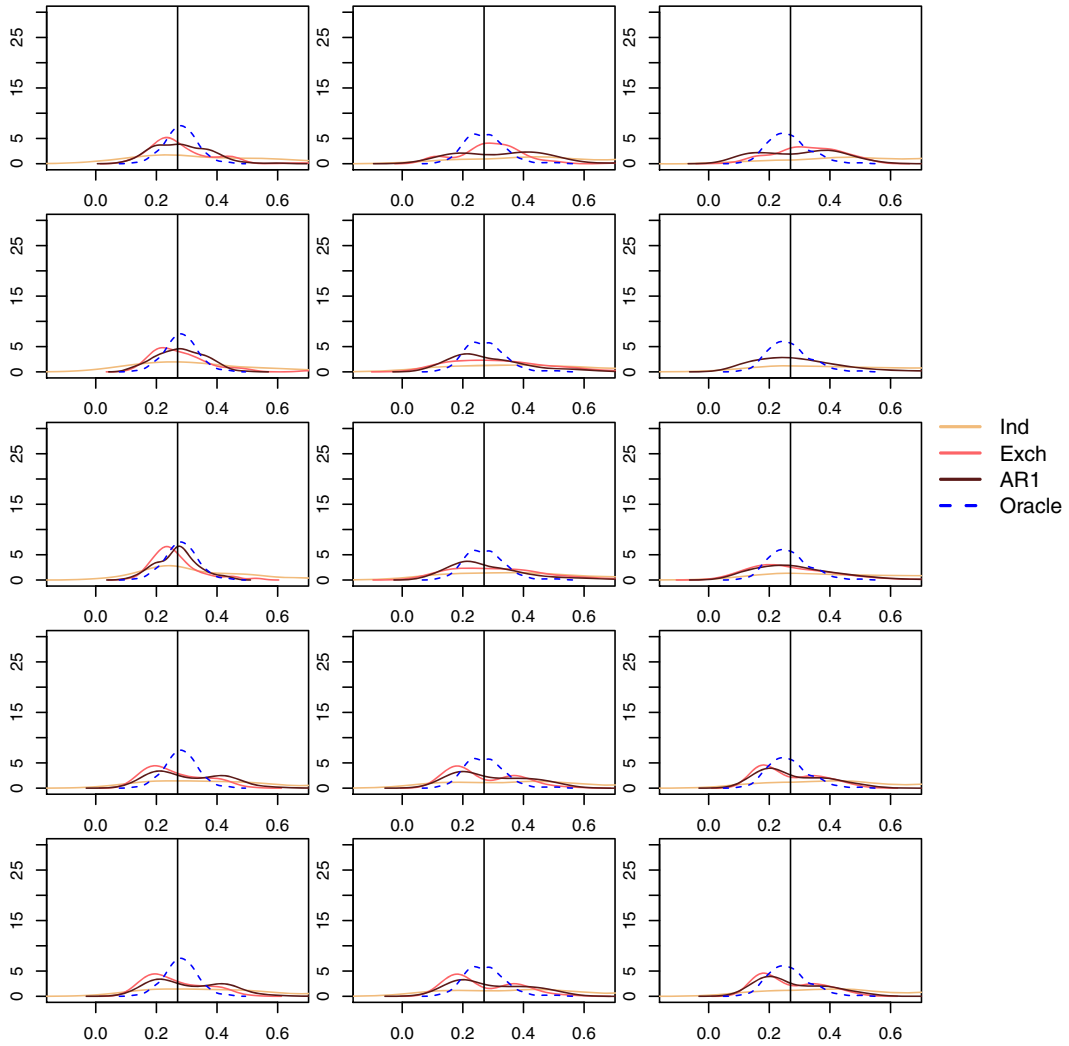


FIGURE 3 The (smoothed) empirical distribution of $\hat{F}_1(2|z_0)$ under the proportional subdistribution hazards model with boosting, for sparse binary covariates. The 3 columns correspond to $p = 20$, 500, and 1000. The rows correspond to different ways of selecting γ, from top to bottom: (1) CV10, (2) minimum Akaike information criterion (AIC), (3) minimum Bayesian information criterion (BIC), (4) elbow AIC, and (5) elbow BIC. The true $F_1(2|z_0) = 0.27$ [Colour figure can be viewed at wileyonlinelibrary.com]

5 | SEER-MEDICARE LINKED DATA

We randomly split the SEER-Medicare data set into approximate equal-sized training ($n = 28{,}505$) and test ($n = 28{,}506$) data sets. As a first step, we excluded binary claim codes having less than 10 ones, with rest all being zero. We then used univariate screening to further reduce dimensionality, eliminate noise, and increase the performance of subsequent variable selection methods.[38]

5.1 | PCSH model with LASSO and adaptive LASSO

Univariate screening under the PCSH model with P value cutoff of .05 gave $p_1 = 2188$ and $p_2 = 1079$ claim codes for noncancer and cancer mortalities, respectively. For each type of mortality, we first applied LASSO under the PCSH model on the training data with the above prescreened claim codes plus the clinical and demographic variables. Based on the simulation results, CV10 was used to choose the penalty parameter. The final model contained 143 predictors for noncancer mortality and 9 predictors for cancer mortality. Since the regression coefficients from LASSO are biased, we refit the PCSH model with the selected predictors.

TABLE 1 Mean integrated squared errors of $\hat{F}_1(t|z_0)$; the penalty parameter was chosen using CV10

Covariates	PCSH		PSDH		
	LASSO	Ad. LASSO	LASSO	Ad. LASSO	Boosting
Continuous					
$p = 20$					
Independence	0.07	0.05	0.08	0.06	0.03
Exchangeable	0.03	0.03	0.03	0.03	0.02
AR1	0.04	0.04	0.04	0.03	0.02
$p = 500$					
Independence	0.11	0.21	0.50	0.19	0.03
Exchangeable	0.27	0.23	0.43	0.28	0.02
AR1	0.30	0.20	0.40	0.31	0.02
$p = 1000$					
Independence	0.12	0.28	0.65	0.24	0.02
Exchangeable	0.25	0.24	0.50	0.39	0.03
AR1	0.29	0.36	0.48	0.34	0.02
Balanced binary					
$p = 20$					
Independence	0.13	0.16	0.11	0.08	0.09
Exchangeable	0.17	0.17	0.11	0.10	0.02
AR1	0.20	0.21	0.10	0.07	0.04
$p = 500$					
Independence	0.62	0.72	0.47	0.52	0.09
Exchangeable	0.71	0.97	0.47	0.50	0.07
AR1	0.74	0.82	0.59	0.66	0.05
$p = 1000$					
Independence	0.71	1.13	0.79	1.12	0.18
Exchangeable	0.68	1.20	0.71	0.90	0.04
AR1	0.81	1.27	1.14	1.28	0.05
Sparse binary					
$p = 20$					
Independence	0.36	0.34	0.30	0.21	0.03
Exchangeable	0.35	0.34	0.23	0.20	0.01
AR1	0.31	0.38	0.30	0.26	0.01
$p = 500$					
Independence	0.95	1.87	1.10	1.40	0.08
Exchangeable	1.28	1.50	1.23	1.70	0.02
AR1	1.32	1.50	1.19	1.98	0.03
$p = 1000$					
Independence	1.29	1.84	2.00	2.70	0.09
Exchangeable	1.30	1.99	2.21	2.15	0.03
AR1	1.39	2.06	1.27	2.30	0.04

Abbreviations: LASSO, least absolute shrinkage and selection operator; PCSH, proportional cause-specific hazards; PSDH, proportional subdistribution hazards.

To evaluate the resulting prediction model on the test data, we first calculated the risk score $\hat{\beta}_j'Z$ for each patient in the test data, $j = 1,2$. For each mortality type j, we then divided the test set into 4 risk strata: low (L), median low (ML), median high (MH), and high (H) according to the quartiles. Combining the 2 types of mortalities, we formed a total of 16 strata for their predicted CIF. Using the average Z values in each of the 16 strata, we plotted their predicted CIFs for both cancer and noncancer mortalities, and the results are shown in the Supporting Information. It was clear that instead of 16 groups, 5 distinct risk groups emerge for both cancer and noncancer mortalities. We note that these are not the same 5 groups for the 2 types of mortality, and the Supporting Information provides the definition for each of them. It is perhaps not surprising to see that each mortality risk was most influenced by the corresponding cause-specific risk and also secondarily by its competing risk.

In Figures 4 and 5, we plot the nonparametric CIF (solid lines) and its 95% confidence intervals (shaded) for each of the above 5 risk groups for noncancer and cancer mortalities, respectively. The clear separation of the 5 groups shows the usefulness of the final PCSH model in classifying patients according to their different prognosis for both cancer and noncancer mortalities. For comparison purposes, we also plot the predicted CIF for each of the 5 risk groups. While the prediction is more accurate for the noncancer CIF, the predicted cancer CIF seems less accurate especially for the high (H) risk group.

We then applied adaptive LASSO under the PCSH model in a similar fashion, and the results are also given in Figures 4 to 5 and the Supporting Information. The final model contained 113 predictors for noncancer mortality and 11 predictors for cancer mortality. Overall, the results were somewhat similar to those from LASSO, with also 5 visual groups from the original 16. The prediction results also appeared generally comparable between these 2 regularization methods, with prediction more accurate using one method sometimes and less so the other times.

5.2 | PSDH model with boosting, LASSO, and adaptive LASSO

Univariate screening with P value cutoff of .05 under the PSDH model initially gave $p_1 = 4634$ and $p_2 = 6088$ claim codes for noncancer and cancer mortalities, respectively. We further reduced the dimension by retaining only the top 2000 claim codes (ranked by P value) for each type of mortality, to be comparable with the fitting of the PCSH model above, as well as for the boosting algorithm to be able to run on our Dell R630 computer (2 Intel Xeon E5-2660 v3 2.6 GHz, each processor with 10 cores [20 threads] for a total of 20 cores [40 threads], and 128 GB of DDR3). We applied boosting under the PSDH model to the training data with these claim codes plus the clinical and demographic variables. Although CV10 performed best in our simulation results, AIC was a close second especially for binary covariates and was less computationally intensive for this procedure where boosting itself was computationally intensive already. Therefore, AIC was used to choose the optimal step. The final model contains 53 predictors for noncancer mortality and 13 predictors for cancer mortality. We refit the PSDH model with the selected predictors to obtain the unbiased estimator.

Similar to the above, we calculated the risk score $\hat{\beta}_j'Z$ for each patient in the test data, $j = 1,2$. For each mortality type j, we divided the test set into 5 risk strata: low (L), median low (ML), median (M), median high (MH), and high (H) according to the quintiles. Again, the classification was not the same 5 groups for the 2 types of morality. In Figures 4 and 5, we plot the nonparametric CIF (solid lines) and its 95% confidence intervals (shaded) for the above 5 risk groups for both noncancer and cancer mortalities. We also plot the predicted CIF for each of the 5 risk groups.

We then applied LASSO and adaptive LASSO under the PSDH model. While LASSO selected 143 predictors for noncancer mortality and 22 predictors for cancer mortality in the final models, adaptive LASSO selected 1045 predictors for noncancer mortality and 153 predictors for cancer mortality. The predicted and nonparametric estimates of the CIFs are also plotted in Figures 4 and 5. It is seen that the prediction results are generally similar to those from boosting. The prediction was still more accurate for the noncancer CIF, and less so for cancer CIF especially for the high (H) risk group.

6 | DISCUSSION

The rapid accumulation of data across many fields, medicine in particular, has created unique challenges in statistics. The distinct issues with high-dimensional data have come to be recognized recently, including, for example, the rapid noise accumulation, the unrealistic independence assumption, and the necessity for novel robust data analysis methods.[39] While researchers work to meet these challenges, some of the methods proposed in the literature do not necessarily scale well to large data sets.

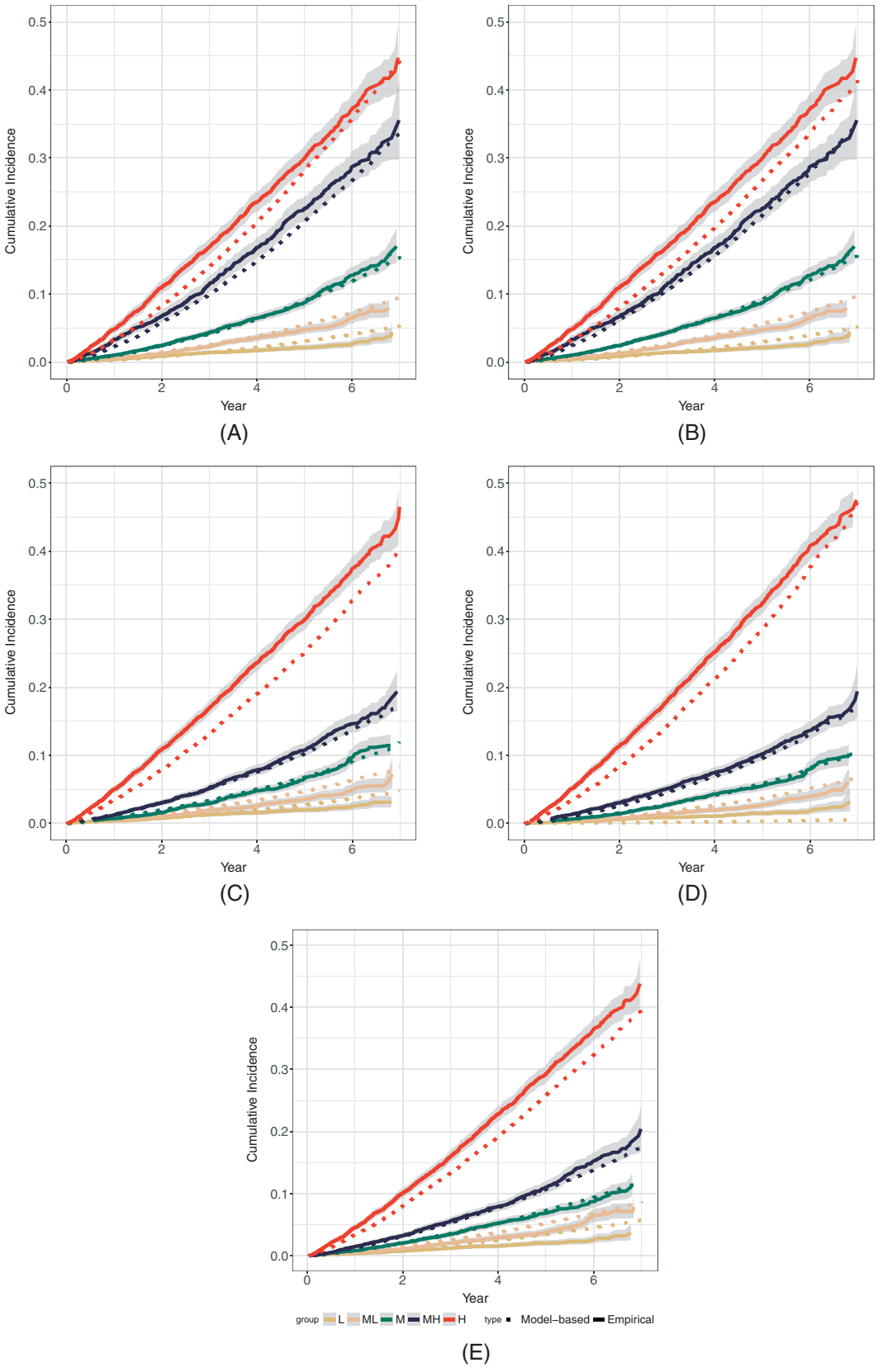


FIGURE 4 Cumulative incidence functions for noncancer mortalities, with classification and prediction using A, LASSO and B, adaptive LASSO under the PCSH model and C, LASSO, D, adaptive LASSO, and E, boosting under the PSDH model. The shaded area is the 95% pointwise confidence intervals based on the nonparametric estimate. LASSO, least absolute shrinkage and selection operator; PCSH, proportional cause-specific hazards; PSDH, proportional subdistribution hazards [Colour figure can be viewed at wileyonlinelibrary.com]

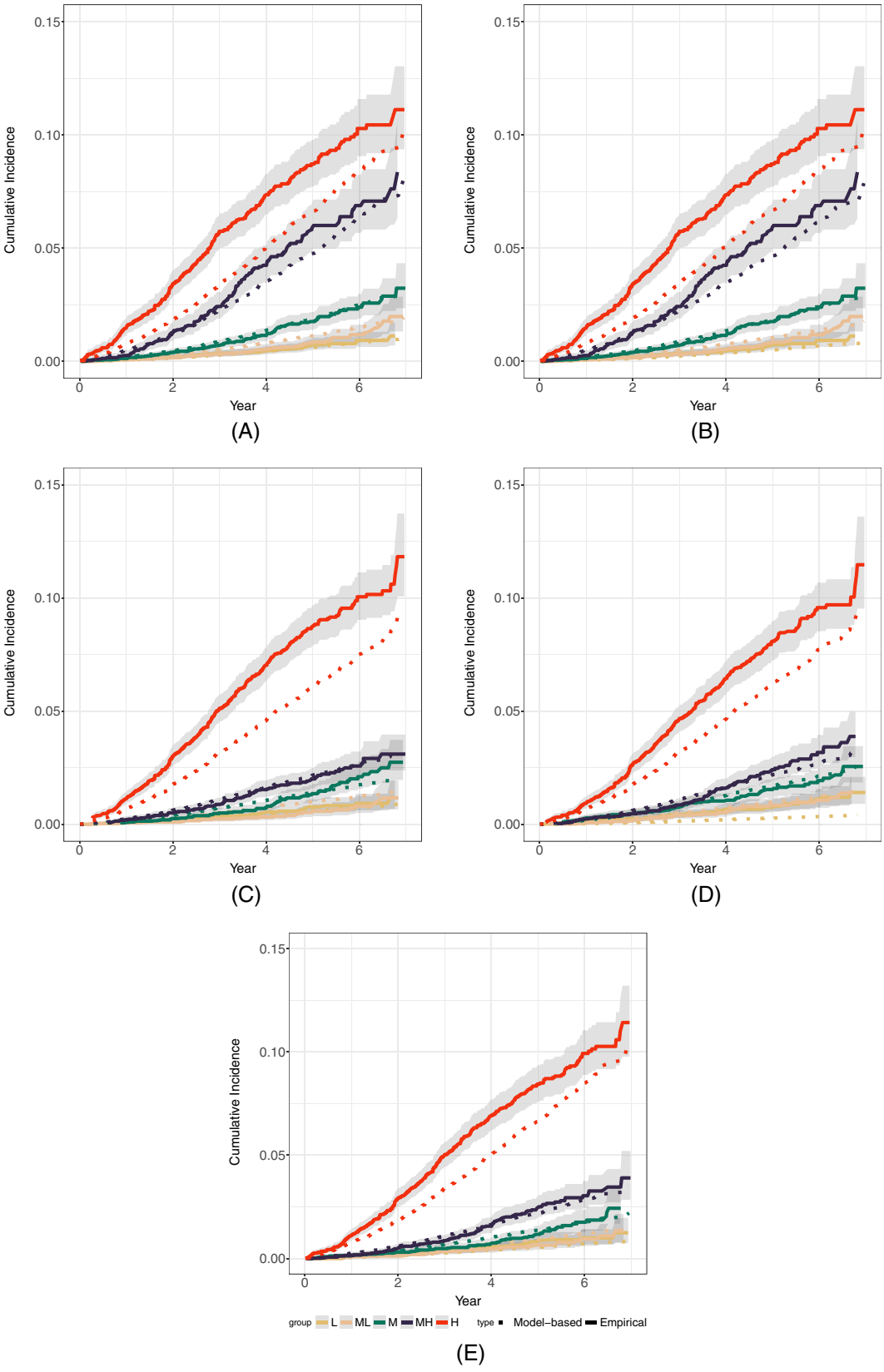


FIGURE 5 Cumulative incidence functions for cancer mortalities, with classification and prediction using A, LASSO and B, adaptive LASSO under the PCSH model and C, LASSO, D, adaptive LASSO, and E, boosting under the PSDH model. The shaded area is the 95% pointwise confidence intervals based on the nonparametric estimate. LASSO, least absolute shrinkage and selection operator; PCSH, proportional cause-specific hazards; PSDH, proportional subdistribution hazards [Colour figure can be viewed at wileyonlinelibrary.com]

In this paper, in addition to existing implementations of machine learning methods, ie, LASSO under the PCSH model and boosting under the PSDH model, we have also developed efficient algorithms for LASSO under the PSDH model and a corresponding R package is expected to be completed in the near future. We empirically studied the performance of these methods in variable selection and prediction through comprehensive simulations in both low- and high-dimensional settings with different covariate structures. From the simulation results, we see that the performance is generally better for continuous covariates than for binary ones, and further worse if the binary covariates are sparse, as in the case of claims data. This finding seems to echo a recent paper by Mukherjee et al[40] who showed that when a binary design matrix is sufficiently sparse, no signal can be detected irrespective of its strength. More work, both methodological and theoretical, appears to be needed to study binary data in high dimension, especially for sparse binary data.

In comparing these machine learning methods, adaptive LASSO had advantage over LASSO in some scenarios, but this was not universally the case especially when the covariates were binary. While it is known that good prediction does not necessarily require exact selection of the true predictors, poor performance in prediction was generally due to overselection of variables that were not true predictors in our simulation. Boosting performed well for prediction, largely due to the fact that there was not severe overselection of the variables.

For each of the machine learning methods, we compared different approaches to choose the penalty parameters, which was an important step in applying these methods. We found the 10-fold cross-validation to be generally as good as the other approaches, if not better. For LASSO-type methods, including adaptive LASSO, AIC including elbow AIC selected overly large numbers of false positive variables, resulting in poor performance for prediction purposes. For boosting however, this did not appear to be the case, and the computational ease of AIC over cross-validation seems a worthwhile advantage.

By applying the above methods to analyze a rich data set with claim codes describing disease diagnoses, surgical procedures, hospitalization, and outpatient activities, we created an individualized patient prediction tool aimed at helping prostate cancer patients and their physicians to better understand the prognosis for both cancer and other morbidities, which can in turn aid in clinical decision making. For the linked SEER-Medicare data that we have considered, the sample size is much larger than what we can afford in simulations. The main differences in prediction results for these data seem to be between the different models and less so between the regularization methods under the same model. As we have described in details, under the PCSH model, we initially had 16 strata that were then combined into 5 strata based on visual inspection, and the 5 strata were of different sizes (see tables in the Supporting Information). On the other hand, under the PSDH model, we directly divided the data into 5 equal-sized strata. How to best divide the strata including deciding the number of strata may worth future investigation. However, the purpose of dividing into strata here was mainly to illustrate that the fitted models have predictive capability. For the purposes of precision medicine, it is the individual prediction given a vector of covariates that is of most importance. We see that for noncancer mortality, both models did a reasonable job in prediction, and it shows the usefulness of claims database such as Medicare in predicting noncancer mortality. For cancer mortality, on the other hand, the clinical information captured in the SEER data might be relatively limited, or the models we considered might be partly misspecified; additional investigation including the use of individual patient electronic medical records might be able to provide more clinical information for better prediction.

Finally, our work here assumed the proportional hazards under both models. We note that there has been recent work considering other modeling approaches such as the additive hazards in the presence of competing risks.[41] Machine learning methods are still yet to be developed and studied under these models.

ACKNOWLEDGMENTS

This research was supported by a grant from the American Society of Clinical Oncology (ASCO) and was partially supported by the National Institutes of Health Clinical and Translational Science Award (CTSA)UL1TR001442.

ORCID


Ronghui Xu http://orcid.org/0000-0002-2822-0561

REFERENCES

1. Thompson I, Thrasher JB, Aus G, et al. Guideline for the management of clinically localized prostate cancer: 2007 update. *J Urol*. 2007;177:2106-2131.

2. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc*. 2001;96:1348-1360.
3. Bühlmann P, van de Geer S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin Heidelberg: Springer; 2011.
4. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. 1997;16:385-395.
5. Zhang HH, Lu W. Adaptive lasso for Cox's proportional hazards model. *Biometrika*. 2007;94:691-703.
6. Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan MJ. Survival ensembles. *Biostatistics*. 2006;7:355-373.
7. Bradic J, Fan J, Jiang J. Regularization for Cox's proportional hazards model with NP-dimensionality. *Ann Stat*. 2011;39:3092.
8. Binder H, Allignol A, Schumacher M, Beyersmann J. Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics*. 2009;25:890-896.
9. Ha ID, Lee M, Oh S, Jeong J, Sylvester R, Lee Y. Variable selection in subdistribution hazard frailty models with competing risks data. *Stat Med*. 2014;33:4590-4604.
10. Fu Z, Parikh CR, Zhou B. Penalized variable selection in competing risks regression. *Lifetime Data Anal*. 2017;23:353-376.
11. Fleming TR, Harrington DP. *Counting Processes and Survival Analysis*, Vol. 169. New York: John Wiley & Sons; 2011.
12. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*, Vol. 360. New York: John Wiley & Sons; 2011.
13. Breslow N. Covariance analysis of censored survival data. *Biometrics*. 1974;30(1):89-99.
14. Cheng S, Fine JP, Wei L. Prediction of cumulative incidence function under the proportional hazards model. *Biometrics*. 1998;54(1):219-228.
15. Hou J, Xu R. CompetingRisk: the semi-parametric cumulative incidence function. R package version 1.0. https://CRAN.R-project.org/package=CompetingRisk; 2017.
16. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc*. 1999;94:496-509.
17. Gray RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Ann Stat*. 1988;16:1141-1154.
18. Geskus RB. Cause-specific cumulative incidence estimation and the Fine-Gray model under both left truncation and right censoring. *Biometrics*. 2011;67:39-49.
19. Geskus RB. *Data Analysis with Competing Risks and Intermediate States*. Boca Raton, FL: Taylor & Francis Group, LLC; 2016.
20. Lambert PC, Wilkes SR, Crowther MJ. Flexible parametric modelling of the cause-specific cumulative incidence function. *Stat Med*. 2017;36:1429-1446.
21. Koller MT, Raatz H, Steyerberg EW, Wolbers M. Competing risks and the clinical community: irrelevance or ignorance. *Stat Med*. 2012;31:1089-1097.
22. Verweij PJ, Van Houwelingen HC. Cross-validation in survival analysis. *Stat Med*. 1993;12:2305-2314.
23. Xu R, Vaida F, Harrington DP. Using profile likelihood for semiparametric model selection with application to proportional hazards mixed models. *Stat Sin*. 2009;19:819-842.
24. Volinsky CT, Raftery AE. Bayesian information criterion for censored survival models. *Biometrics*. 2000;56:256-262.
25. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a dataset via the gap statistic. *J R Stat Soc, Ser B*. 2001;63:411-423.
26. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc. Ser B (Methodological)*. 1996;58(1):267-288.
27. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. *Ann Stat*. 2006;34(3):1436-1462.
28. Zhao Peng, Yu Bin. On model selection consistency of Lasso. *J Mach Learn Res*. 2006;7:2541-2563.
29. Zhang CH, Zhang SS. Confidence intervals for low dimensional parameters in high dimensional linear models. *J R Stat Soc, Ser B*. 2014;76:217-242.
30. van de Geer S, Bühlmann P, Ritov Y, Dezeure R. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann Stat*. 2014;42:1166-1202.
31. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101:1418-1429.
32. Freund Y, Schapire RE. A desicion-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci*. 1997;55:119-139.
33. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29:1189-1232.
34. Mallat SG, Zhang Z. Matching pursuits with time-frequency dictionaries. *IEEE Transa Signal Process*. 1993;41:3397-3415.
35. Bühlmann P. Boosting for high-dimensional linear models. *Ann Stat*. 2006;34:559-583.
36. Bühlmann P, Yu B. Boosting with the L_2 loss: regression and classification. *J Am Stat Assoc*. 2003;98:324-339.
37. Beyersmann J, Latouche A, Buchholz A, Schumacher M. Simulating competing risks data in survival analysis. *Stat Med*. 2009;28:956-971.
38. Wasserman L, Roeder K. High dimensional variable selection. *Ann Stat*. 2009;37:2178-2201.
39. Fan J, Han F, Liu H. Challenges of big data analysis. *Nat Sci Rev*. 2014;1:293-314.

40. Mukherjee R, Pillai NS, Lin X. Hypothesis testing for high-dimensional sparse binary regression. *Ann Stat*. 2015;43(1):352.

41. Zheng C, Dai R, Hari PN, Zhang M. Instrumental variable with competing risk model. *Stat Med*. 2017;36:1240-1255.

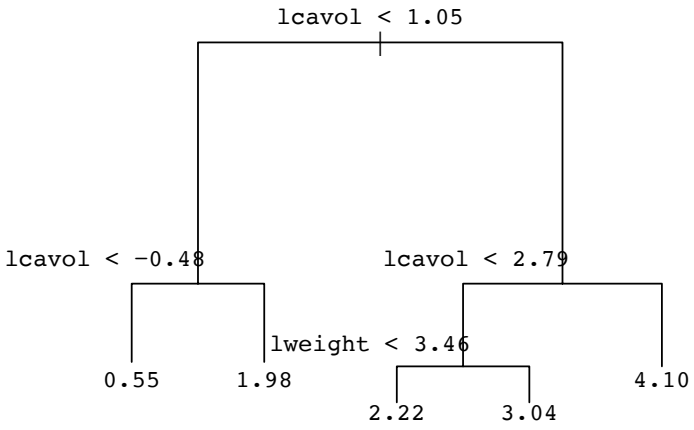
SUPPORTING INFORMATION

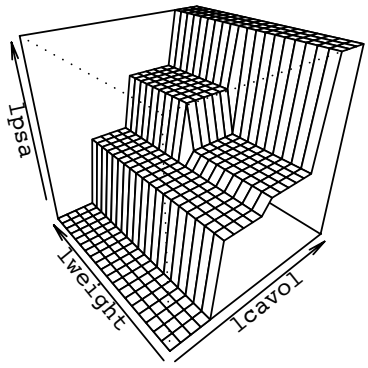
Additional supporting information may be found online in the Supporting Information section at the end of the article.

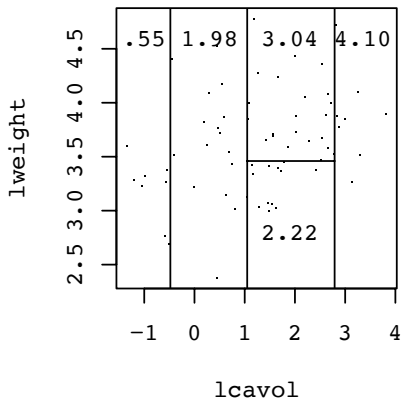
How to cite this article: Hou J, Paravati A, Hou J, Xu R, Murphy J. High-dimensional variable selection and prediction under competing risks with application to SEER-Medicare linked data. *Statistics in Medicine*. 2018;37:3486–3502. https://doi.org/10.1002/sim.7822

Tree-based Methods

What is a tree?







Growing the Tree

- Tree here refers to a *recursive partition* (eg. R package 'rpart') in the regressor/predictor space into regions called *nodes*, containing more homogeneous response values.

- The partition (i.e. split) is **binary**, one predictor at a time:

  - starting from the *root*, i.e. a single node, the tree algorithm searches through all possible splits for all predictors, to find the best split according to some criterion;

    * for a continuous predictor, the split is between every two consecutive observed values;

    * for a categorical predictor, the split is between any plausible combinations of the categories.

  - the *splitting criterion* is typically based on some loss or cost function, eg. to minimize prediction error, or to maximize a test statistic for whether an 'effect' in the **two** *daughter nodes* are equal:

    * suppose the effect of X_1 on Y is β_1 in the *left node*, and β_2 in the *right node*, $H_0 : \beta_1 = \beta_2$ (why)

    * this is related to change point detection.

  - as the search is exhaustive, it is advantageous if the criterion can be quickly updated as the search moves through the predictor space, eg. of the form:

    cumsum(daughter1) - cumsum(daughter2);

    * score test statistic mentioned before often has this property, while likelihood ratio test statistic does not (why).

- Once the root has been split into two daughter nodes, the splitting process repeats itself within each of the daughter nodes. Hence it is called recursive partitioning.

- Why binary splits?

  - easy to implement;
  - multiway splits can be achieved by a series of binary splits.

Stopping Rules

When do we stop growing the tree?

These stopping rule are relatively straightforward; for example:

1. run out of data: often each node needs to have enough sample for estimation;

2. pre-specified tree *depth* is reached;

3. certain *purity* threshold is reached, eg. the maximized score test statistic falls under the pre-specified threshold.

We tend to grow a large tree, because stopping too early we might miss a good split down the road.

CART: pruning the tree

Trees can be used for categorical or continuous outcomes, corresponding to *classification and regression trees* (CART, Breiman *et al.* 1984).

- Binary splitting is in fact not so hard for a researcher to come up with;
- Once you have a fully grown tree, it might be necessary to prune (why);
- How many possibly ways can you prune the tree?
- The real smart thing about CART, is that Breiman *et al.* (1984) figured out the following.

- Let T_0 be the full tree

- T is a *subtree* of T_0:
  - T has the same root as T_0;
  - every node of T is a node of T_0;
  - denote $T \preceq T_0$

- Let $C(T)$ be the *cost* of T, eg. error rate or $-$log likelihood

- $|T|$ denotes the number of *terminal nodes* of T

- For a given *penalty* α, the optimally pruned subtree has:

$$T^*(\alpha) = \text{argmin}_{T \preceq T_0}\{C(T) + \alpha|T|\}$$

Q: the larger the α, the larger or smaller the $T^*(\alpha)$?

- CART has the following property:

  There exists a finite sequence $0 = \alpha_1 \leq \alpha_2 \leq ... \leq \alpha_K$ such that, for $k = 1, 2, ..., K-1$,

  $$T^*(\alpha) = T^*(\alpha_k) = T_k$$

  for $\alpha_k \leq \alpha < \alpha_{k+1}$, and

  $$T^*(\alpha) = T_K \text{ the root}$$

  for $\alpha \geq \alpha_K$.

- In other words, there are only K optimally pruned subtrees we need to consider.

- See Xu and Adak (2002) for an example of optimally pruned subtrees.

Q: how to choose a subtree?

Selection of a subtree

A: based on prediction error.

- Because the splitting is *adaptive* based on data, $C(T_k)$ is smaller than it would be with pre-chosen splits; i.e. it is the 'apparent' cost as opposed to the true predictive cost.

Q: how do we correct for such bias?

Correction for optimism

- There are multiple ways:
  1. test - training sample;
  2. cross-validation;
  3. bootstrap.

- Afterwards, additional criteria such as AIC etc. may be used to choose a *final tree*.

Pros and Cons of Trees

From Hastie *et al.* (2001) book *The Elements of Statistical Learning*:

- Can capture interactions (why)
- Naturally handle both continuous and categorical predictors
- Handle missing values in the predictor variables
- Robust to outliers in the predictors
- Insensitive to monotone transformations of the predictors
- Scale well for large sample sizes
- Deal well with irrelevant predictors

  Other approaches like support vector machines (SVM) or neural networks do not have the above strengths.

- People also find the results easy to interpret in practice, than say from a regression model, eg. in medical prognosis.

- But trees are known to be unstable:
  - a small perturbation in data can change the root split;
  - and that changes the whole tree.

- Not smooth

- Not good at capturing linear combinations of predictors

- Not as accurate as some of the more recently developed methods.

Example

kyphosis {rpart} R Documentation

Data on Children who have had Corrective Spinal Surgery

Description

The kyphosis data frame has 81 rows and 4 columns. representing data on children who have had corrective spinal surgery

This data frame contains the following columns:

Kyphosis
a factor with levels absent present indicating if a kyphosis (a type of deformation) was present after the operation.

Age
in months

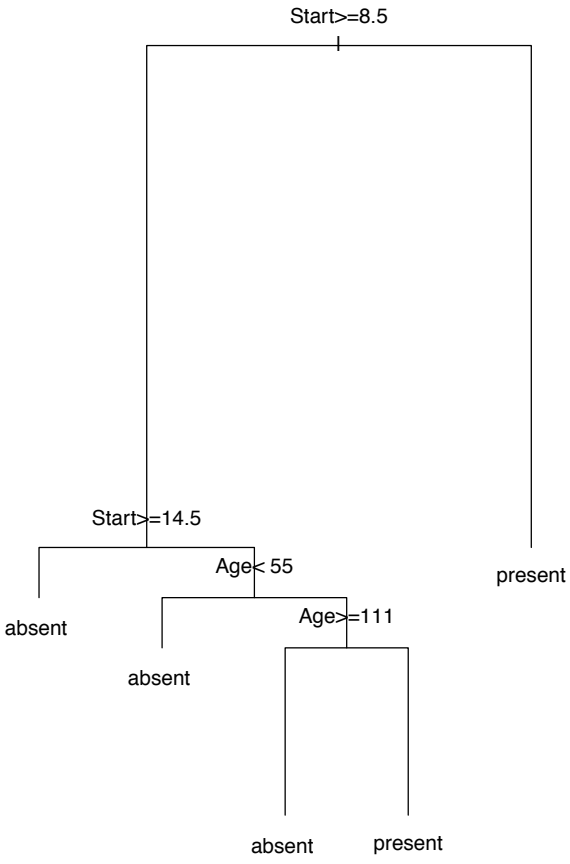
Number
the number of vertebrae involved

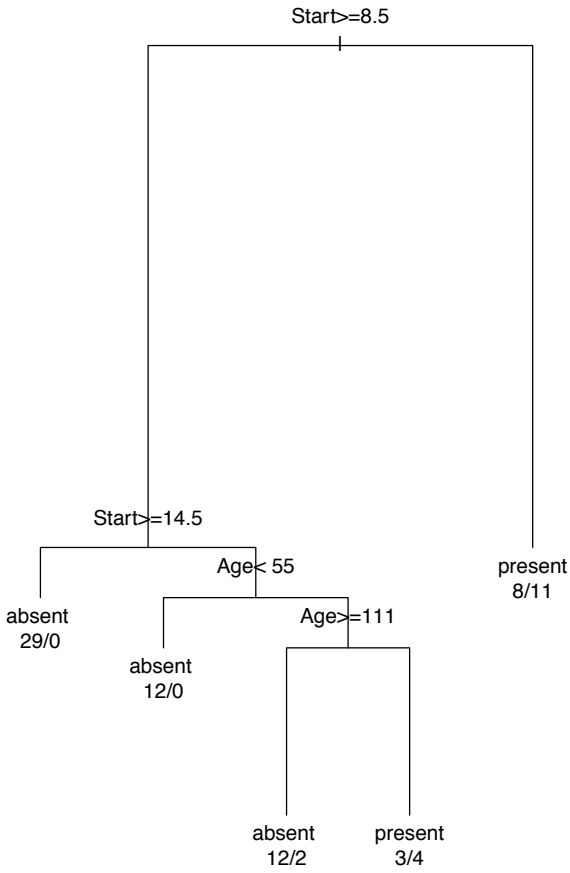
Start
the number of the first (topmost) vertebra operated on.

Source

John M. Chambers and Trevor J. Hastie eds. (1992) Statistical Models in S, Wadsworth and Brooks/Cole, Pacific Grove, CA.

```
fit <- rpart(Kyphosis ~ Age + Number + Start, data = kyphosis)
par(mfrow = c(1,2), xpd = NA)
plot(fit)
text(fit)
plot(fit)
text(fit, use.n = TRUE)
```

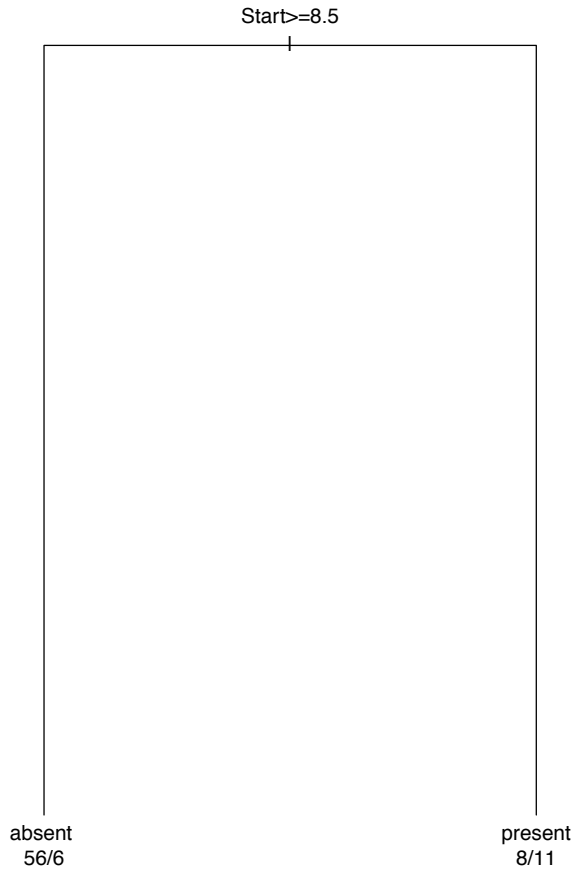


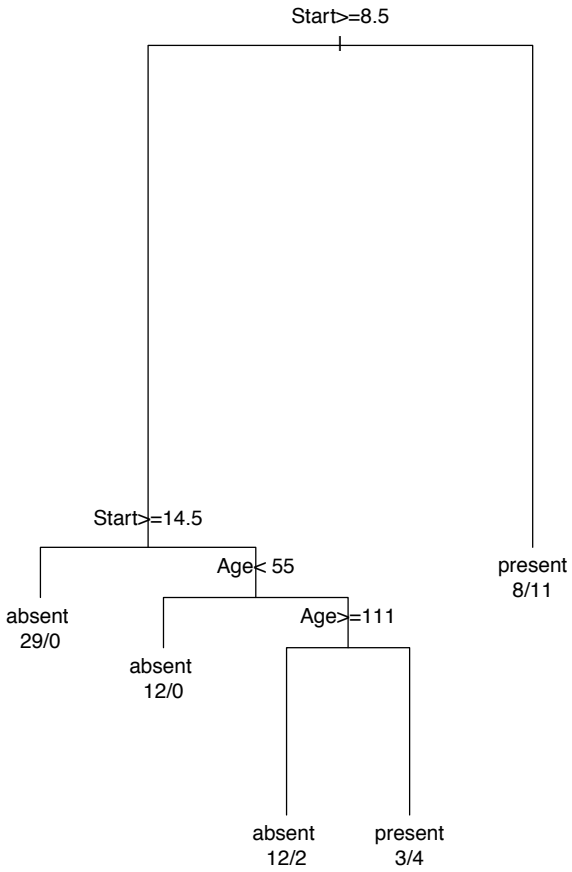


```
> fit$cptable
## a matrix of information on the optimal prunings based on
## a complexity parameter (CP), i.e. penalty.

          CP nsplit rel error    xerror      xstd
1 0.17647059      0 1.0000000 1.0000000 0.2155872
2 0.01960784      1 0.8235294 0.8823529 0.2056488
3 0.01000000      4 0.7647059 0.8823529 0.2056488
```

Q: how many optimally pruned subtrees?

```
fit1 = prune(fit, cp=0.1)
plot(fit1)
text(fit1, use.n = TRUE)
fit2 = prune(fit, cp=0.015)
plot(fit2)
text(fit2, use.n = TRUE)
```


Start>=8.5
absent
56/6
present
8/11


Start>=8.5
Start>=14.5
Age< 55
present
8/11
absent
29/0
Age>=111
absent
12/0
absent
12/2
present
3/4

```
## xpred.rpart() gives the predicted values for an rpart fit, under
## cross validation, for a set of complexity parameter values.

> cbind(kyphosis$Kyphosis, xpred.rpart(fit))

## Complexity penalties are actually ranges, not values. If the cp values found
## in the table were .36, .28, and .13, for instance, this means that the first row
## of the table holds for all complexity penalties in the range [.36, 1], the
## second row for cp in the range [.28, .36) and the third row for [.13,.28).
## By default, the geometric mean of each interval is used for cross validation.

     0.58823529 0.05882353 0.01400280
1  1          1          2          2
2  1          1          1          1
3  2          1          2          2
4  1          1          2          2
5  1          1          1          1
6  1          1          1          1
7  1          1          1          1
8  1          1          1          1
9  1          1          1          1
10 2          1          1          1
11 2          1          1          1
12 1          1          1          1
13 1          1          2          2
14 1          1          1          1
15 1          1          1          1
16 1          1          1          1
17 1          1          1          1
18 1          1          1          1
19 1          1          1          1
20 1          1          1          1
....
```