Exercises & Lecture Notes:

STK 390: Bootstrapping and Resampling Theory

Nils Lid Hjort, Spring 1990

These form a dynamically growing set of exercises and lecture notes, during my course.

Exercise No. 1

This exercise concerns the empirical distribution $\widehat{F}$ for an observed data set. $\widehat{F}$ is quite central in the theory and practice of bootstrapping techniques.

Let F be a probability distribution on $\mathcal{R}$, the real line. It is customary and convenient to use F to denote two slightly different quantities: it can be viewed as a *distribution function*, that is

$$F(t) = \Pr\{X \le t\}, \quad t \in \mathcal{R},$$

or as the accompanying *probability measure*, that is

$$F(A) = \Pr\{X \in A\}, \quad A \text{ a Borel set.}$$

(a) So F has two different connotations; why is this acceptable?

(b) Assume that independent observations $X_1, \ldots, X_n$ have been drawn from F. The most usual and indeed natural nonparametric estimator of F is *the empirical distribution* $\widehat{F}$, given as

$$\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le t\}, \quad t \in \mathcal{R}, \quad \text{or} \quad \widehat{F}(A) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \in A\}, \quad A \text{ a set.}$$

Find $\mathrm{E}_F\widehat{F}(t)$ and $\mathrm{Var}_F\widehat{F}(t)$, and comment. [Here and in the following $I\{\ldots\}$ denotes the *indicator function* for $\{\ldots\}$, so that $I\{X_i \le t\} = 1$ or 0 according to whether $X_i \le t$ or $X_i > t$. Furthermore, Var_F means 'variance evaluated under the model F', *&c.*]

(c) Find also mean value and variance (under F) for the random variable $\widehat{F}(A)$, for a fixed set A.

(d) How accurate is the estimator $\widehat{F}$ for F? One out of several possible 'overall measures' for how close $\widehat{F}$ is to F, is the distance $D(F,\widehat{F}) = \int\{\widehat{F}(t)-F(t)\}^2F(\mathrm{d}t)$. A reasonable quality measure for $\widehat{F}$ is accordingly the 'risk function'

$$R(\widehat{F},F) = \mathrm{E}_F D(F,\widehat{F}) = \mathrm{E}_F \int\{\widehat{F}(t)-F(t)\}^2F(\mathrm{d}t).$$

Show that this risk function is identical to $1/(6n)$, for all continuous F's.

(e) Remember Glivenko & Cantelli, not to mention Kolmogorov, from 1933? What can you say about $D(F,\widehat{F})$, and other distance measures, when the sample size n is large?

(f) (PONDER & ponder) Does $\widehat{F}$ have any shortcomings or inadequacies? What can constitute good alternative estimators? Are there competing methods that are 'better'? How much better can F be estimated under parametric circumstances? See Exercises 15, 16, 17.

Exercise No. 2

The following data set was artificially generated, for the sake of illustrating several important concepts:

```
1.3455, 0.3667, 0.4845, 0.7166, 0.3155,
1.0561, 1.7350, 1.1957, 1.7310, 3.6730,
0.1582, 1.9139, 0.6522.
```

The data points were actually drawn as an i.i.d. sample from the unit exponential distribution (with density $g(t) = e^{-t}$ for positive t). The point of view to be taken in the present exercise is however the nonparametric one, so the sole assumption being made is that the data set constitutes an i.i.d. sample of $n = 13$ from an unknown distribution F.

Consider the following parameter, which has been proposed as a measure of spread of the underlying distribution:

$$\theta = \theta(F) = \text{med}_F\{|X - \text{med}(F)|\}, \quad X \sim F.$$

(If $\theta = 1.377$ and μ is the median of F, then X's from F are within 1.377 of μ half of the time and more than 1.377 away from μ the other half of the time.)

(a) To get a feeling for the spread parameter θ, show that $\theta = .675\,\sigma$ in a Gaussian (μ, σ^2) distribution, and that $\theta = \log(\frac{1}{2} + \frac{1}{2}\sqrt{5})\,\lambda = .481\,\lambda$ for the exponential distribution with density $(1/\lambda)e^{-x/\lambda}$. Show also that $\theta(F) = F_0^{-1}(\frac{3}{4})\,\sigma$ when $F(x) = F_0(\frac{x-\mu}{\sigma})$ for a 'basis distribution' F_0 that is symmetric around zero.

(b) Discuss the merits of θ as an alternative to the more familiar and tradition-bound standard deviation parameter, as a measure of spread.

(c) Make a histogram of the data, and plot the empirical cumulative distribution $\widehat{F}(t)$, along with the underlying $F(t) = 1 - e^{-t}$. Use MINITAB or something else you might have available. (In particular, a nice exercise is to make up a MINITAB-macro that for given data column `c1` produces a plot of $\widehat{F}(t)$ against t.)

(d) Compute the natural estimate

$$\widehat{\theta} = \theta(\widehat{F}) = \text{med}\{|X_1 - \widehat{\mu}|, \ldots, |X_n - \widehat{\mu}|\},$$

in which $\widehat{\mu}$ is the sample median. (I got $\widehat{\theta} = 0.6749$.)

(e) Of essential practical importance is the ability of statistical methodology to complement the estimate $\widehat{\theta}$ above with a *measure of uncertainty*, say an estimate of its standard deviation, or its root mean square error. Let $\tau = \tau(F)$ be this root mean square error, i.e.

$$\begin{aligned}\tau^2 = \tau(F)^2 &= \text{E}_F\{\theta(\widehat{F}) - \theta(F)\}^2 \\ &= \text{E}_F\{\widehat{\theta}(X_1, \ldots, X_n) - \theta(F)\}^2 \\ &= \int \{\widehat{\theta}(x_1, \ldots, x_n) - \theta(F)\}^2 \,\text{d}F(x_1) \cdots \text{d}F(x_n).\end{aligned}$$

This is certainly a complicated functional, since $\widehat{\theta}$ is so complicated, and a closed form expression seems unattainable. We can nevertheless consider

$$\begin{aligned}\widehat{\tau}^2 = \tau(\widehat{F})^2 &= \mathrm{E}_{\widehat{F}}\{\theta(\widehat{\widehat{F}}) - \theta(\widehat{F})\}^2 \\ &= \mathrm{E}_{\widehat{F}}\{\widehat{\theta}(X_1^*, \ldots, X_n^*) - \theta(\widehat{F})\}^2 \\ &= \int \{\widehat{\theta}(x_1^*, \ldots, x_n^*) - \theta(\widehat{F})\}^2 \, \mathrm{d}\widehat{F}(x_1^*) \cdots \mathrm{d}\widehat{F}(x_n^*).\end{aligned}$$

$X_1^*, \ldots, X_n^*$ in the next-to-last expression are an i.i.d. sample from $\widehat{F}$, i.e. randomly drawn, with replacement, from the original data points $\{X_1, \ldots, X_n\}$. — Explain why $\widehat{\tau}$ is an explicit estimator, and that it can be computed, in principle, as a sum over n^n terms.

(f) From what n on would you say such a computational procedure, evaluating your estimate as a sum of n^n terms, is prohibitive?

(g) But there is another numerical scheme to compute $\widehat{\tau}$: $\widehat{\tau}^2 = \tau(\widehat{F})^2 = \mathrm{E}_* Z^*$, say, where

$$Z^* = (\widehat{\theta}^* - \widehat{\theta})^2 = \{\widehat{\theta}(X_1^*, \ldots, X_n^*) - \theta(\widehat{F})\}^2$$

is a random variable which can be simulated easily, e.g. with the help of a simple MINITAB-macro. Obtain a (large) number boot of independent values $Z_1^*, \ldots, Z_{\mathrm{boot}}^*$, with

$$Z_b^* = (\widehat{\theta}^{*b} - \widehat{\theta})^2 = \{\widehat{\theta}(X_1^{*b}, \ldots, X_n^{*b}) - \theta(\widehat{F})\}^2 = \{\widehat{\theta}(X_1^{*b}, \ldots, X_n^{*b}) - 0.6749\}^2,$$

and use $\widehat{\tau}^2 \approx \frac{1}{\mathrm{boot}} \sum_{b=1}^{\mathrm{boot}} Z_b^*$ (an approximation guaranteed by Kolmogorov (1903–1987) and his law of large numbers). Do this in the present situation, with *boot* equal to 20, 100, and 1000. — Here $\{X_1^{*b}, \ldots, X_n^{*b}\}$ constitutes bootstrap sample no. b, and is an i.i.d. sample, with replacement, from the original data points. '$\mathrm{E}_*\{\ldots\}$' signals mathematical expectation of $\{\ldots\}$ within the bootstrap framework, in which the X_i^*'s are i.i.d. from $\widehat{F}$. In particular, $\mathrm{E}_*\{\ldots\}$ refers to a stochastic framework entirely in the hands of the statistician and her electronical computer, in which the data values are given and fixed, as opposed to the 'outer' statistical model from which the data points were generated.

Exercise No. 3

The point to be made now is that the *bias* and *median-bias* of a given estimator $\widehat{\theta} = \theta(\widehat{F})$ can be estimated (and later on corrected for), for a given set of data values.

Let

$$\beta = \beta(F) = \mathrm{E}_F \widehat{\theta} - \theta(F) = \mathrm{E}_F \widehat{\theta}(X_1, \ldots, X_n) - \theta(F)$$

be the bias (w.r.t. expectation) of $\widehat{\theta}$. For concreteness let the $\theta = \theta(F)$ functional be as in the previous exercise. Then no closed form expression for the bias can be found. It can still be estimated, however, in a bootstrap way: let

$$\widehat{\beta} = \beta(\widehat{F}) = \mathrm{E}_{\widehat{F}} \widehat{\theta}^* - \theta(\widehat{F}) = \mathrm{E}_* W^*,$$

say, where $W^* = \widehat{\theta}^* - \widehat{\theta}$ and $\widehat{\theta}^* = \widehat{\theta}(X_1^*, \ldots, X_n^*)$. Obtain a number boot of bootstrap estimates $\widehat{\theta}^*$ from your computer, resulting in independently simulated values $W_1^*, \ldots, W_{\rm boot}^*$, and let their average constitute your numerical approximation to $\widehat{\beta}$.

(a) Do this for boot = 20, 100, 1000, in the situation of the previous exercise, and compute along the way the bias-corrected estimate $\widehat{\theta}_{\rm BC} = \widehat{\theta} - \widehat{\beta}$, which purports to have zero bias.

(b) Statisticians aiming at a track record with the admirable property that they overestimate parameters about as often as they underestimate them, need *median-unbiased* estimators. Let

$$\gamma = \gamma(F) = \text{med}_F \widehat{\theta} - \theta(F)$$

be the median-bias, subsequently to be estimated and removed. Devise a bootstrap way of doing this, and do it! in the situation of the previous exercise. Compare your median-corrected estimate $\widehat{\theta}_{\rm MC}$ with the expectation-corrected $\widehat{\theta}_{\rm BC}$.

> *What we have to learn to do we learn by doing.*
>
> — Aristotle, Ethica Nicomachea II (c. 325 B.C.)

Exercise No. 4

The bootstrap computations can be considered as simple numerical devices to evaluate estimates of the empirical-functional type $\tau(\widehat{F})$. For a small number of simple functionals these numbers can be computed explicitly, without resampling strategies. Consider in particular the familiar parameters

$$\theta = \theta(F) = \text{E}_F X_i = \int x \, \text{d}F(x),$$

$$\sigma^2 = \sigma(F)^2 = \text{Var}_F X_i = \int \{x - \theta(F)\}^2 \, \text{d}F(x).$$

(a) Find explicit (and familiar!) expressions for $\widehat{\theta} = \theta(\widehat{F})$ and $\widehat{\sigma} = \sigma(\widehat{F})$.

(b) Let $\beta_1 = \beta_1(F)$ be the bias for $\widehat{\theta}$, and $\beta_2 = \beta_2(F)$ the bias for $\widehat{\sigma}^2$. Give expressions for β_1 and β_2, and for the natural estimates $\widehat{\beta}_1 = \beta_1(\widehat{F})$ and $\widehat{\beta}_2 = \beta_2(\widehat{F})$. Show that the bootstrap scheme, involving bootstrap samples $X_1^*, \ldots, X_n^*$, if applied here, leads to the same results! What are the resulting bias-corrected estimators for θ and σ^2?

(c) Next consider

$$\tau_1 = \tau_1(F) = \text{stdev}_F(\widehat{\theta}) = \sqrt{\text{Var}_F \, \widehat{\theta}}, \quad \tau_2 = \tau_2(F) = \text{stdev}_F(\widehat{\sigma}^2) = \sqrt{\text{Var}_F \, \widehat{\sigma}^2}.$$

Find expressions for $\widehat{\tau}_1 = \tau_1(\widehat{F})$ and $\widehat{\tau}_2 = \tau_2(\widehat{F})$, and show again that the bootstrap scheme also leads to these expressions.

Exercise No. 5

Let $x_1, \ldots, x_n$ be the original data set, supposed to be realisations of random variables $X_1, \ldots, X_n$ that were i.i.d. $\sim F$, and suppose that there are no ties in the data. Let $X_1^*, \ldots, X_n^*$ be a bootstrap sample, i.e. they are i.i.d. $\sim \widehat{F}$, the empirical distribution.

(a) Find

$$\mathrm{E}_* X_i^*, \quad \mathrm{Var}_* X_i^*, \quad \mathrm{E}_* \frac{1}{n}\sum_{i=1}^n X_i^*, \quad \mathrm{Var}_* \frac{1}{n}\sum_{i=1}^n X_i^*,$$

in which the subscript $_*$ again refers to the bootstrap framework, conditional on the data points.

(b) Letting $\bar{X}^* = \frac{1}{n}\sum_{i=1}^n X_i^*$, find

$$\mathrm{E}\{\mathrm{E}_* \bar{X}^*\}, \quad \mathrm{E}\{\mathrm{Var}_* \bar{X}^*\}, \quad \mathrm{Var}\{\mathrm{E}_* \bar{X}^*\}.$$

(c) Finally find

$$\mathrm{E}\, X_i^*, \quad \mathrm{Var}\, X_i^*, \quad \mathrm{E}\bar{X}^*, \quad \text{and} \quad \mathrm{Var}\, \bar{X}^*.$$

Exercise No. 6

Let $X_1^*, \ldots, X_n^*$ be a bootstrap sample, as in the previous exercise. Let M be the number of the original data points x_i that manage to escape the looming bootstrap, i.e. $M = \sum_{i=1}^n M_i$, where M_i is indicator for $\{x_i$ is not in the bootstrap sample$\}$.

(a) Show that $\mathrm{E}_* M = np_n$, where $p_n = (1 - \frac{1}{n})^n \doteq e^{-1} = .368$. Accordingly, an average bootstrap sample includes only 63.2% of the original data points.

(b) Why is M not binomial (n, p_n)? Show that

$$\mathrm{Var}_* \left[(M - np_n)/\sqrt{n}\right] \to e^{-1}(1 - e^{-1}) - e^{-2}$$

as n grows. In particular the distribution of M is distinctively different from the binomial, even asymptotically. Can you supply a limit distribution result?

Exercise No. 7

The previous exercise showed that in an average bootstrap sample, about 36.8% of the original data points will not be included, so $X_1^*, \ldots, X_n^*$ will most probably contain replicates of some of those that are included. Let N_i be the number of replicates of sample point x_i in the bootstrap sample.

(a) Show that $N_i \sim \mathrm{Bin}(n; \frac{1}{n})$, and put down $\mathrm{E}_* N_i$ and $\mathrm{Var}_* N_i$.

(b) More generally, convince yourself that $\mathbf{N} = (N_1, \ldots, N_n)'$ is multinomial $(n; \frac{1}{n}, \ldots, \frac{1}{n})$. What is $\mathrm{cov}_*(N_i, N_j)$?

(c) Let $\mathbf{P}^0 = (\frac{1}{n}, \ldots, \frac{1}{n})'$ and $\mathbf{P}^* = (P_1^*, \ldots, P_n^*) = (N_1/n, \ldots, N_n/n)'$. We might call $\mathbf{P}^*$ a *(bootstrap) resampling vector*. Note that $\mathrm{E}_* \mathbf{P}^* = \mathbf{P}^0$. Show that $\mathrm{VAR}_* \mathbf{P}^* = \frac{1}{n^2}[I - \frac{1}{n}ee']$, where $e = (1, \ldots, 1)'$.

(d) How far is $\{X_1^*, \ldots, X_n^*\}$ from being equal to the original data set $\{x_1, \ldots, x_n\}$? This question matters for parts of the theoretical support behind the bootstrap method, and can be rephrased to questions involving the distance $\|\mathbf{P}^* - \mathbf{P}^0\|$, the square of which is $\sum_{i=1}^n (N_i/n - 1/n)^2$. Compute the expected squared distance. Show that the distance $\|\mathbf{P}^* - \mathbf{P}^0\|$ is $O_p^*(1/\sqrt{n})$. Show, on the other hand, that the corresponding

distance $\|\mathbf{P}_{(i)}-\mathbf{P}^0\|$ is $O(1/n)$, i.e. much smaller, for jackknife resampling vectors $\mathbf{P}_{(i)} = (\frac{1}{n-1},\ldots,0,\ldots,\frac{1}{n-1})$. In a sense the jackknife estimates $\widehat{F}_{(i)}$ of F are too close to the simple $\widehat{F}$, while the random bootstrap estimates $\widehat{F}^*$ come sufficiently far away to display the genuine uncertainty.

(e) Devise a MINITAB-macro that produces outcomes of $\mathbf{N} = (N_1,\ldots,N_n)$, and have a look.

Exercise No. 8

Let $\widehat{\theta} = \widehat{\theta}_n(X_1,\ldots,X_n)$ estimate $\theta = \theta(F)$, where the $\widehat{\theta}_n$-function is defined for every n, and is symmetric in its n arguments. Write

$$\mathrm{E}_F\widehat{\theta} = \theta(F) + b(F),$$

i.e. $b(F)$ is the *bias* of the estimator. Statisticians have devised several general schemes that intend to estimate the bias, based directly on the data, producing $\widehat{b}$, say, so that a new and hopefully bias-corrected estimate $\widetilde{\theta} = \widehat{\theta} - \widehat{b}$ can be put forward. One scheme is the bootstrap one, and another, historically preceding the bootstrap, is the jackknife method.

Define $\widehat{\theta}_{(i)} = \widehat{\theta}_{n-1}(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)$ and $\widehat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^n \widehat{\theta}_{(i)}$. The *jackknife estimator for bias* is $\widehat{b}_{\mathrm{JACK}} = (n-1)(\widehat{\theta}_{(\cdot)} - \widehat{\theta})$. The accompanying *jackknife (bias-corrected) estimate* for θ becomes

$$\widehat{\theta}_{\mathrm{JACK}} = n\widehat{\theta} - (n-1)\widehat{\theta}_{(\cdot)}.$$

It is important to observe that $\widehat{\theta}$ does not need to be of the functional estimator form $\widehat{\theta} = \theta(\widehat{F})$ here (but one needs $\theta = \theta(F)$ in order to define the bias $b(F) = b_n(F) = \mathrm{E}_F\{\widehat{\theta}_n - \theta(F)\}$ properly).

The great practical advantage is of course the generality of the proposed method; it can be put to use even in situations where no closed form expression can be derived for $\widehat{\theta}_{(\cdot)}$. It is instructive to find such expressions in not-so-complex situations, however. Find explicit formulae for the jackknife estimator, and for the jackknife estimate of bias, in the following situations, and find out, if possible, whether the resulting procedure really succeeds in getting smaller bias than the original estimator:

(a) $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$, an estimator for $\theta = \mathrm{E}_F X_i = \int x\,\mathrm{d}F(x)$.

(b) $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$, an estimator for $\sigma^2 = \mathrm{Var}_F\, X_i = \int\{x-\theta(F)\}^2\,\mathrm{d}F(x)$.

(c) $\widehat{\gamma} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^3$, an estimator for $\gamma = \mathrm{E}_F\{X_i - \theta(F)\}^3$.

(d) $\widehat{\mu} = \mathrm{median}\{X_1,\ldots,X_n\}$, an estimator for $\mu = \mathrm{median}(F) = F^{-1}(\frac{1}{2})$.

(e) $\widehat{\tau}$ = upper quartile − lower quartile, an estimator for the spread parameter $\tau = F^{-1}(\frac{3}{4}) - F^{-1}(\frac{1}{4})$. [To do this particular sub-exercise properly, one needs to decide on which order statistic to use, or what combination of which two order statistics, to estimate $F^{-1}(p)$. There is no universal agreement on this issue in the statistical community. One reasonable argument is the following: One has $\mathrm{E}_F F(X_{(i)}) = \frac{i}{n+1}$ (prove it!), so that $X_{(i)} \doteq F^{-1}(\frac{i}{n+1})$. Let $i = i(p)$ be the smallest integer $\geq (n+1)p$, so that $\frac{i}{n+1} = p + \varepsilon$ and $\frac{i-1}{n+1} = p - (\frac{1}{n+1} - \varepsilon)$, with $0 \leq \varepsilon < \frac{1}{n+1}$. Now show that the linear combination $cX_{(i-1)} + (1-c)X_{(i)}$ becomes approximately unbiased for the

sought-after $F^{-1}(p)$, with the choice $c = i - (n+1)p$, $1 - c = (n+1)p - (i-1)$. — Using this strategy, what are the estimates for τ based on a 100-sample, a 99-sample, and a 98-sample?] In particular, find expressions for $\widehat{\tau}$, $\widehat{\tau}_{(\cdot)}$, $\widehat{b}_{\text{JACK}}$, and $\widehat{\tau}_{\text{JACK}}$, based on an ordered 99-sample $X_{(1)}, \ldots, X_{(99)}$.

(f) Your own choice.

Exercise No. 9

Consider the general framework of the previous exercise, involving an estimator $\widehat{\theta} = \widehat{\theta}_n$ for a parameter θ. Assume that

$$E_n = \mathrm{E}_F \widehat{\theta}_n(X_1, \ldots, X_n) = \theta(F) + \frac{c_1(F)}{n} + \frac{c_2(F)}{n^2} + \frac{c_3(F)}{n^3} + \cdots;$$

in particular, $\widehat{\theta}$ has bias of the order $O(1/n)$. There are many examples of this form for the bias; show, for example, that $\widehat{\gamma} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^3$ has expectation $(1 - \frac{3}{n} + \frac{2}{n^2})\gamma$, where $\gamma = \mathrm{E}_F\{X_i - \mathrm{E}_F(X_i)\}^3$.

(a) Show that the bias-corrected jackknife estimator $\widehat{\theta}_{\text{JACK}} = n\widehat{\theta} - (n-1)\widehat{\theta}_{(\cdot)}$ has

$$\mathrm{E}_F \widehat{\theta}_{\text{JACK}} = \theta(F) - \frac{c_2(F)}{n(n-1)} - c_3(F)\Big[\frac{1}{(n-1)^2} - \frac{1}{n^2}\Big] - \cdots,$$

i.e. the bias has been reduced to order $O(1/n^2)$.

(b) It is possible to go one step further: Let

$$\widehat{\theta}_{(i,j)} = \widehat{\theta}_{n-2}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n),$$

where both X_i and X_j have been removed from the data set, and let $\widehat{\theta}_{(\cdot,\cdot)} = \sum_{i<j} \widehat{\theta}_{(i,j)} / \binom{n}{2}$ be their average. Devise a 'double jackknife' estimator of the form

$$\widehat{\theta}_{\text{DOUBLEJACK}} = a\widehat{\theta} + b\widehat{\theta}_{(\cdot)} + c\widehat{\theta}_{(\cdot,\cdot)}$$

that has bias of order $O(1/n^3)$! [ANSWER: $a = \frac{1}{2}n^2$, $b = -(n-1)^2$, $c = \frac{1}{2}(n-2)^2$.] Write down the first couple of terms in the bias expansion.

(c) Find an expression for the double jakknife estimator whose point of departure is $\widehat{\gamma} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^3$. Does it succeed in lowering the bias?

(d) Too laborious de-biasing can quickly lead to overkilling, however; correcting too heavily for bias can lead to too much to pay in variance. Construct an example and carry out a small simulation study to see for yourself.

Exercise No. 10

In the best spirit & tradition of Icelandic and Hebrew, construct or invent proper and fitting Norwegian terms for 'jackknife' and 'bootstrap'. Bradley Efron, in the historical début paper for the bootstrap (Bootstrap methods: another look at the jackknife, *Annals of Statistics* 1979), graciously put forward alternative terms for the method, including

'Shotgun', 'Swan-Dive', 'Swiss Army Knife', 'Jack-Rabbit', and 'Meat Axe'. (And what about 'Jack of Diamonds', 'Slaughterknife Five', and 'Jack the Crusher'?) – Among recent proposals are 'kjøttøks', 'kjøttkvern', and 'Food Processor' *(sic)*.

A Euro-cultural case in point is *Freiherr Baron Karl Friedrich Hieronymus von Münchhausen* (1720–1797), who [as we all remember] managed to save both himself and his horse from sinking into a quagmire by pulling his pigtail.

> *Når ein skal studera det norske ordfanget, må ein fyrst og fremst ha kjennskap til det levande bruket av ordi og til det miljøet der ordi vert nytta. Og nemningsbruket må granskast i samband med etterrøkjingar um dei ting, skikkar, truer og fyrestellingar som det høyrer i hop med. — Ein skynar ikkje eit mål utan at ein kjenner levevilkåri til det folket som nyttar dette målet. Heller ikkje kann ein til fullnads skyna trudom og seder hjå eit folk um ein ikkje kjenner målet til det same folket.*
>
> — Nils Lid, Norske Slakteskikkar (1924)

Exercise No. 11

Hjort (1986, *Annals of Statistics*) derived, rather in passing, and following Bayesian nonparametric considerations, the following estimator for the median $\theta = F^{-1}(\frac{1}{2})$ of a continuous distribution:

$$\widehat{\theta} = \sum_{i=1}^{n} \binom{n-1}{i-1} (\tfrac{1}{2})^{n-1} X_{(i)},$$

in which $X_{(1)} \le \ldots \le X_{(n)}$ are the order statistics. Of course it cannot possibly be unbiased; —

(a) Does there exist any unbiased estimator for the median at all?

— so there is some interest in estimating its bias, based on the data points themselves, for subsequent removal, as in $\widetilde{\theta} = \widehat{\theta} - \widehat{\text{bias}}$. The present exercise looks into two methods for estimating the bias

$$b(F) = b_n(F) = \mathrm{E}_F \widehat{\theta} - \theta(F),$$

namely the jackknife method and the bootstrap method.

(b) Show that, with the usual jackknife notation,

$$\widehat{\theta}_{(\cdot)} = \sum_{j=1}^{n} \binom{n-1}{j-1} (\tfrac{1}{2})^{n-1} 2\frac{n-1}{n} \Big[\Big(\frac{j-1}{n-1} \Big)^2 + \Big(\frac{n-j}{n-1} \Big)^2 \Big] X_{(j)}.$$

Put up an expression for $\widehat{b}_{\mathrm{JACK}}$.

(c) Describe the bootstrap procedure that leads to $\widehat{b}_{\mathrm{boot}}$.

(d) Assume now that $F = F_0$ is the unit exponential distribution, $F_0(t) = 1 - e^{-t}$ for positive t, and take $n = 17$, for concreteness. Under the exponential model it holds that $X_{(i)}$ has expected value $\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{n+1-i}$. Compute the bias $b = b(F_0)$, the parameter to be estimated. Compute the expected values of $\widehat{\theta}$, $\widehat{\theta}_{(\cdot)}$, $\widehat{b}_{\mathrm{JACK}}$, and $\widehat{\theta}_{\mathrm{JACK}}$.

(e) Carry out a simulation experiment, consisting of 100 sets of outcomes $X_1, \ldots, X_{17}$, look at the distribution of the 100 realisations of $\widehat{b}_{\mathrm{JACK}}$, and in particular, assess its mean, standard deviation, and coefficient of variation (standard deviation divided by mean). (One could presumably do the theoretical calculations, with a lot of algebraic effort, for example to find $\mathrm{Var}\,\widehat{b}_{\mathrm{JACK}}$, but doing simulations is simply cheaper, at NKr. 550 an hour.)

(f) For each of the 100 simulated sets $\{X_1, \ldots, X_{17}\}$, carry out 250 bootstrap simulations $\{X_1^*, \ldots, X_{17}^*\}$ in order to arrive at a value of $\widehat{b}_{\mathrm{boot}}$. Look at the distribution of these 100 values, and assess its mean, standard deviation, and coefficient of variation. Reach a conclusion: which one of the two methods for bias estimation performed best, for $F = F_0$ and $n = 17$?

(g) And: which one of the three estimators $\widehat{\theta}$, $\widehat{\theta}_{\mathrm{JACK}} = \widehat{\theta} - \widehat{b}_{\mathrm{JACK}}$, $\widehat{\theta}_{\mathrm{boot}} = \widehat{\theta} - \widehat{b}_{\mathrm{boot}}$ for the median performed best, for $F = F_0$ and $n = 17$?

Exercise No. 12

The jackknife machinery provides not only an estimate of the bias, for any given parameter estimator, but also an estimate of its variance:

$$\widehat{\mathrm{Var}}_{\mathrm{JACK}} = \frac{n-1}{n} \sum_{i=1}^{n} \Big[\widehat{\theta}_{(i)} - \widehat{\theta}_{(\cdot)}\Big]^2.$$

This estimate intends to be close to $\mathrm{Var}_F \widehat{\theta}$ (as opposed, for example, to $\mathrm{Var}_F \widehat{\theta}_{(\cdot)}$).

(a) Find $\widehat{\mathrm{Var}}_{\mathrm{JACK}}$ when $\widehat{\theta} = \bar{X}$. Comment on the result.

(b) Find similarly $\widehat{\mathrm{Var}}_{\mathrm{JACK}}$ for the variance estimator $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, and comment.

Exercise No. 13

Consider once more the jackknife apparatus, the point of departure for which is a given estimator $\widehat{\theta}$. We have seen how the bias and the variance of $\widehat{\theta}$ can be estimated using jackknife values $\widehat{\theta}_{(1)}, \ldots, \widehat{\theta}_{(n)}$. But what about the jackknife estimate of the variance of the jackknife-bias-corrected estimator for θ? That is, consider $\widehat{\theta}_{\mathrm{JACK}} = n\widehat{\theta} - (n-1)\widehat{\theta}_{(\cdot)}$ as the basic estimator, and find the jackknife variance estimator for this estimator, by first finding $\widehat{\theta}_{\mathrm{JACK},(\mathrm{i})}$ and so on.

Exercise No. 14

This exercise provides an asymptotic justification for the jackknife estimator of variance, for a certain class of estimators. This is the reasonably large class of estimators that are smooth functions of averages or that can be approximated by such.

Let $X_1, \ldots, X_n$ be i.i.d. from F, perhaps in a higher-dimensional space, and assume that $\widehat{\theta} = h(\bar{A}_n, \bar{B}_n)$, where $\bar{A}_n$ and $\bar{B}_n$ are averages of respectively A_i's and B_i's, and where A_i and B_i are functions of X_i. The classical large-sample solution to the problem of assessing the variability of $\widehat{\theta}$ is the *delta method*, essentially based on a first order Taylor series approximation. It works as follows:

Observe first that $\sqrt{n}(\bar{A}_n - a, \bar{B}_n - b)'$ converges in distribution to $(M, N)'$, say, by the central limit theorem, a Gaussian two-vector with means zero and a covariance matrix Σ, the elements of which are $\sigma_1^2 = \operatorname{Var} A_i$, $\sigma_2^2 = \operatorname{Var} B_i$, and $\sigma_{12} = \operatorname{cov}(A_i, B_i)$. Also, $a = \mathrm{E}\, A_i$ and $b = \mathrm{E}\, B_i$. It then follows that

$$\sqrt{n}(\widehat{\theta} - \theta) \to_d N(0, \kappa^2),$$

where

$$\kappa^2 = Dh(a,b)'\, \Sigma\, Dh(a,b) = \big[\tfrac{\partial h}{\partial a}(a,b)\big]^2 \sigma_1^2 + \big[\tfrac{\partial h}{\partial b}(a,b)\big]^2 \sigma_2^2 + 2\, \tfrac{\partial h}{\partial a}(a,b)\, \tfrac{\partial h}{\partial b}(a,b)\, \sigma_{12}.$$

$Dh(a,b)$ here is the 2-row vector containing the partial derivatives of h, evaluated at the point (a,b).

(a) The delta method approximation to $\operatorname{Var}\widehat{\theta}$ is κ^2/n, which can be estimated by

$$\widehat{\kappa}^2/n = Dh(\bar{A}_n, \bar{B}_n)'\, \widehat{\Sigma}\, Dh(\bar{A}_n, \bar{B}_n).$$

Write down an explicit expression.

(b) Show that

$$\widehat{\theta}_{(i)} \doteq \widehat{\theta} - \frac{\partial h}{\partial a}(\bar{A}_n, \bar{B}_n)\frac{A_i - \bar{A}_n}{n-1} - \frac{\partial h}{\partial b}(\bar{A}_n, \bar{B}_n)\frac{B_i - \bar{B}_n}{n-1}.$$

(c) Give an expression for $\widehat{\operatorname{Var}}_{\mathrm{JACK}}$, and show that it coincides with the variance estimate arrived at by the delta method.

(d) Do things over again, but explicitly, in the following case: Pairs (X_i, Y_i) are i.i.d., and $\mathrm{E}\, X_i = \mu_1$, $\mathrm{E}\, Y_i = \mu_2$. The parameter $\theta = \mu_1 \exp(\mu_2)$ is of interest, and the natural estimator is $\widehat{\theta} = \bar{X}_n \exp(\bar{Y}_n)$.

(e) And do them over again, but in a more general and compactly-written way, with q averages instead of two, and perhaps with a p-dimensional θ-parameter.

Exercise No. 15

The ordinary, nonparametric bootstrap is based on drawing bootstrap samples $X_1^*, \ldots, X_n^*$ from the empirical distribution $\widehat{F}$. The success of the subsequent bootstrap analysis is critically dependent on the quality of the nonparametric estimate $\widehat{F}$ for F. About how much better can one fare in parametric waters?

If $Z_n(t) = \sqrt{n}\{\widehat{F}(t) - F(t)\}$, then Z_n converges in distribution to the process Z, where $Z(t) = W^0\{F(t)\}$ and $W^0(.)$ is the *Brownian bridge*; it is Gaussian, $W^0(0) = W^0(1) = 0$, has $\mathrm{E}\, W^0(u) = 0$, and $\operatorname{cov}\{W^0(u), W^0(v)\} = u(1-v)$ for $u \le v$. The convergence in question takes place in the space $D[-\infty, \infty]$ of all functions on the line that are right continuous with left hand limits and which possesses limits at $+\infty$ and $-\infty$, equipped by the Skorohod metric, see e.g. Billingsley (1968). It ensures in particular that $g(Z_n)$ converges in distribution to $g(Z)$ for every continuous functional g.

(a) Show that

$$\sqrt{n}\|\widehat{F} - F\| = \sqrt{n} \max_{t \in \mathcal{R}} |\widehat{F}(t) - F(t)| \to_d \|W^0\| = \max_{0 \le u \le 1} |W^0(u)|,$$

the distribution of which is tabulated in standard collections. This result gives rise to the celebrated *Kolmogorov-Smirnov confidence band* for the unknown F, and one can test hypotheses of the type $F = F_0$. The point to make presently is that it gives precise information about how well $\widehat{F}$ estimates F, in terms of the maximal error

$$D(F, \widehat{F}) = \|\widehat{F} - F\| = \max_{t \in \mathcal{R}} |\widehat{F}(t) - F(t)|.$$

(b) The limit distribution above can be given in the form

$$\lim_{n\to\infty} \Pr_F\{\sqrt{n} \max_{t\in\mathcal{R}} \|\widehat{F} - F\| \le y\} = 1 - 2\sum_{j=1}^{\infty} (-1)^{j-1} e^{-2j^2y^2}, \quad y \ge 0.$$

Derivations can be found in Billingsley (1968, Section 11) and in Hájek and Šidák (1967, pp. 199–200). Owen's *Handbook of Statistical Tables* (1962, pp. 439–440) has a wrong formula but a correct table of probabilities. — Why hasn't anybody told me that

$$\mathrm{E}\,\|W^0\| = \sqrt{\pi/2}\,\log 2, \quad \mathrm{E}\,\|W^0\|^2 = \frac{\pi^2}{12}\,?$$

Show this, and conclude that the mean and standard deviation are 0.8687 and 0.2605 respectively. The median, by interpolation in Owen's table, is 0.8267. [HINTS: $\mathrm{E}\,Y = \int_0^\infty \{1 - G(y)\}\,\mathrm{d}y$ and $\mathrm{E}\,Y^2 = \int_0^\infty \{1 - G(\sqrt{y})\}\,\mathrm{d}y$ for non-negative variables Y. Also, $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots = \log 2$, and $1 - \frac{1}{4} + \frac{1}{9} - \frac{1}{16} + \cdots = \pi^2/12$ (prove it!).]

(c) On the other and parametric hand, suppose that $F(t) = F_\theta(t)$ for some unknown parameter θ, a priori. Then

$$V_{n,\theta}(t) = \sqrt{n}\{F_{\widehat{\theta}}(t) - F_\theta(t)\} \doteq G_\theta(t)'\,\sqrt{n}(\widehat{\theta} - \theta),$$

where

$$G_\theta(t) = \tfrac{\partial}{\partial\theta} F_\theta(t) = \int_{-\infty}^{t} f_\theta(s) \tfrac{\partial \log f_\theta(s)}{\partial\theta}\,\mathrm{d}s.$$

Utilise this to find explicit expressions for the limiting process $V_\theta(t)$, say, in the following cases:

(i) $X_i \sim \mathrm{Exp}(\theta)$. [ANSWER: $V_\theta(t) = \theta t \exp(-\theta t)\,N$, where N is a $N(0,1)$ variable.]
(ii) $X_i \sim N(\mu, \sigma^2)$, σ^2 known.
(iii) $X_i \sim N(\mu, \sigma^2)$, μ known.
(iv) $X_i \sim N(\mu, \sigma^2)$, both parameters unknown. [ANSWER: $V_{\mu,\sigma}(t) = -\phi(\frac{t-\mu}{\sigma})\big[N + \frac{t-\mu}{\sigma} M/\sqrt{2})\big]$, where N and M are independent $N(0,1)$ variables.]
(v) Your own choice. [ANSWER: Go confidently in the direction of your dreams.]

(d) In these examples, $\sqrt{n}\|F_{\widehat{\theta}} - F_\theta\|$ converges to $\|V_\theta\| = \max_{t\in\mathcal{R}} |V_\theta(t)|$ in distribution. Find this limit distribution as explicitly as possible, in each of the five cases considered above.

(e) Try to assess how much smaller $\sqrt{n}\|F_{\widehat{\theta}} - F\|$ will be in these parametric examples, for large n, than the nonparametric $\sqrt{n}\|\widehat{F} - F\|$. You might e.g. compare mean values

and/or median values in the limit distributions. [ANSWERS, for the limiting ratio $\mathrm{E}_F\|F_{\widehat{\theta}} - F\| \,/\, \mathrm{E}_F\|\widehat{F} - F\|$:

(i) $2/(e\pi \log 2) = 1/2.9596 = .3379$.

(ii) $\sqrt{2}/(\pi\sqrt{\pi}\log 2) = 1/2.7292 = .3664$.

(iii) $1/(\pi\sqrt{\pi e}\log 2) = 1/6.3635 = .1571$.

(iv) $\mathrm{E}\, J(\alpha)/\log 2$, where $J(\alpha) = \max\{|h(s_1)|, |h(s_2)|\}$, $h(s) = \{\cos\alpha + (s/\sqrt{2})\sin\alpha\}$ $\phi(s)$, where s_1 and s_2 are the two roots of $h'(s) = 0$, and where $\alpha \sim$ uniform on $[0, 2\pi]$; ratio $= 1/2.1724 = .4603$, evaluated based on numerical integration $\int_0^{2\pi} J(\alpha)\,\mathrm{d}\alpha/(2\pi)$. The value of $\mathrm{E}\, J(\alpha)$ is mysteriously close to $1/\pi$, which would have given ratio $= 1/(\pi \log 2)$.

(v) Live the life you have imagined.]

(f) In the idealised parametric situation, suppose one uses n observations, thereby achieving $\mathrm{E}_F\|F_{\widehat{\theta}} - F\| \doteq a/\sqrt{n}$, where $a = \mathrm{E}\,\|V_\theta\|$ is the constant appropriate for the parametric situation studied. About how much larger must the number m of observations be if the same accuracy is to be obtained using the nonparametric $\widehat{F}$? [ANSWERS: (i) $m \doteq 8.76\,n$; (ii) $m \doteq 7.45\,n$; (iii) $m \doteq 40.49\,n$ (but then that situation is hardly realistic); (iv) $m \doteq 4.72\,n$.]

Exercise No. 16

We continue the theme of the previous exercise. We know how well the distribution F can be estimated nonparametrically, and try to understand and assess how much better it can be estimated in idealised parametric situations. Instead of the quality measure $\|\widetilde{F} - F\|$ considered there, let us now study

$$D(F, \widetilde{F}) = \int |\widetilde{F}(t) - F(t)|\,\mathrm{d}F(t).$$

It is interesting to study the difference between the nonparametric and parametric situation w.r.t. this distance measure in the first place, since it is a very natural loss function, and secondly it is of separate interest for a practicing decision theorist if there is such a thing to see whether two natural but very different loss functions, that of maximal absolute error and that of expected absolute error, give approximately the same qualitative answers.

(a) For the nonparametric case, show that

$$\sqrt{n}\int |\widehat{F}(t) - F(t)|\,\mathrm{d}F(t) \to_d J = \int_0^1 |W^0(u)|\,\mathrm{d}u.$$

Resist trying to find the very intricate probability distribution of J, leave it rather to Larry Shepp (*Annals of Probability*, 1982), but show that $\mathrm{E}\, J = \frac{\pi}{8}\,\mathrm{E}\,|N(0,1)| = \sqrt{2\pi}/8$.

(b) For the general parametric case, where $F = F_\theta$ for some underlying θ, show that

$$\sqrt{n}\int |F_{\widehat{\theta}}(t) - F(t)|\,\mathrm{d}F(t) \to_d \int |V_\theta(t)|\,\mathrm{d}F_\theta(t),$$

in the notation of Exercise 15. Find as explicit expressions as possible for this limit distribution variable in the five parametric cases considered there. [ANSWERS: (i) $|N|/4$; (ii) $|N|/(2\sqrt{\pi})$; (iii) $|N|/(2\sqrt{2}\pi)$; (iv) $\int |N + \frac{1}{2}xM|\phi(x)\,\mathrm{d}x/(2\sqrt{\pi})$.]

(c) Find the limiting ratios $\mathrm{E}_F D(F, F_{\widehat{\theta}}) \,/\, \mathrm{E}_F D(F, \widehat{F})$, and compare with 15(e). [ANSWERS: (i) $2/\pi = .6366$; (ii) $4/(\pi\sqrt{\pi}) = 0.7183$; (iii) $4/(\sqrt{2}\pi^2) = 0.2866$; (iv) $4c/(\pi\sqrt{\pi}) = 0.7968$. Here c is $\mathrm{E}\,(1 + \frac{1}{4}N^2)^{1/2}$, with a numerical value of 1.1093, obtained through numerical integration in MINITAB. Do this yourself, and compare with the value you get from 100,000 simulations, and with what you get, still in MINITAB, after first writing c as an infinite sum.]

(d) As in 15(f), the number m of data points needed to achieve the same accuracy with the nonparametric method as one does with n data points using the parametric method, still assuming that the parametric model is exactly correct which is unrealistically optimistic, needs to be larger than n, but by how much? [ANSWERS: (i) $2.47\,n$; (ii) $1.93\,n$; (iii) $12.17\,n$; (iv) $1.58\,n$.]

(e) Stop to think.

Exercise No. 17

Decision theorists have usually employed quadratic loss functions, like the Gaussian $(\widetilde{\alpha} - \alpha)^2$, but mostly for reasons of mathematical ease. They have offered hopeful remarks that other loss functions, which like the Laplacean $|\widetilde{\alpha} - \alpha|$ might have even stronger intuitive appeal, but are much more complicated to work with (in the classicist sense of obtaining explicit solutions, and so on), ought to give approximately the same qualitative results: estimators derived from the same principle, for loss functions 1 and 2, should be reasonably similar; the difference in quality between estimators 1 and 2 should be similar from the point of view of loss functions 1 and 2; and so on.

Since this is a chance to investigate these matters, however briefly, let's muster the stamina to work through the problems of Exercises 15 and 16, but this time entertaining the quadratic distance measure

$$D(F, \widetilde{F}) = \int \{\widetilde{F}(t) - F(t)\}^2 \,\mathrm{d}F(t).$$

We should admit at the outset that putting up a meaningful distance measure between distribution functions $\widetilde{F}$ and F is much more difficult than measuring distance between real numbers $\widetilde{\alpha}$ and α. Hence one should not expect too much similarity regarding qualitative conclusions drawn from working with $\int |\widetilde{F} - F|\,\mathrm{d}F$ on one side and $\int (\widetilde{F} - F)^2\,\mathrm{d}F$ on the other.

(a) For the nonparametric case, show that

$$nD(F, \widehat{F}) = n \int \{\widehat{F}(t) - F(t)\}^2 \,\mathrm{d}F(t) \to_d \int_0^1 W^0(u)^2 \,\mathrm{d}u = CvM.$$

The distribution of CvM is complicated, but can be expressed in terms of an infinite linear combination of independent χ^2_1-variables, and is tabulated several places. It is

commonly used to test hypotheses of the form $F = F_0$. $D(F, \widehat{F})$ is called the *Cramér–von Mises* test statistic. — Anyway, show that CvM has mean value $1/6$ and standard deviation $1/3$.

(b) For the parametric cases, show that

$$nD(F, F_{\widehat{\theta}}) = n \int \{F_{\widehat{\theta}}(t) - F(t)\}^2 \, \mathrm{d}F(t) \to_d \int V_\theta(t)^2 \, \mathrm{d}F_\theta(t).$$

Find explicit expressions for the limit distribution variable in the five cases you have studied above. [Answers: (i) $N^2 \frac{2}{27}$; (ii) $N^2/(2\sqrt{3}\pi)$; (iii) $N^2/(12\sqrt{3}\pi)$; (iv) $(N^2 + \frac{1}{6}M^2)/(2\sqrt{3}\pi)$; (v) tell me.]

(c) Using the distance measure under consideration, give the limiting ratios $\mathrm{E}_F D(F, F_{\widehat{\theta}})$ / $\mathrm{E}_F D(F, \widehat{F})$, in the five parametric situations. [Answers: (i) .4444; (ii) .5513; (iii) .0919; (iv) .6432.]

(d) And, finally, find the limiting sample-size ratios m/n, in the notation of 15(f) and 16(d). Delight yourself by comparing these numbers to those of 16(d). [Answers: (i) 2.25 (exact!); (ii) 1.81; (iii) 10.88; (iv) 1.55.]

(e) Try to sum up the experience of Exercises 15, 16, and 17. Put up a small table of limiting sample-size ratios m/n, sorted according to parametric model and loss function! Please include other loss functions as well, if you have time, like $D(F, \widetilde{F}) = \left[\int (\widetilde{F} - F)^2 \, \mathrm{d}F\right]^{1/2}$. You might also usefully include sample-size ratios based on studying $\mathrm{med}_F D(F, F_{\widehat{\theta}})$ *vs.* $\mathrm{med}_F D(F, \widehat{F})$.

(f) Speculate about the value of parametric bootstrapping as an alternative to ordinary, nonparametric bootstrapping.

Exercise No. 18

Let $\theta = \theta(F)$ be a functional, defined for all distributions F on the real line. Define

$$I(x) = I(F, x) = \lim_{\varepsilon \to 0} \frac{\theta(F_\varepsilon) - \theta(F)}{\varepsilon},$$

where $F_\varepsilon = (1 - \varepsilon)F + \varepsilon\delta(x)$, and $\delta(x)$ is the point mass probability measure at point x. Thus F_ε is the distribution of a variable Y that with probability $1 - \varepsilon$ is an X drawn from F and with probability ε is equal to x. Such an F_ε is also referred to as a *contamination of* F. The function $I(F, x)$ of x is called the *influence function* for the functional $\theta(.)$.

I hate definitions.

— Benjamin Disraeli

Find as explicit expressions as possible for the influence function in each of the following cases:

(a) $\xi = \xi(F) = \mathrm{E}_F X = \int x \, \mathrm{d}F(x)$.

(b) $\sigma^2 = \sigma^2(F) = \mathrm{Var}_F X = \int \{x - \xi(F)\}^2 \, \mathrm{d}F(x)$.

(c) $\sigma = \sigma(F) = \mathrm{stdev}_F X$. Formulate and prove a general chain rule for the influence function of a functional $\nu(F) = g(\theta(F))$.

(d) $\nu = \nu(F) = \sigma(F)/\xi(F)$, the *coefficient of variation* for F. Formulate and prove a more general chain rule, applicable to functionals $\nu(F) = g(\theta_1(F), \ldots, \theta_p(F))$.

(e) $\gamma = \gamma(F) = \mathrm{E}_F\{X - \xi(F)\}^3$.

(f) $\mu = \mu(F) = F^{-1}(\frac{1}{2})$, the median. Generalise to $F^{-1}(p)$. – Explain, in terms of influence functions, why the median is more robust than the mean.

(g) $\tau = \tau(F) = F^{-1}(\frac{3}{4}) - F^{-1}(\frac{1}{4})$, the interquartile distance, a useful nonparametric measure of spread.

(h) $\theta = \theta(F) = \mathrm{E}_F\,|X - \mu(F)| = \int |x - \mu(F)|\,\mathrm{d}F(x)$, the spread measure studied in Exercises 2 and 3.

Exercise No. 19

The maximum likelihood estimator $\widehat{\theta}$, under a given parametric model $f_\theta(x)$, maximises $(1/n)\sum_{i=1}^n \log f_\theta(X_i) = \int \log f_\theta(x)\,\mathrm{d}\widehat{F}(x)$, and accordingly takes aim at the parameter value $\theta = \theta(F)$ that maximises $\mathrm{E}_F \log f_\theta(X) = \int \log f_\theta(x)\,\mathrm{d}F(x)$. In other and insightful words, the maximum likelihood estimator can be viewed as $\widehat{\theta} = \theta(\widehat{F})$, where $\theta(F)$ is the parameter value that minimises the Kullback-Leibler information distance $\Delta(f, f_\theta) = \int f \log(f/f_\theta)\,\mathrm{d}x$. Find its influence function.

Exercise No. 20

Let $\theta = \theta(F)$ be a parameter functional, and let $\widehat{\theta} = \theta(\widehat{F})$ be the natural nonparametric plug-in estimator. Under suitable regularity conditions,

$$\widehat{\theta} - \theta = \theta(\widehat{F}) - \theta(F) = \frac{1}{n}\sum_{i=1}^n I(F, X_i) + O_p(\frac{1}{n}),$$

where $I(F, .)$ is the influence function that befriended you above. Also, $\mathrm{E}_F I(F, X) = \int I(F, x)\,\mathrm{d}F(x) = 0$ under regularity.

(a) Show that $\sqrt{n}(\widehat{\theta} - \theta)$ converges in distribution to $N(0, \kappa^2)$, where

$$\kappa^2 = \kappa^2(F) = \mathrm{Var}_F I(F, X) = \int I(F, x)^2\,\mathrm{d}F(x).$$

For precise regularity conditions, see for example Boos and Serfling (*Annals of Statistics*, 1980) or Huber's *Robust Statistics* (1981), or James Reed's Ph. D. thesis *On the definition of a von Mises functional* (1976), or Liusa Fernholz' *von Mises calculus for statistical functionals* (1983).

(b) Find the limit distribution for $\sqrt{n}(\widehat{\theta} - \theta)$ in each of the examples of the Exercise 18, by evaluating the κ^2 expression.

Exercise No. 21

And find the limit distribution for $\sqrt{n}(\widehat{\theta} - \theta)$, where $\widehat{\theta}$ is the maximum likelihood estimator, again by evaluating the $\kappa^2 = \int I(F, x)^2\,\mathrm{d}F(x)$ expression, where the influence function $I(F, x)$ was found in Exercise 19. — This amounts to an important discovery in the theory

of parametric inference: You have found the limit distribution of the maximum likelihood estimator when the model is incorrect!, thereby generalising the classical textbook-result, which invariably assumes that the model is right. See also Exercises 22, , and .

Exercise No. 22

Extend the definition of Exercise 18 to define multi-parameter influence functions in multi-dimensional sample spaces. For example, what is the influence function of $\xi = \mathrm{E}_F X$ and of $\Sigma = \mathrm{VAR}_F X = \mathrm{E}_F\{(X-\xi)(X-\xi)'\}$, when $X = (X_1, X_2)'$ is two-dimensional, with a distribution F in $\mathcal{R}^2$? And what is the influence function for the one-dimensional correlation parameter $\rho = \rho(F) = \mathrm{corr}(X_1, X_2)$?

Formulate multi-parameter and multi-dimensional versions of Exercises 19, 20, 21 as well. Show in particular that the influence function for the maximum likelihood functional becomes

$$I(F,x) = J(\theta_0)^{-1}\frac{\partial \log f_{\theta_0}(x)}{\partial\theta},$$

in which $\theta_0 = \theta(F)$ is the parameter value for which $\widehat{\theta}_{\mathrm{ML}}$ is consistent, and $J(\theta)$ is the familiar Fisher information matrix, with elements

$$J_{i,j}(\theta) = -\mathrm{E}_F\frac{\partial^2 \log f_\theta(X)}{\partial\theta_i\partial\theta_j} = -\int \frac{\partial^2 \log f_\theta(x)}{\partial\theta_i\partial\theta_j}\,\mathrm{d}F(x).$$

Exercise No. 23

One can define *empirical influence functions*, for functional parameters or for general estimators, in various ways. Consider in this exercise the nonparametric plug-in estimator $\widehat{\theta} = \theta(\widehat{F})$ for a parameter $\theta = \theta(F)$. Let $\mathbf{P}^* = (P_1^*, \ldots, P_n^*)$ be a general *resampling vector*, a probability distribution on the n data points $x_1, \ldots, x_n$. Thus $\widehat{F} = \widehat{F}(\mathbf{P}^0)$, say, in which $\mathbf{P}^0 = (\frac{1}{n}, \ldots, \frac{1}{n})$. The more general $\widehat{F}(\mathbf{P}^*)$ gives rise to $\widehat{\theta}^* = \theta\big(\widehat{F}(\mathbf{P}^*)\big) = \widehat{\theta}(\mathbf{P}^*)$. — Having established this framework, where the data points are fixed but the probability weights attached to them can vary, introduce

$$U_i(\varepsilon) = \frac{1}{\varepsilon}\big\{\widehat{\theta}(\mathbf{P}_{i,\varepsilon}) - \widehat{\theta}(\mathbf{P}^0)\big\},$$

in which $\mathbf{P}_{i,\varepsilon} = (1-\varepsilon)\mathbf{P}^0 + \varepsilon\Delta_i$, the probability distribution with weight $(1-\varepsilon)/n + \varepsilon$ on x_i and weights $(1-\varepsilon)/n$ on each of the $n-1$ remaining data points.

The influence function concept is tied to the notion of *linearisation*, or the art of finding linear approximations to given functionals or estimators. The idea of a (first order or second order) Taylor expansion of a given estimator can be made precise in several ways. The traditional *delta method* linearises a function of averages by computing partial derivatives w.r.t. these averages. Presently we have been led to consider $\widehat{\theta}$ as a function of resampling weights $\mathbf{P}^*$, and we should look for Taylor expansions of $\widehat{\theta}(\mathbf{P}^*)$ around $\mathbf{P}^* = \mathbf{P}^0$.

(a) The version of $U = (U_1, \ldots, U_n)$ that most naturally matches the idea of a Taylor expansion around $\mathbf{P}^* = \mathbf{P}^0$ uses the partial derivatives

$$U_i = U_i(0) = \lim_{\varepsilon\to 0} U_i(\varepsilon), \quad i = 1, \ldots, n.$$

These are the *tangential influence factors*, making up the *tangential influence vector* U. Compute U_i for $\theta(F)$ being the mean, the variance, the standard deviation, and the variation coefficient.

(b) Show that indeed

$$U_i = U_i(0) = I(\widehat{F}, x_i),$$

with $I(F, x)$ defined as in Exercise 18. You can therefore find expressions for U_i by simply inserting $\widehat{F}$ and x_i in the expressions you were required to find in Exercise 18.

(c) Observe that U_i can be numerically computed in practice by simply putting e.g. $\varepsilon = .000001$ in the definition of $U_i(\varepsilon)$, if an exact expression is hard to derive. Do this, and compare with the explicit solution, for the data set of Exercise 2, for a small list of easy and not-so-easy functionals.

(d) Another choice corresponds to letting $\varepsilon = -1/(n-1)$. Show that the result becomes

$$U_{\text{JACK},i} = U_i(-\frac{1}{n-1}) = (n-1)\{\widehat{\theta} - \widehat{\theta}_{(i)}\},$$

in which $\widehat{\theta}_{(i)} = \widehat{\theta}_{n-1}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ avoids x_i, and is the familiar one from jackknife analysis. Involved in $U_{\text{JACK},i}$ are therefore the n jackknife resampling vectors $\mathbf{P}_{(i)} = (\frac{1}{n-1}, \ldots, 0, \ldots, \frac{1}{n-1})$, in addition to the central $\mathbf{P}^0$.

(e) And still another choice arises by putting $\varepsilon = 1/(n+1)$. Show that this leads to

$$U_{\text{PLUSJACK},i} = (n+1)(\widehat{\theta}_{[i]} - \widehat{\theta}),$$

where $\widehat{\theta}_{[i]} = \widehat{\theta}_{n+1}(x_1, \ldots, x_i, x_i, \ldots, x_n)$ doubles x_i instead of avoiding it. This procedure is called the *positive jackknife* method.

(f) Compute explicitly $U_i(\varepsilon)$, U_i, $U_{\text{JACK},i}$, $U_{\text{PLUSJACK},i}$ for the cases $\xi = \mathrm{E}_F X$ and $\sigma = \mathrm{stdev}_F X$.

Exercise No. 24

The previous exercise established a framework in which (many) nonparametric estimators $\widehat{\theta}$ can be viewed as functions of the resampling vector $\mathbf{P}^* = (P_1^*, \ldots, P_n^*)$. Now consider the bootstrap resampling scheme, in which $P_i^* = N_i/n$ and $(N_1, \ldots, N_n)$ is multinomial $(n; \frac{1}{n}, \ldots, \frac{1}{n})$.

(a) Show (again) that $\mathrm{E}_* \mathbf{P}^* = \mathbf{P}^0$, $\mathrm{VAR}_* \mathbf{P}^* = (I - \frac{1}{n} ee')/n^2$, where $e = (1, \ldots, 1)'$. (We adopt a slightly confusing convention here and in what follows: we follow Efron (SIAM, 1982) in taking the resampling vectors to be line vectors, but let all other non-transposed vectors, like e above and U below, be column vectors.)

(b) Let $\widehat{\theta} = \mu + (1/n)\sum_{i=1}^n \alpha(x_i)$ be a *linear functional statistic*. Show that

$$\widehat{\theta}(\mathbf{P}^*) = \widehat{\theta} + (\mathbf{P}^* - \mathbf{P}^0)U,$$

where $U_i = \alpha_i - \alpha_\cdot = \alpha(x_i) - \sum_{j=1}^n \alpha(x_j)/n$. Note that a constant can be added to all components of U without affecting the result (why?).

(c) Show that $\mathrm{E}_*\widehat{\theta}(\mathbf{P}^*) = \widehat{\theta}$ and that

$$\mathrm{Var}_*\widehat{\theta}(\mathbf{P}^*) = \frac{1}{n^2}\sum_{i=1}^{n}\{\alpha(x_i) - \alpha_.\}^2.$$

This is also the bootstrap estimate of the variance of $\widehat{\theta}$, so in this linear case there is no need to actually carry out bootstrapping by computer simulation.

(d) More generally, let $\widehat{\theta}$ be a *quadratic functional statistic*,

$$\widehat{\theta}(x_1, \ldots, x_n) = \mu + \frac{1}{n}\sum_{i=1}^{n}\alpha(x_i) + \frac{1}{n^2}\sum_{i<j}\beta(x_i, x_j).$$

Here μ, $\alpha(.)$, and $\beta(.,.)$ are allowed to depend upon n, see e.g. page 24 of Efron (SIAM, 1982). Give an expression for the underlying $\theta(F)$ functional, in terms of F. Show that $\widehat{\theta}$ also is a quadratic function of $\mathbf{P}^*$,

$$\widehat{\theta}(\mathbf{P}^*) = a + (\mathbf{P}^* - \mathbf{P}^0)U + \tfrac{1}{2}(\mathbf{P}^* - \mathbf{P}^0)V(\mathbf{P}^* - \mathbf{P}^0)',$$

where $U_i = \alpha_i - \alpha_. + \beta_{i.} - \beta_{..}$ and $V_{ij} = \beta_{ij} - \beta_{i.} - \beta_{.j} + \beta_{..}$. Note that $\mathbf{P}^0 U = 0$, $\mathbf{P}^0 V = 0$.

(e) Obtain the following expression for the bootstrap expectation of $\widehat{\theta}^*$:

$$\mathrm{E}_*\widehat{\theta}(\mathbf{P}^*) = \widehat{\theta}(\mathbf{P}^0) + \frac{1}{2n^2}\sum_{i=1}^{n}V_{ii}.$$

What is the bootstrap estimate of the bias for $\widehat{\theta}$?

(f) Why should one stop? Go on to the third order.

Exercise No. 25

The previous exercise considered bootstrap properties of linear and quadratic functional statistics. What happens if one approximates a given, complicated $\widehat{\theta}(x_1, \ldots, x_n)$ with a linear or a quadratic statistic? — The notation quickly gets involved here; we use

$$\widehat{\mathrm{bias}}^{\mathrm{methodA}}\{\mathrm{estimatorB}\} \quad \text{and} \quad \widehat{\mathrm{Var}}^{\mathrm{methodC}}\{\mathrm{estimatorD}\}$$

to denote respectively the methodA-based estimate for the bias of estimatorB and the methodC-based estimate for the variance of estimatorD.

(a) Let $\widehat{\theta}_{\mathrm{LIN}}$ be any suitable linear approximation to $\widehat{\theta}$, of the form $\widehat{\theta}_{\mathrm{LIN}}(\mathbf{P}^*) = a + (\mathbf{P}^* - \mathbf{P}^0)U$. Show that the bootstrap method estimate of the variance of $\widehat{\theta}_{\mathrm{LIN}}$ can be written

$$\widehat{\mathrm{Var}}^{\mathrm{boot}}\{\widehat{\theta}_{\mathrm{LIN}}\} = \frac{1}{n^2}\sum_{i=1}^{n}(U_i - \bar{U})^2.$$

(b) One particular linear approximation is

$$\widehat{\theta}_{\mathrm{TAN}} = \widehat{\theta}_{\mathrm{TAN}}(\mathbf{P}^*) = \widehat{\theta} + (\mathbf{P}^* - \mathbf{P}^0)U,$$

in which $U = U(0)$ is the tangential influence vector studied in Exercise 23, also called the infinitesimal jackknife factors. Demonstrate that this leads to

$$\widehat{\mathrm{Var}}^{\mathrm{boot}}\{\widehat{\theta}\} \approx \widehat{\mathrm{Var}}^{\mathrm{boot}}\{\widehat{\theta}_{\mathrm{TAN}}\} = \frac{1}{n^2}\sum_{i=1}^{n} I(\widehat{F}, x_i)^2.$$

(c) Another linear approximation is available:

$$\widehat{\theta}_{\mathrm{JACK}} = \widehat{\theta}_{\mathrm{JACK}}(\mathbf{P}^*) = a + (\mathbf{P}^* - \mathbf{P}^0)U,$$

where a and U are chosen such that the approximation and the given $\widehat{\theta}$ agree for $\mathbf{P}^* = \mathbf{P}_{(i)}$, for $i = 1, \ldots, n$. The linear functional is uniquely determined by these requirements, although the representation above is not, since a constant can be added to each element of U without changing the result. Show that one specification that works is $a = \widehat{\theta}_{(\cdot)}$ and $U_i = (n-1)\{\widehat{\theta}_{(\cdot)} - \widehat{\theta}_{(i)}\}$; it satisfies $\bar{U} = \mathbf{P}^0 U = 0$. Note the connection to Exercise 23(d). Show that

$$\begin{aligned}\widehat{\mathrm{Var}}^{\mathrm{boot}}\{\widehat{\theta}\} &\approx \widehat{\mathrm{Var}}^{\mathrm{boot}}\{\widehat{\theta}_{\mathrm{JACK}}\} \\ &= \Big(\frac{n-1}{n}\Big)^2 \sum_{i=1}^{n}\Big[\widehat{\theta}_{(i)} - \widehat{\theta}_{(\cdot)}\Big]^2 = \frac{n-1}{n}\,\widehat{\mathrm{Var}}^{\mathrm{JACK}}\{\widehat{\theta}\}.\end{aligned}$$

(d) Explore the analogous linear approximation $\widehat{\theta}_{\mathrm{PLUSJACK}}$ that is defined by requiring it to agree with $\widehat{\theta}(\mathbf{P}^*)$ for $\mathbf{P}_{[i]} = (\frac{1}{n+1}, \ldots, \frac{2}{n+1}, \ldots, \frac{1}{n+1})$, cf. Exercise 23(e). Write down a formula for

$$\widehat{\mathrm{Var}}^{\mathrm{PLUSJACK}}\{\widehat{\theta}\} = \widehat{\mathrm{Var}}^{\mathrm{boot}}\{\widehat{\theta}_{\mathrm{PLUSJACK}}\}.$$

(e) Define a quadratic approximation to the given $\widehat{\theta}$ of the form

$$\widehat{\theta}_{\mathrm{QUAD}} = \widehat{\theta}_{\mathrm{QUAD}}(\mathbf{P}^*) = a + (\mathbf{P}^* - \mathbf{P}^0)U + \tfrac{1}{2}(\mathbf{P}^* - \mathbf{P}^0)V(\mathbf{P}^* - \mathbf{P}^0)',$$

where a and U and V are such that the approximating statistic agrees with $\widehat{\theta}(\mathbf{P}^*)$ for $\mathbf{P}^0$ and for $\mathbf{P}_{(1)}, \ldots, \mathbf{P}_{(n)}$. One can always arrange matters so that $\mathbf{P}^0 U = 0$, $\mathbf{P}^0 V = 0$, cf. Exercise 24(d). — Obtain the following useful approximation formula for a general bootstrap expectation:

$$\begin{aligned}\mathrm{E}_*\Big[\widehat{\theta}(\mathbf{P}^*) - \widehat{\theta}(\mathbf{P}^0)\Big] &\approx \mathrm{E}_*\Big[\widehat{\theta}_{\mathrm{QUAD}}(\mathbf{P}^*) - \widehat{\theta}_{\mathrm{QUAD}}(\mathbf{P}^0)\Big] \\ &= \frac{n-1}{n}\,(n-1)\{\widehat{\theta}_{(\cdot)} - \widehat{\theta}\} = \frac{n-1}{n}\,\widehat{\mathrm{bias}}^{\mathrm{JACK}}\{\widehat{\theta}\}.\end{aligned}$$

(f) Researchers in the field seem to tend to drop the $(n-1)/n$ factor here, and accordingly promote $(n-1)\{\widehat{\theta}_{(\cdot)} - \widehat{\theta}\}$ as the second order jackknife approximation to the bootstrap expected value of $\widehat{\theta}(\mathbf{P}^*) - \widehat{\theta}(\mathbf{P}^0)$. This leads to the classical jackknife estimate of bias, and adds authority and interpretational substance to this method, but there are otherwise no good reasons for dropping the $(n-1)/n$ factor. — How should one construct a third order approximation to the bootstrap bias?

Exercise No. 26

Prove that the equation $x^n + y^n = z^n$ cannot have integer solutions when the integer exponent n is greater than or equal to three. HINT: Consult Yoichi Miyaoka, October 1988. Deduce, as a *Corollary*, that the margin sometimes is narrower than one thinks.

Exercise No. 27

Consider the framework of Exercise 14: $X_1, \ldots, X_n$ are i.i.d. from F, and the statistic under consideration, not necessarily an estimator, is a smooth function of averages, or can be approximated by such a function. Write $\widehat{\theta} = h(\bar{A}_1, \ldots, \bar{A}_p)$, where $\bar{A}_j$ is the average of $A_{j1} = g_j(X_1), \ldots, A_{jn} = g_j(X_n)$, say. — There are now a variety of nonparametric methods available to the statistician for estimating the standard deviation of $\widehat{\theta}$.

♢ The *delta method* reviewed in Exercise 14 gives

$$\widehat{\mathrm{Var}}^{\mathrm{DELTA}}\{\widehat{\theta}\} = \frac{1}{n}\frac{\partial h}{\partial a}(\bar{A})' \,\widehat{\Sigma}\, \frac{\partial h}{\partial a}(\bar{A}),$$

where $\widehat{\Sigma} = \sum_{i=1}^n (A_i - \bar{A})(A_i - \bar{A})'/(n-1)$.

♢ The *influence function approach* utilises the limit distribution result of Exercise 20, which says that $\kappa^2(F)/n = \int I(F,x)^2 \,\mathrm{d}F(x)/n$ approximates $\mathrm{Var}\,\widehat{\theta}$. The natural nonparametric estimate of this asymptotic variance is

$$\widehat{\mathrm{Var}}^{\mathrm{INFLUENCE}}\{\widehat{\theta}\} = \frac{1}{n}\kappa^2(\widehat{F}) = \frac{1}{n}\int I(\widehat{F},x)^2 \,\mathrm{d}\widehat{F}(x).$$

♢ The *jackknife method* produces

$$\widehat{\mathrm{Var}}^{\mathrm{JACK}}\{\widehat{\theta}\} = \frac{n-1}{n}\sum_{i=1}^n \Big[\widehat{\theta}_{(i)} - \widehat{\theta}_{(\cdot)}\Big]^2.$$

♢ The *tangential influence viewpoint* or the *infinitesimal jackknife* yields

$$\widehat{\mathrm{Var}}^{\mathrm{TAN}}\{\widehat{\theta}\} = \widehat{\mathrm{Var}}^{\mathrm{boot}}\{\widehat{\theta}_{\mathrm{TAN}}\} = \frac{1}{n^2}\sum_{i=1}^n I(\widehat{F},x_i)^2.$$

♢ The *bootstrap method* estimate is

$$\widehat{\mathrm{Var}}^{\mathrm{boot}}\{\widehat{\theta}\} = \mathrm{E}_*\{\widehat{\theta}^* - \mathrm{E}_*\widehat{\theta}^*\}^2.$$

♢ And let us throw in the *jackknife approximation to the bootstrap estimate*, to boot, namely

$$\widehat{\mathrm{Var}}^{\mathrm{BOOTJACK}}\{\widehat{\theta}\} = \widehat{\mathrm{Var}}^{\mathrm{BOOT}}\{\widehat{\theta}_{\mathrm{JACK}}\} = \Big(\frac{n-1}{n}\Big)^2 \sum_{i=1}^{n}\Big[\widehat{\theta}_{(i)} - \widehat{\theta}_{(\cdot)}\Big]^2.$$

Explore the degree of equivalence of these approaches, and prove that all formulae are asymptotically equivalent, and in fact consistent, under mild regularity. In this case, where it is known that $n\mathrm{Var}\,\widehat{\theta}$ converges to some limit κ^2, consistency means that n times the variance estimate also converges, in probability, to κ^2.

Exercise No. 28

Sometimes interest focusses on a quantity that depends upon both the data and the unknown parameters of the model. In the nonparametric framework this means a function $R = R(F, \mathbf{x})$, where $\mathbf{x} = (x_1, \ldots, x_n)$ is the vector of observed data, being i.i.d. $\sim F$. The bootstrap equivalent to R is

$$R^* = R(\widehat{F}, \mathbf{x}^*) = R(\widehat{F}, x_1^*, \ldots, x_n^*),$$

where the x_i^*'s are i.i.d. $\sim \widehat{F}$.

One particular use of R^* is to approximate $\mathrm{E}_F R(F, \mathbf{X})$ with $\mathrm{E}_* R(\widehat{F}, \mathbf{X}^*)$. Re-employ the reasoning of Exercise 25(e) to arrive at the following jackknife approximation:

$$\mathrm{E}_* R(\widehat{F}, \mathbf{X}^*) = \mathrm{E}_* R(\mathbf{P}^*) \doteq \frac{(n-1)^2}{n}\{R_{(\cdot)} - R(\mathbf{P}^0)\} \doteq (n-1)\{R_{(\cdot)} - R(\mathbf{P}^0)\},$$

where $R_{(\cdot)}$ is the average of the n values of $R(\mathbf{P}_{(i)})$.

Exercise No. 29

As a simple example, assume estimates

$$\widehat{\xi} = \widehat{\xi}(x_1, \ldots, x_n) = \bar{x} \quad \text{and} \quad \widehat{\sigma}^2 = \widehat{\sigma}^2(x_1, \ldots, x_n) = \sum_{i=1}^{n}(x_i - \bar{x})^2/n$$

have been extracted from the data sample, and let

$$\begin{aligned}\pi = \pi(F, \mathbf{x}) &= \Pr_F\{X_{\mathrm{new}} > \widehat{\xi} + 1.645\,\widehat{\sigma}\} \\ &= \mathrm{E}_F I\{X_{\mathrm{new}} > \widehat{\xi}(\mathbf{x}) + 1.645\,\widehat{\sigma}(\mathbf{x})\} \\ &= \int I\{x_{\mathrm{new}} > \widehat{\xi} + 1.645\,\widehat{\sigma}\}\,\mathrm{d}F(x_{\mathrm{new}}).\end{aligned}$$

Here X_{new} denotes a future observation, independent of the given training set $x_1, \ldots, x_n$ of data, and $\Pr_F$ and E_F refer to probability statements w.r.t. X_{new}, with the training data fixed at their observed values. In prediction situations one is interested in precisely

such probability statements w.r.t. forthcoming data, as opposed to statements about the average probability over all possible training sets.

(a) Suppose for a minute that F is in fact Gaussian (ξ, σ^2). Give an expression for $\pi(F, \mathbf{x})$, and contrast it with $\pi(F) = \mathrm{E}_{\mathrm{all}}\, \pi(F, \mathbf{X})$, where $X_1, \ldots, X_n, X_{\mathrm{new}}$ are i.i.d. F in the $\mathrm{E}_{\mathrm{all}}$ statement.

(b) A simplistic estimator of $\pi(F, \mathbf{x})$ is

$$\widehat{\pi}^{\mathrm{NAIVE}} = \pi(\widehat{F}, \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} I\{x_i > \widehat{\xi}(\mathbf{x}) + 1.645\, \widehat{\sigma}(\mathbf{x})\}.$$

Why is it naïve?
Write

$$\pi = \widehat{\pi}^{\mathrm{NAIVE}} + \text{naïvité},$$

where

$$\text{naïvité} = \text{naïvité}(F, \mathbf{x}) = \pi(F, \mathbf{x}) - \widehat{\pi}^{\mathrm{NAIVE}}(\mathbf{x})$$

again is a random quantity depending upon both F and the training data $x_1, \ldots, x_n$. Let

$$\omega = \omega(F) = \mathrm{E}_{\mathrm{all}}\, \text{naïvité}(F, \mathbf{X})$$

be the average naïvité, over all possible training sets. The idea is to estimate $\pi(F, \mathbf{x})$ by correcting the naïve estimator for its average naïvité:

$$\widehat{\pi} = \widehat{\pi}^{\mathrm{NAIVE}} + \widehat{\omega}.$$

(c) Show that the bootstrap estimate of the average naïvité becomes

$$\widehat{\omega}^{\mathrm{BOOT}} = \omega(\widehat{F}) = \mathrm{E}_* \Big[\pi(\widehat{F}, \mathbf{X}^*) - \widehat{\pi}^{\mathrm{NAIVE}}(\mathbf{X}^*) \Big] = \mathrm{E}_* R(\mathbf{P}^*),$$

in which

$$\begin{aligned} R(\mathbf{P}^*) &= \frac{1}{n} \sum_{i=1}^{n} I\{x_i > \widehat{\xi}^* + 1.645\, \widehat{\sigma}^*\} - \frac{1}{n} \sum_{i=1}^{n} I\{x_i^* > \widehat{\xi}^* + 1.645\, \widehat{\sigma}^*\} \\ &= \sum_{i=1}^{n} \Big(\frac{1}{n} - P_i^* \Big)\, I\{x_i > \widehat{\xi}^* + 1.645\, \widehat{\sigma}^*\}, \end{aligned}$$

and $\widehat{\xi}^* = \widehat{\xi}(x_1^*, \ldots, x_n^*)$, $\widehat{\sigma}^* = \widehat{\sigma}(x_1^*, \ldots, x_n^*)$. The net result is

$$\widehat{\pi}^{\mathrm{BOOT}} = \widehat{\pi}^{\mathrm{NAIVE}} + \widehat{\omega}^{\mathrm{BOOT}}.$$

Can you construct a MINITAB macro that computes this?

(d) A simpler alternative to $\widehat{\pi}^{\mathrm{BOOT}}$ is the cross validation or leave-one-out estimate

$$\widehat{\pi}^{\mathrm{CROSS}} = \frac{1}{n} \sum_{i=1}^{n} I\{x_i > \widehat{\xi}_{(i)} + 1.645\, \widehat{\sigma}_{(i)}\},$$

where $\widehat{\xi}_{(i)}$ and $\widehat{\sigma}_{(i)}$ are computed from $\mathbf{x}_{(i)} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$. Give motivation for this estimator. — Note that the cross validation estimator for π corresponds to a cross validation estimator $\widehat{\omega}^{\mathrm{CROSS}}$ for ω.

(e) Following the scheme of the previous exercise, evaluate

$$\widehat{\omega}^{\mathrm{BOOTJACK}} = (n-1)\{R_{(\cdot)} - R(\mathbf{P}^0)\}.$$

Show that this leads to

$$\widehat{\pi}^{\mathrm{BOOTJACK}} = \widehat{\pi}^{\mathrm{CROSS}} + \widehat{\pi}^{\mathrm{NAIVE}} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} I\{x_j > \widehat{\xi}_{(i)} + 1.645\,\widehat{\sigma}_{(i)}\},$$

and find out how small the difference between the cross validation estimator and the bootjack estimator is.

Exercise No. 30

Exercises 24 and 25 looked into jackknife type approximations to bootstrap expectations, up to second order. The present exercise discusses third order approximations.

(a) Consider a third order functional statistic of the form

$$\widehat{\theta}_{\mathrm{THIRD}}(\mathbf{P}^*) = a + (\mathbf{P}^* - \mathbf{P}^0)U + \tfrac{1}{2}(\mathbf{P}^* - \mathbf{P}^0)V(\mathbf{P}^* - \mathbf{P}^0)' + \tfrac{1}{6}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} W_{ijk}(P_i^* - \tfrac{1}{n})(P_j^* - \tfrac{1}{n})(P_k^* - \tfrac{1}{n}),$$

where V_{ij} and W_{ijk} are symmetric in their arguments. Show that one without loss of generality can take each of the averages $U_\cdot$, $V_{i\cdot}$, $V_{\cdot j}$, $V_{\cdot\cdot}$, $W_{ij\cdot}$, $W_{i\cdot k}$, $W_{\cdot jk}$, $W_{i\cdot\cdot}$, $W_{\cdot j\cdot}$, $W_{\cdot\cdot k}$, $W_{\cdot\cdot\cdot}$ to be zero. [Hint: Consider $W_{ijk}^{new} = W_{ijk} - W_{ij\cdot} - W_{i\cdot k} - W_{\cdot jk} + W_{i\cdot\cdot} + W_{\cdot j\cdot} + W_{\cdot\cdot k} - W_{\cdot\cdot\cdot}$.]

(b) Recall that $P_i^* = N_i/n$, where $(N_1, \ldots, N_n)$ is multinomial $(n; \frac{1}{n}, \ldots, \frac{1}{n})$. Show that

$$\mathrm{E}_*\Big(P_i^* - \frac{1}{n}\Big)\Big(P_j^* - \frac{1}{n}\Big)\Big(P_k^* - \frac{1}{n}\Big) = \Big(\frac{1}{n}\Big)^3 \mathrm{E}_*(N_i - 1)(N_j - 1)(N_k - 1)$$
$$= \begin{cases} \big(\frac{1}{n}\big)^3\big(1 - \frac{1}{n}\big)\big(1 - \frac{2}{n}\big) & \text{if } i,j,k \text{ are equal;} \\ -\big(\frac{1}{n}\big)^4\big(1 - \frac{2}{n}\big) & \text{if two among } i,j,k \text{ are equal;} \\ 2\big(\frac{1}{n}\big)^5 & \text{if } i,j,k \text{ are distinct.} \end{cases}$$

(c) Show that

$$\sum_{\text{two equal}} W_{ijk} = -3\sum_{i=1}^{n} W_{iii}, \qquad \sum_{\text{three distinct}} W_{ijk} = 2\sum_{i=1}^{n} W_{iii}.$$

(d) Arrive safely at

$$\mathrm{E}_*\widehat{\theta}_{\mathrm{THIRD}}(\mathbf{P}^*) = a + \frac{A}{2n} + \frac{B}{6n^2},$$

where

$$A = \frac{1}{n}\sum_{i=1}^{n} V_{ii} \quad \text{and} \quad B = \frac{1}{n}\sum_{i=1}^{n} W_{iii}.$$

(e) Now let $\widehat{\theta}$ be any given, complicated functional statistic, which we want to approximate with the third order functional statistic $\widehat{\theta}_{\text{THIRD}}$ above. There are several candidates, one of which is the following: Specify a, U, V_{ij}, W_{ijk} above such that $\widehat{\theta}_{\text{THIRD}}(\mathbf{P}^*)$ agrees with $\widehat{\theta}(\mathbf{P}^*)$ for $\mathbf{P}^*$ equal to $\mathbf{P}^0$, $\mathbf{P}_{(i)}$, $\mathbf{P}_{[i]}$ for $i = 1, \ldots, n$, cf. Exercises 23(e) and 25(d). Show that this leads to

$$J_{(\cdot)} =_{\text{def}} \frac{(n-1)^2}{n}\{\widehat{\theta}_{(\cdot)} - \widehat{\theta}\} = \frac{A}{2n} - \frac{n}{n-1}\frac{B}{6n^2},$$

$$J_{[\cdot]} =_{\text{def}} \frac{(n+1)^2}{n}\{\widehat{\theta}_{[\cdot]} - \widehat{\theta}\} = \frac{A}{2n} + \frac{n}{n+1}\frac{B}{6n^2}.$$

The quantities $J_{(\cdot)}$ and $J_{[\cdot]}$ entering here are the natural second-order estimates of the bias $\mathrm{E}_*\{\widehat{\theta}^* - \widehat{\theta}\}$, based on respectively the ordinary jackknife and the positive jackknife; in particular $J_{(\cdot)}$ is the usual bootjack estimator of bias recommended by Efron.

(f) Show that this strategy leads to

$$\begin{aligned}\mathrm{E}_*\{\widehat{\theta}(\mathbf{P}^*) - \widehat{\theta}(\mathbf{P}^0)\} &\doteq \mathrm{E}_*\{\widehat{\theta}_{\text{THIRD}}(\mathbf{P}^*) - \widehat{\theta}_{\text{THIRD}}(\mathbf{P}^0)\} \\ &= J_{(\cdot)} + \frac{(n+1)(2n-1)}{2n^2}\big[J_{[\cdot]} - J_{(\cdot)}\big].\end{aligned}$$

(g) An alternative specification of a third order approximation to $\widehat{\theta}$ arises by requiring the $\widehat{\theta}(\mathbf{P}^*)$ and its approximation to agree for $\mathbf{P}^0$, $\mathbf{P}_{(i)}$, $\mathbf{P}_{(i,j)}$ for all $i, j = 1, \ldots, n$, where $\mathbf{P}_{(i,j)}$ comes from the double jackknife that deletes both x_i and x_j, cf. Exercise 9. Show that the natural second-order bias approximation estimate based on the double jackknife is

$$J_{(\cdot\,\cdot)} =_{\text{def}} \frac{(n-1)(n-2)}{2n}\{\widehat{\theta}_{(\cdot\,\cdot)} - \widehat{\theta}\} = \frac{A}{2n} - \frac{n(n-4)}{(n-2)^2}\frac{B}{6n^2}.$$

(h) And obtain finally a third-order jack *&* double-jack approximation to a general bootstrap expectation:

$$\begin{aligned}\mathrm{E}_*\{\widehat{\theta}(\mathbf{P}^*) - \widehat{\theta}(\mathbf{P}^0)\} &\doteq \mathrm{E}_*\{\widehat{\theta}_{\text{THIRD}}(\mathbf{P}^*) - \widehat{\theta}_{\text{THIRD}}(\mathbf{P}^0)\} \\ &= J_{(\cdot)} + \frac{(n-2)^2(2n-1)}{n^2}\big[J_{(\cdot\,\cdot)} - J_{(\cdot)}\big].\end{aligned}$$

Exercise No. 31

Here comes another example of the methods of Exercises 28 and 29, this time in a context related to nonparametric density estimation, a topic that will be discussed more fully in

exercises to come. For given data $x_1, \ldots, x_n$, realisations of variables $X_1, \ldots, X_n$ from F, define

$$\widehat{f}(x) = \widehat{f}(x; x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} \phi\Big(\frac{x - x_i}{h}\Big)\frac{1}{h}.$$

Here $\phi(x)$ is the standard normal density, and h is a smoothing parameter. The following quantity will be of interest later, as a function of h:

$$\rho = \rho(h) = \rho(F, x_1, \ldots, x_n) = \int f\widehat{f}\,\mathrm{d}x = \mathrm{E}_F\, \widehat{f}(X_{\mathrm{new}}; \mathbf{x}).$$

For the moment we keep h fixed.

(a) Why is

$$\widehat{\rho}^{\,\mathrm{NAIVE}} = \rho(\widehat{F}, \mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}(x_i; \mathbf{x}) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \phi\Big(\frac{x_i - x_j}{h}\Big)\frac{1}{h}$$

naïve?

(b) And why is

$$\widehat{\rho}^{\,\mathrm{CROSS}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}(x_i; \mathbf{x}_{(i)}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{(i)}(x_i)$$

a better estimator?

(c) In this case there is a simple connection between $\widehat{\rho}^{\,\mathrm{NAIVE}}$ and $\widehat{\rho}^{\,\mathrm{CROSS}}$. Work out an expression for $\widehat{f}_{(i)}(x)$, and show that

$$\widehat{\rho}^{\,\mathrm{CROSS}} = \frac{n}{n-1}\widehat{\rho}^{\,\mathrm{NAIVE}} - \frac{1}{n-1}\frac{\phi(0)}{h}.$$

(d) Write $\rho = \widehat{\rho}^{\,\mathrm{NAIVE}} + \text{naïvité}$, where naïvité $= \text{naïvité}(F, \mathbf{x}) = \rho(F, \mathbf{x}) - \widehat{\rho}^{\,\mathrm{NAIVE}}(\mathbf{x})$, with average naïvité

$$\omega = \omega(F) = \mathrm{E}_{\mathrm{all}}\, \text{naïvité}(F, X_1, \ldots, X_n),$$

in analogy with Exercise 29, suggesting estimators of the type $\widehat{\rho} = \widehat{\rho}^{\,\mathrm{NAIVE}} + \widehat{\omega}$. Study in particular

$$\widehat{\rho}^{\,\mathrm{BOOT}} = \widehat{\rho}^{\,\mathrm{NAIVE}} + \widehat{\omega}^{\,\mathrm{BOOT}},$$

and demonstrate that $\widehat{\omega}^{\mathrm{BOOT}} =_{\mathrm{def}} \omega(\widehat{F}) = \mathrm{E}_* R(\mathbf{P}^*)$, in which

$$\begin{aligned} R(\mathbf{P}^*) &= \frac{1}{n}\sum_{i=1}^{n} \widehat{f}(x_i; \mathbf{x}^*) - \frac{1}{n}\sum_{i=1}^{n} \widehat{f}(x_i^*; \mathbf{x}^*) \\ &= \sum_{i=1}^{n} \Big(\frac{1}{n} - P_i^*\Big)\, \widehat{f}(x_i; \mathbf{x}^*). \end{aligned}$$

(e) Use the jackknife approximation $(n-1)\{R_{(\cdot)} - R(\mathbf{P}^0)\}$ of Exercise 28 to obtain a jackknife approximation $\widehat{\omega}^{\mathrm{BOOTJACK}}$ to $\widehat{\omega}^{\mathrm{BOOT}}$. Show that this in fact leads to

$$\widehat{\rho}^{\mathrm{BOOTJACK}} \equiv \widehat{\rho}^{\mathrm{CROSS}}.$$

The more accurate $(n-1)^2/n$ factor, see Exercise 28 again, indicates that an even better approximation might be $\{(n-1)/n\}\widehat{\rho}^{\mathrm{CROSS}}$.

(f) To illustrate, draw a sample of size 30 from the standard normal $f = \phi$, and compute and plot three curves: the true $\rho(h)$ and the two estimates $\widehat{\rho}^{\mathrm{NAIVE}}(h)$, $\widehat{\rho}^{\mathrm{CROSS}}(h)$. Plot the fourth curve $\widehat{\rho}^{\mathrm{BOOT}}(h)$ as well, if you have time. [Show first that

$$\rho(h) = \frac{1}{n}\sum_{i=1}^{n} \phi\Big(\frac{x_i}{\sqrt{1+h^2}}\Big)\frac{1}{\sqrt{1+h^2}}.]$$

I got the following figure from my own simulated 30-sample, where the true curve is the dotted one, the top solid line is the naïve estimate, and the bottom solid line is the cross validation estimate:

Exercise No. 32

The previous and several of the forthcoming exercises concern resampling techniques applied to problems in nonparametric density estimation, so let's learn about density estimates first.

Assume that $X_1, \ldots, X_n$ are i.i.d. with unknown probability density f on (some portion of) the real line. The *kernel method* is among the simpler ones of the multitude of nonparametric methods for estimating f that have been proposed since 1956, and works as follows. Let $K(.)$ be *a kernel function*, taken here to be a probability density itself which is symmetric about zero, and with $\int yK(y)\,\mathrm{d}y = 0$, $\int y^2K(y)\,\mathrm{d}y = 1$. Let

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big)\frac{1}{h}$$

for x in the region of interest. To choose kernel function $K(.)$ and smoothing parameter (or window width) h is part of the estimation problem.

(a) Show that $\widehat{f}$ indeed is a density. Show that $\widehat{f}$ corresponds to the probability distribution $\widehat{F} * K_h$, the convolution of the empirical distribution $\widehat{F}$ with the distribution with density $K_h(x) = K(x/h)/h$. In still other words: if X is sampled from $\widehat{f}$, then $X \sim Y + Z$, where $Y \sim \widehat{F}$ (a randomly sampled data point X_i), $Z \sim K_h$, and these two are independent.

(b) If X is sampled from $\widehat{f}$ and how would you do that?, what is its mean value and variance?

(c) Show that

$$\mathrm{E}\,\widehat{f}(x) \doteq f(x) + \tfrac{1}{2}h^2 f''(x) + \tfrac{1}{24}h^4 \int y^4K(y)\,\mathrm{d}y\, f^{(iv)}(x).$$

How must h behave for the bias of $\widehat{f}(x)$ to go to zero when n grows?

(d) And show that

$$\operatorname{Var}\widehat{f}(x) \doteq \frac{B}{nh}\Big[f(x) + \tfrac{1}{2}h^2 \int y^2 K(y)^2 \,\mathrm{d}y\, f''(x)\Big] - \frac{1}{n}\Big[f(x)^2 + O(h^4)\Big],$$

where $B = \int K(y)^2 \,\mathrm{d}y$ is a constant determined by the kernel function. How must h behave for the variance of $\widehat{f}(x)$ to go to zero as n grows?

(e) Combine your efforts to obtain an expression for the mean squared error, under regularity conditions on the true density:

$$\mathrm{E}\,\{\widehat{f}(x) - f(x)\}^2 = \tfrac{1}{4}h^4 f''(x)^2 + \frac{B}{nh}f(x) - \frac{1}{n}f(x)^2 + O(h/n + h^6).$$

How should h behave with n?

(f) Try to reach similar conclusions with L_1-based criteria: each of the three terms appearing in

$$E\,\big|\widehat{f}(x) - f(x)\big| \le \mathrm{E}\,\big|\widehat{f}(x) - \mathrm{E},\widehat{f}(x)\big| + \big|\mathrm{E}\,\widehat{f}(x) - f(x)\big|$$

should ideally be small, and at least converge to zero as n grows.

Exercise No. 33

The task of estimating and predicting permeability in prospective reservoirs is central for oil companies. One strategy is to construct a prediction or estimation rule as a function of well log measurements. This is a difficult problem, and will not be dwelt on here. Our aim is only to consider a particular sub-problem, by means of density estimation and various connected cross validation and bootstrap techniques, namely the following: Are there (more or less homogeneous) sub-groups in the distribution of permeability, in the region of interest, and can they be identified? Geologists disagree over this question. An affirmative answer would suggest that prediction rules in some way should take these sub-groups into account.

The following 70 data points are *transformed* permeability measurements, taken from a certain well in a certain area along the Norwegian coast, by a certain oil company. They have been made unrecognisable and uninterpretable for reasons of security: the particular transformation used has no interpretation whatsoever, is only known to the present writer, and cannot be guessed at; and the data points have been ordered, so that the spatial information inherent in the original ordering has been lost. — In this way the secrets of the original and very expensive data set are guarded, while the less-sensitive problem about clusters still can be discussed in a meaningful way.

```
Transformed perm-data
6.033   6.125   6.196   6.436   6.516   6.568   6.751   6.796
6.854   7.009   7.044   7.068   7.097   7.108   7.150   7.205
7.253   7.415   7.444   7.570   7.576   7.597   7.724   7.784
7.808   8.094   8.108   8.125   8.126   8.303   8.371   8.450
8.483   8.495   8.562   8.570   8.691   8.816   8.900   9.024
9.079   9.195   9.308   9.573   9.617   9.696  10.064  11.278
```

```
11.620   11.653   11.831   11.838   11.954   12.235   12.460   12.838
12.854   12.870   12.895   14.101   14.222   14.222   14.298   14.308
14.391   14.743   14.875   15.023   15.046   15.195
```

```
Histogram of transformed perm-data   N = 70
Midpoint   Count
       6       4  ****
       7      15  ***************
       8      15  ***************
       9       9  *********
      10       4  ****
      11       1  *
      12       7  *******
      13       4  ****
      14       6  ******
      15       5  *****
```

(a) Construct a MINITAB-macro (or something equivalent in your own environment) that allows you to compute and display the nonparametric estimate $\widehat{f}(x)$, for a given smoothing parameter h, using the ordinary Gaussian kernel $K(x) = \phi(x) = \exp(-\frac{1}{2}x^2)/\sqrt{2\pi}$. Every statistician should have such an algorithm in her collection of ever-useful tricks & gadgets.

(b) Use such an algorithm to evaluate $\widehat{f}(x)$ for the transformed permeability data. Try out different h values. Pinpoint what is wrong with too small values and with too large values. What range of h values seems to give satisfactory pictures?

(c) What is the accompanying cumulative distribution $\widehat{F}_h(x)$? And what happens to $\widehat{F}_h(.)$ when h tends to zero? — Have a look at these, for some values of h, in the present example. What are the benefits of studying pictures of $\widehat{f}$ instead of their cumulative sisters?

(d) Discuss the information content of a smooth density estimate compared to that of a histogram. Discuss also the use of a picture of $\widehat{f}$ as a data summary.

(e) Display in a figure three density estimates, corresponding to $h = .20$ (too ragged), $h = .636$ (which turns out to be the optimal value based on least squares cross validation, see Exercise 35), and $h = 1.50$ (too smooth). What do you think: Are there identifiable sub-groups in the distribution of (transformed) permeability?

Exercise No. 34

It can be demonstrated in various ways that the choice of a good smoothing parameter h in the kernel-type density estimator is much more crucial than the choice of kernel function $K(.)$. In the examples considered here the standard normal kernel $\phi(.)$ will be used.

In many cases where a nonparametric density estimate is needed it will suffice to choose h subjectively, by considering graphs of $\widehat{f}(x)$ for a small selection of h values, and then select one that matches prior beliefs (conscious or subconscious) about the phenomenon under study. This is obviously not quite satisfactory in general, however, and the present and

the following two exercises look into the tricky problem of obtaining a sensible, automatic, objective, data-based smoothing parameter h. They should be worked through only by persons who are willing to increase their sorrow:

For where there is much wisdom is much grief,
and he that increaseth his knowledge increaseth his sorrow.
— Ecclesiastes (The Preacher) 1:18 (c. 200 B.C.)

(a) Consider the integrated mean squared error

$$\text{IMSE} = E_f \int \{\widehat{f}(x) - f(x)\}^2 \, \mathrm{d}x,$$

as a function of h. Show that the leading terms are $\frac{1}{4}h^4 A + B/(nh) - \int f^2 \, \mathrm{d}x/n$, in which $B = B(K) = \int K(y)^2 \, \mathrm{d}y$ and $A = A(f) = \int \{f''(x)\}^2 \, \mathrm{d}x$, cf. Exercise 32. Note that A is a measure of the 'ruggedness' or 'roughness' of the underlying density.

(b) Minimise this approximation to IMSE w.r.t. h. Show that the optimal h becomes

$$h_0 = h_0(f) = \frac{(B/A)^{1/5}}{n^{1/5}}.$$

(c) And show that the corresponding minimum IMSE has leading terms

$$\text{minimum IMSE } = \frac{\frac{5}{4}B^{4/5}A^{1/5}}{n^{4/5}} - \frac{1}{n}\int f^2 \, \mathrm{d}x.$$

(d) Compute B for $K = \phi$ and A for $f = N(\mu, \sigma^2)$, and conclude that the best value to use if the underlying density is normal, is

$$h_0 = \frac{(4/3)^{1/5}\sigma(f)}{n^{1/5}} = \frac{1.059\,\sigma(f)}{n^{1/5}}.$$

(e) Find the best possible kernel $K(.)$: minimise the minimum IMSE w.r.t. the choice of K. This amounts to minimising $B = \int K(y)^2 \, \mathrm{d}y$ under the constraints $\int yK(y) \, \mathrm{d}y = 0$, $\int y^2 K(y) \, \mathrm{d}y = 1$. Compare the resulting minimum value of $B^{4/5}$ with what one gets from other natural kernels.

(f) It has been suggested that the $1.059\,\sigma(f)/n^{1/5}$ rule is relatively robust against departures from normality. Hence $h = 1.059\,\widehat{\sigma}/n^{1/5}$, with a robust standard deviation estimate, comes forward as a reasonable rule of thumb. Investigate this proposal, to some extent, by considering $f =$ a mixture of two normals. For such an f, what is $(B/A)^{1/5}$, compared to $1.059\,\sigma(f)$?

(e) Another data-based procedure arises naturally: Estimate $A(f)$ from the data, and use $\widehat{h} = (B/\widehat{A})^{1/5}/n^{1/5}$. Comment on this proposal.

Exercise No. 35

Here is another approach to choosing h: Consider the integrated squared error

$$\text{ISE} = \int \{\widehat{f}(x) - f(x)\}^2 \, \mathrm{d}x = \int \widehat{f}^2 \, dx - 2\int f\widehat{f} \, \mathrm{d}x + \int f^2 \, \mathrm{d}x.$$

The first term here is an explicit function of h, the second term can be estimated, as in Exercise 31, and the third term is not affected by data at all. A natural program is therefore to minimise an estimate of the first two terms w.r.t. h.

(a) With kernel function $K = \phi$, show that

$$\int \widehat{f}(x)^2 \, \mathrm{d}x = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi\Big(\frac{x_i - x_j}{\sqrt{2}h}\Big) \frac{1}{\sqrt{2}h}.$$

(b) Motivate the following estimator for $\lambda(h) = \int \widehat{f}^2 \, dx - 2 \int f \widehat{f} \, \mathrm{d}x$:

$$\widehat{\lambda}(h)^{\mathrm{CROSS}} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \phi\Big(\frac{x_i - x_j}{\sqrt{2}h}\Big) \frac{1}{\sqrt{2}h} - 2\widehat{\rho}(h)^{\mathrm{CROSS}},$$

where $\widehat{\rho}^{\mathrm{CROSS}}$ is given in Exercise 31. Find the expected value of $\widehat{\lambda}(h)^{\mathrm{CROSS}}$ and compare it to $E_f\mathrm{ISE} = \mathrm{IMSE}$.

(c) Think through other approaches to estimating $\lambda(h)$, including $\widehat{\lambda}(h)^{\mathrm{NAIVE}}$, $\widehat{\lambda}(h)^{\mathrm{BOOT}}$, and $\widehat{\lambda}(h)^{\mathrm{BOOTJACK}}$.

(d) Compute and display the curve $\widehat{\lambda}(h)^{\mathrm{CROSS}}$ for the transformed permeability data of the previous example. Find the minimising value of h, and draw the resulting density estimate curve $\widehat{f}(x)$. [Answer: The least squares cross validation curve is displayed on the following page, and is minimised by $h = .636$. The resulting least-squares optimal kernel-type density estimate is drawn, with a solid line, in the figure of Exercise 33.]

(e) Generalise the approach developed here to one appropriate for the loss function $\int \{\widehat{f}(x) - f(x)\}^2 w(x) \, \mathrm{d}x$, where $w(.)$ is a given weight function.

(f) Obtain a simulated 30-sample from $N(0, 1)$, and draw curves $\lambda(h)^{\mathrm{TRUE}}$, $\widehat{\lambda}(h)^{\mathrm{NAIVE}}$, and $\widehat{\lambda}(h)^{\mathrm{CROSS}}$.

Exercise No. 36

Kullback-Leibler instead. gamma(h). naive, cross, boot, jackboot. Figures. Quasi-data 30.

Exercise No. 37

Ten bootstraps of perm-data. Comments and speculations.

Exercise No. 38

Ten bootstraps of h-values, perm-data: h(LS) and h(KL).

Exercise No. 39

Other density estimation schemes. k-NN for example. Find now simply a discrete curve and an estimate.

Exercise No. 40

Silverman and Young. The ridge regression syndrom again?

Exercise No. 41

The purpose of the present exercise is to understand what happens with maximum likelihood estimators when the parametric model on which they are based is incorrect. The textbook treatment almost invariably stops with a discussion of what goes on when the idealised model is absolutely correct. Generalising this to the agnostic case is obviously of general importance, and will also help us understand, in later exercises, how parametric bootstrapping and nonparametric bootstrapping work for parametric models, including regression models.

The i.i.d. framework is as follows: The data points X_i come in reality from a distribution F, with density f, but are fitted to a p-dimensional parametric model $\{f_\theta : \theta \in \Theta\}$. Below you are asked for heuristic proofs which will work under sufficient regularity. The estimator considered is $\widehat{\theta} = \widehat{\theta}_{\mathrm{ML}}$, the maximum likelihood one. The same ideas can however be used to characterise the behaviour of also other estimators.

(a) Show that $\widehat{\theta}$ converges in probability to the parameter value $\theta_0 = \theta_0(F)$ that makes f_θ come closest to the true f as measured by the Kullback–Leibler distance, $\Delta(f, f_\theta) = E_F \log\{f(X)/f_\theta(X)\} = \int f \log(f/f_\theta)\,\mathrm{d}x$. See also Exercise No. 19. — A sufficient condition that is often easy to verify is concavity of the log likelihood: If $A_n(\theta)$ converges to $A(\theta)$ in probability (almost surely), for each θ, and $A_n(.)$ is concave, then the maximiser of $A_n(.)$ converges in probability (almost surely) to the maximiser of $A(.)$. See Andersen and Gill (*Annals of Statistics* 1982, Appendix II).

(b) Let

$$U_n(\theta) = \sum_{i=1}^n \frac{\partial}{\partial\theta} \log f_\theta(X_i) \quad \text{and} \quad I_n(\theta) = \sum_{i=1}^n \frac{\partial^2}{\partial\theta\partial\theta} \log f_\theta(X_i).$$

Show that

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \Big\{-\frac{1}{n} I_n(\widetilde{\theta})\Big\}^{-1} \frac{U_n(\theta_0)}{\sqrt{n}},$$

where $\widetilde{\theta}$ lies somewhere between the least false parameter θ_0 and its estimate $\widehat{\theta}$.

(c) And deduce that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d N_p(0, J^{-1}KJ^{-1}),$$

in which $J = J(\theta_0)$, $K = K(\theta_0)$, and

$$J(\theta) = -E_F \frac{\partial^2}{\partial\theta\partial\theta} \log f_\theta(X) = -\int \frac{\partial^2}{\partial\theta\partial\theta} \log f_\theta(x)\, f(x)\,\mathrm{d}x,$$

$$K(\theta) = \mathrm{VAR}_F \frac{\partial}{\partial\theta} \log f_\theta(X) = \int \Big(\frac{\partial}{\partial\theta} \log f_\theta(x)\Big)\Big(\frac{\partial}{\partial\theta} \log f_\theta(x)\Big)' f(x)\,\mathrm{d}x.$$

(d) Convince yourself that the above results constitute a generalisation of what you have seen in textbooks by considering the idealised case, in which $f = f_{\theta_0}$ for a certain true value θ_0. Show in particular that $J = K$ and that $J^{-1}KJ^{-1} = J^{-1}$ in this case.

(e) Rededuce the result of (c) using influence functions, cf. Exercises 19–22.

(f) Define

$$\widehat{J}_n = -\frac{1}{n} I_n(\widehat{\theta}) = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta\partial\theta} \log f_{\widehat{\theta}}(X_i),$$

$$\widehat{K}_n = \frac{1}{n} \sum_{i=1}^{n} \Big(\frac{\partial}{\partial\theta} \log f_{\widehat{\theta}}(X_i) \Big) \Big(\frac{\partial}{\partial\theta} \log f_{\widehat{\theta}}(X_i) \Big)'.$$

Show that these are consistent for J and K respectively. This suggests that the traditional estimator $\widehat{J}^{-1}/n$ of the covariance matrix for $\widehat{\theta}$ could be replaced by the model-robust $\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}/n$.

(g) See how the above results look like in some traditional parametric models.

(h) Consider generalising what you have seen here to regression models.

Exercise No. 42

Find the maximum likelihood estimator $\widehat{\theta}$ and its limiting distribution in each of the following parametric models. Find also the robust variance-covariance estimator $\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}$ and compare it with $\widehat{J}^{-1}$.

(a) $X_i \sim$ exponential $(1/\theta)$, i.e. $f_\theta(x) = (1/\theta)\exp(-x\theta), \quad x > 0.$

(b) $X_i \sim N(\xi, \sigma^2)$, both parameters unknown.

(c) $X_i \sim$ Gamma (α, β), i.e. $f_{\alpha,\beta}(x) = \big(\beta^\alpha/\Gamma(\alpha)\big)\, x^{\alpha-1}e^{-\beta x}, \quad x > 0.$

(d) Unlike virtue, courage is not its own reward: it brings results. Explore your own example.

Exercise No. 43

Back to bootstrapping: When analysing a parametric model there are at least two ways of bootstrapping. The *nonparametric bootstrap* draws random X_i^*'s from the nonparametric estimate of the model structure, i.e. from the usual empirical distribution $\widehat{F}$. The *parametric bootstrap*, on the other hand, trusts the model and draws $X_1^*, \ldots, X_n^*$ from the parametric estimate $f_{\widehat{\theta}}$. Let us see how these work in a specific and illuminating example, namely the exponential distribution with parameter $1/\theta$.

(a) The maximum likelihood estimator is of course $\widehat{\theta} = \bar{X}$. If the model is perfect, show that $\widehat{\theta}$ is distributed as $\theta\,\chi^2_{2n}/2n$, and that $\widehat{\theta}$ is approximately $N(\theta_0, \frac{1}{n}\theta_0^2)$. If the model is incorrect, show that $\widehat{\theta}$ is approximately $N(\theta_0, \frac{1}{n}\sigma_0^2)$, where θ_0 is the true mean and σ_0 is the true standard deviation.

(b) Consider parametric bootstrapping: Then $X_1^*, \ldots, X_n^*$ are drawn from the exponential $(1/\widehat{\theta})$ model. Show that $\widehat{\theta}^* \sim \widehat{\theta}\,\chi^2_{2n}/2n$ and that $\widehat{\theta}^*$ is approximately $N(\widehat{\theta}, \frac{1}{n}\widehat{\theta}^2)$, in the conditional situation given data. Conclude that parametric bootstrapping works fine when the model is correct but that it goes astray when the model is incorrect with $\sigma_0^2 \neq \theta_0^2$.

(c) And consider nonparametric bootstrapping: In this case $X_1^*, \ldots, X_n^*$ are drawn from $\widehat{F}$. Demonstrate that $\widehat{\theta}^*$ is approximately $N(\widehat{\theta}, \frac{1}{n}\widehat{\sigma}^2)$. Conclude that nonparametric bootstrapping should work fine both when the model is correct and when it is incorrect!

Exercise No. 44

Let us generalise. Assume only that i.i.d data $X_1, \ldots, X_n$ are fitted to the p-dimensional parametric family $\{f_\theta\}$ with the maximum likelihood method.

(a) If the model is correct, then $\widehat{\theta}$ is close to $N_p(\theta_0, \frac{1}{n}J^{-1})$. If the model is incorrect, then J^{-1} must be replaced by $J^{-1}KJ^{-1}$. Make sure you know the precise definitions of J and K here.

(b) Parametric bootstrapping: Draw $X_1^*, \ldots, X_n^*$ from $f_{\widehat{\theta}}$. The result is a $\widehat{\theta}^*$ that is close to $N_p(\widehat{\theta}, \frac{1}{n}\widehat{J}^{-1})$. Bootstrap analysis works under ideal model conditions but may go wrong outside the model. Specifically, show that the bootstrap estimate of the (true) standard deviation for $\widehat{\theta}_i$ usually will be inconsistent.

(c) But show that nonparametric bootstrapping works in either case: The distribution of $\widehat{\theta}^*$ is close to $N_p(\widehat{\theta}, \frac{1}{n}\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1})$.

Exercise No. 45

Does this mean that we always should use nonparametric bootstrapping, i.e. resample from the original data points, even in parametric models? Give some *pro*'s and *con*'s.

If the model is right, or reasonably right, then much can be lost in estimating efficiency when assessing variability. When estimating the variance of $\widehat{\theta}_i$, for example, both of the bootstrapping schemes give consistent estimates, but the nonparametric version may have much larger sampling variability than the parametric version.

(a) Show that the last point brought up roughly corresponds to studying the estimation efficiency of $\widehat{J}^{-1}\widehat{K}\widehat{J}^{-1}$ versus $\widehat{J}^{-1}$ for estimating

$$J = J(\theta_0) = -\int \Big[\frac{\partial^2}{\partial\theta\partial\theta}\log f_{\theta_0}(x)\Big] f_{\theta_0}(x)\,\mathrm{d}x,$$

under model conditions.

(b) In the exponential example, the two estimators for θ_0^2 are $\widehat{\sigma}^2$ (nonparametric) and $\widehat{\theta}^2$ (parametric). Show this, and show that $\operatorname{Var}\widehat{\sigma}^2 \doteq 8\theta_0^4/n$, $\operatorname{Var}\widehat{\theta}^2 \doteq 4\theta_0^4/n$ (uner model conditions). Transform this to a statement about sample sizes.

(c) In the normal (ξ, σ^2) example, show that the two estimators for the variance of $\widehat{\sigma}$ are $(\frac{1}{2} + \frac{1}{4}\widehat{\beta})\,\widehat{\sigma}^2$ (nonparametric) and $\frac{1}{2}\widehat{\sigma}^2$ (parametric). Here $\widehat{\beta} = V_n/\widehat{\sigma}^4 - 3$ is the sample kurtosis, with $V_n = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^4$. Show further that $\widehat{\sigma}^2$ has variance close to $2\sigma_0^4/n$ while $(1 + \frac{1}{2}\widehat{\beta})\widehat{\sigma}^2$ has variance close to $8\sigma_0^4/n$, under home turf conditions for the model.

(d) Invent another example.

— One should probably carry out both parametric and nonparametric bootstrapping, and scrutinise any significant differences that might result. The examples above show that some of these differences might be incidental and due to large sampling variability on the part of the nonparametric bootstrap.

Exercise No. 46

Consider inventing an asymptotic mathematical framework for evaluating the different estimation schemes, and the different bootstrapping schemes, which is able to reflect the important case of a 'reasonably but not exactly correct model'. Thus the true f is close to but not equal to the best-fitting f_θ.

One idea is the following. Take $f(x) = f_{\theta,\nu_n}(x)$, where $f_{\theta,\nu}(.)$ is an enlarged parametric family, with $f_\theta(.) = f_{\theta,\nu_0}(.)$. Let now ν_n be close to but not equal to ν_0, for example $\nu_n = \nu_0 + \delta/\sqrt{n}$. As an example, suppose data $X_1, \ldots, X_n$ come from a distribution which is Weibull and close to but not quite exponential, with

$$\Pr X_i \le x = 1 - \exp\Big[-(\theta x)^{\beta_n}\Big], \quad x > 0,$$

where $\beta_n = 1 + \delta/\sqrt{n}$.

In this framework, try to characterise the behaviour of the maximum likelihood estimator $\widehat{\theta}$, and try to sort out the differences in behaviour between parametric and non-parametric bootstrapping.

Exercise No. 47

Resampling methods in regression models is a very important subject, and a difficult one. There are several ways to carry out both jackknifing and bootstrapping for a given regression model, and the properties of each scheme, and the differences between schemes, are not sufficiently well understood at present.

To cover some ground, let us begin in this and the following three exercises by trying to understand how the ordinary estimation methods fare in general and specific regression models. Then one can go on to bootstrapping and jackknifing afterwards.

A reasonably general regression framework for data (X_i, Y_i) is as follows: A parametric model postulates that $Y \mid x \sim f_\theta(y|x)$, whereas the true conditional distribution is some unknown $f(y|x)$. The prime example would be $Y \mid x \sim N(x'\beta, \sigma^2)$ for a covariate vector x of length p, with $\theta = (\beta, \sigma)$ being $(p+1)$-dimensional. Observe that interest centres on the conditional distribution of Y given its associate x, and the distribution of X alone is not modelled at all. This distribution will nevertheless necessarily enter some of the evaluations and expressions below, and likewise the simultaneous distribution of (X, Y). We write the former in the form of a density $f(x)\,dx$ on its sample space and the latter in the form of $f(x, y)\,dxdy$.

(a) Show that the maximum likelihood estimator $\widehat{\theta}$, which maximises the observed x-conditional likelihood $\Pi_{i=1}^n f_\theta(y_i|x_i)$, is consistent for a certain least false parameter value θ_0. This best fitting value is the one that minimises the distance function

$$\Delta[f, f_\theta] = \int \Delta_x[f(.|x), f_\theta(.|x)]\, f(x)\,\mathrm{d}x,$$

in which $\Delta_x[f(.|x), f_\theta(.|x)] = \int f(y|x) \log\big(f(y|x)/f_\theta(y|x)\big)\,\mathrm{d}y$ happens to be the x-conditional Kullback–Leibler distance between the true and the modelled density for Y given x. — Note as in Exercise 41 (a) that concavity of the log likelihood is a simple sufficient condition for consistency of $\widehat{\theta}$ towards θ_0.

(b) Aiming at a result parallelling that of Exercise 41 (c), expose yourself to *(i)* art, *(ii)* the matrices

$$J_x(\theta) = -E_{f(.|x)} \frac{\partial^2}{\partial\theta\partial\theta} \log f_\theta(Y|x) = -\int \frac{\partial^2}{\partial\theta\partial\theta} \log f_\theta(y|x)\, f(y|x)\, \mathrm{d}y,$$

$$\begin{aligned} K_x(\theta) &= E_{f(.|x)} \Big(\frac{\partial}{\partial\theta} \log f_\theta(Y|x) \Big) \Big(\frac{\partial}{\partial\theta} \log f_\theta(Y|x) \Big)' \\ &= \int \Big(\frac{\partial}{\partial\theta} \log f_\theta(y|x) \Big) \Big(\frac{\partial}{\partial\theta} \log f_\theta(y|x) \Big)' f(y|x)\, \mathrm{d}y, \end{aligned}$$

$$J(\theta) = \mathrm{E}\, J_X(\theta) = \int J_x(\theta)\, f(x)\, \mathrm{d}x,$$

$$K(\theta) = \mathrm{E}\, K_X(\theta) = \int K_x(\theta)\, f(x)\, \mathrm{d}x.$$

Show that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \to_d N(0, J^{-1} K J^{-1}),$$

in which $J = J(\theta_0)$ and $K = K(\theta_0)$.

(c) Demonstrate that $J = K$ if the model holds true.

(d) Let

$$\widehat{J}_n = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial\theta\partial\theta} \log f_{\widehat{\theta}}(Y_i|X_i),$$

$$\widehat{K}_n = \frac{1}{n} \sum_{i=1}^n \Big(\frac{\partial}{\partial\theta} \log f_{\widehat{\theta}}(Y_i|X_i) \Big) \Big(\frac{\partial}{\partial\theta} \log f_{\widehat{\theta}}(Y_i|X_i) \Big)'.$$

Show that these are consistent for J and K.

(e) Explain how the results above may be used to construct model-robust confidence regions for the least false parameters of the model.

(f) Quick: See how the machinery above works in the traditional normal linear regression model. Study in particular the $N(\beta x, \sigma^2)$ and $N(\alpha + \beta x, \sigma^2)$ cases, where x is one-dimensional.

Exercise No. 48

Let us apply the method and results of the previous exercise to the traditional linear regression model. *The model* postulates that Y given a p-vector of covariate measurements $x = (x_1, \ldots, x_p)'$ is normal with mean $x'\beta = \sum_{j=1}^p \beta_j x_j$ and variance σ^2. Let *us*, conservatively and counterbalancedly, postulate only that $Y \mid x$ has some density $f(y|x)$. Define

$$L = \mathrm{E}\, XX' = \int xx'\, f(x)\, \mathrm{d}x \quad \text{and} \quad \beta_0 = L^{-1} \mathrm{E}\, XY = L^{-1} \int x y_0(x)\, f(x)\, \mathrm{d}x,$$

where $y_0(x) = E(Y|x) = \int y f(y|x)\, \mathrm{d}y$ is the true conditional expectation of Y given x. Define also

$$\sigma_0^2(x) = E(Y - x'\beta_0)^2 | x = \int (y - x'\beta_0)^2\, f(y|x)\, \mathrm{d}y.$$

Observe that this is not quite the true conditional variance of Y, it is rather equal to $\{y_0(x) - x'\beta_0\}^2 + \mathrm{Var}(Y|x)$.

(a) Find the maximum likelihood estimators:

$$\widehat{\beta} = \Big(\sum_{i=1}^n X_i X_i'\Big)^{-1} \sum_{i=1}^n X_i Y_i = \widehat{L}^{-1} \frac{1}{n} \sum_{i=1}^n X_i Y_i,$$

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - X_i'\widehat{\beta})^2.$$

Here the familiar matrix $\widehat{L} = \frac{1}{n}\sum_{i=1}^n X_i X_i'$ is the natural estimator for L.

(b) What are the least false or best fitting parameter values $\beta_{\text{l.f.}}$, $\sigma_{\text{l.f.}}$? Show that $\beta_{\text{l.f.}}$ is indeed equal to β_0 defined above, and can be defined as the value that minimises $E(Y - X'\beta)^2 = \int\int (y - x'\beta)^2 \, f(x,y)\,\mathrm{d}x\,\mathrm{d}y$. And show that the best fitting σ is σ_0 given by

$$\sigma_0^2 = \mathrm{E}\,\sigma_0^2(X) = \int \sigma_0^2(x)\, f(x)\,\mathrm{d}x.$$

(c) In the general notation of the previous exercise, show that

$$J = \frac{1}{\sigma_0^2} \begin{pmatrix} L & 0 \\ 0 & 2 \end{pmatrix} \quad \text{and} \quad K = \frac{1}{\sigma_0^2} \begin{pmatrix} M & a \\ a' & b^2 \end{pmatrix},$$

in which

$$M = \frac{1}{\sigma_0^2} \mathrm{E}\, XX'(Y - X'\beta_0)^2 = \int xx' \frac{\sigma_0^2(x)}{\sigma_0^2}\, f(x)\,\mathrm{d}x,$$

$$a = E\frac{(Y - X'\beta_0)^3}{\sigma_0^3} X, \quad \text{and} \quad b^2 = E\frac{(Y - X'\beta_0)^4}{\sigma_0^4} - 1.$$

(d) Infer that

$$\widehat{\beta} \text{ is asymptotically } \sim N_p(\beta_0, \tfrac{1}{n} L^{-1} M L^{-1}).$$

L is estimated by $\widehat{L}$. In this framework, where $E(Y - x'\beta_0)^2|x$ may depend upon x, we also need a consistent estimate of M. Show that

$$\widehat{M} = \frac{1}{\widehat{\sigma}^2} \frac{1}{n} \sum_{i=1}^n X_i X_i' (Y_i - X_i'\widehat{\beta})^2$$

is one such.

(e) In particular, the textbook method for making inference about the β_j coefficients, which is based on

$$\mathrm{VAR}\,\widehat{\beta} \doteq \widehat{\sigma}^2 \Big(\sum_{i=1}^n X_i X_i'\Big)^{-1},$$

is only valid in the variance-homogeneous case $\sigma_0^2(x) \equiv \sigma_0^2$.

(f) On the other hand you ought to admit that the textbook method, based on (i) exact expectation $\mathrm{E}\,(Y|x) = x'\beta_0$, (ii) exact variance homogeneity, and (iii) exact normality, is impressively robust (for large samples) against failures of (i) and (iii). Observe that the β_0 parameter has a perfectly acceptable statistical interpretation even without (i), cf. (a).

(g) Construct a model-robust confidence interval for σ with confidence coefficient close to 90%.

Exercise No. 49

It may be instructive to actually calculate the various least false parameters in a couple of constructed examples, in which the true probability mechanisms are known. Here is one such:

The model is the simple linear one-dimensinal normal regression, where Y given x is $N(\alpha+\beta x, \sigma^2)$. The true state of affairs, however, is as follows: The conditional expectation is $y_0(x) = 1 + x + cx^2$; the conditional variance is $\mathrm{Var}(Y|x) = v^2(1 + dx^2)$; and the conditional distribution is exactly normal. To find the relevant quantities exactly one also needs the distribution of x's, as made clear in the previous exercises. Take X to be standard normal.

Based on these assumptions, find explicitly the best fitting line (or the least false line) $\alpha_0 + \beta_0 x$; the standard deviation parameter σ_0; the limit distribution of $(\widehat{\alpha}, \widehat{\beta})$; and the limit distribution of $\widehat{\sigma}$. Give your answers in terms of the given constants c, d, v.

Construct an example yourself, in which the distribution of Y given x is taken to be something non-normal.

Exercise No. 50

Consider the logistic regression model, involving a 0–1 variable Y whose probability of being equal to 1 is thought to be influenced by covariate measurements $x_1, \ldots, x_p$, in the following way:

$$\Pr\{Y = 1 \mid x\} = \frac{\exp(\alpha + \sum_{j=1}^p \beta_j x_j)}{1 + \exp(\alpha + \sum_{j=1}^p \beta_j x_j)}.$$

The interpretation is that the relative proportion of $\{Y = 1\}$ events in a homogeneous subgroup of the population under study, where all individuals have the same x-vector, should be close to the right hand side, for a suitable set of $\alpha, \beta_1, \ldots, \beta_p$.

Of course one cannot expect such a relation to hold exactly, and we are lead to view it merely as an approximation to some true, underlying function $\Pr\{Y = 1 \mid x\} = q(x)$, say. Try to work out the analogue of Exercise 48 for the logistic regression model, once again employing the general machinery of Exercise 47. How can the least false parameters that the maximum likelihood estimators are aiming at, be characterised? What is the model-robust estimator of the covariance matrix of $\widehat{\beta}$? — For details and further comments, see Hjort (1988, NCC–report).

Exercise No. 51

Back to resampling again: There are several ways in which to carry out jackknifing and bootstrapping in regression models. Consider the ordinary linear regression model, with data (x_i, Y_i) that ideally satisfy $Y_i = x_i'\beta + \varepsilon_i$, where the unobservable residuals ε_i come from a distribution F with mean zero.

Here are four different bootstrapping schemes.

BOOT 1: View the $(p+1)$-tuples $(x_1, y_1), \ldots, (x_n, y_n)$ as having come from a common distribution G in $\mathcal{R}^{p+1}$. The nonparametric estimate of this distribution is $\widehat{G}$, placing equal mass $\frac{1}{n}$ on each $(p+1)$-tuple. Draw therefore a bootstrap sample $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ from $\widehat{G}$, i.e. from the n observed tuples, and compute

$$\widehat{\beta}^* = \Big(\sum_{i=1}^n X_i^* X_i^{*'}\Big)^{-1} \sum_{i=1}^n X_i^* Y_i^*$$

and other relevant bootstrap statistics from these.

BOOT 2: View y_i as having come from a signal term $x_i'\beta$ plus a noise term ε_i. Estimate the first by $x_i'\widehat{\beta}$ and simulate the second by drawing an ε_i^* from an estimate of their distribution. One possibility is the nonparametric $\widehat{F}$ which places equal mass $\frac{1}{n}$ on each estimated residual $\widehat{\varepsilon}_j = y_j - x_j'\widehat{\beta}$. The bootstrap sample on which to base this particular scheme's $\widehat{\beta}^*$ is therefore $(x_1, Y_1^*), \ldots, (x_n, Y_n^*)$, where $Y_i^* = x_i'\widehat{\beta} + \varepsilon_i^*$ and where the ε_i^*'s are drawn independently from $\widehat{F}$.

BOOT 3: The previous method is semi-parametric: It uses a parametrically estimated signal part and a nonparametrically estimated noise part. A fully parametric bootstrap method is the similar version that instead draws residuals ε_i^*'s from $N(0, \widehat{\sigma}^2)$.

BOOT 4: The first method is utterly nonparametric, and in a way the special characteristics of the regression situation, concerned with modelling the distribution of Y given its associate x, were lost. A smoother and more regressionistic but still nonparametric scheme is as follows: By some nonparametric smoothing method, estimate the conditional expectation $E(Y|x)$ by some $\widehat{y}_0(x)$. Hundreds of such methods have been discussed in the literature during the last ten years. Let $\widehat{\varepsilon}_i = y_i - \widehat{y}_0(x_i)$ this time, and let $Y_i^* = \widehat{y}_0(x_i) + \varepsilon_i^*$ with a simulated residual from the empirical or smoothed distribution of $\widehat{\varepsilon}_j$'s.

How do these schemes perform?

Exercise No. 52

52: Preliminaries. (a) Test for normality:

$$T_n = \frac{1}{n}\sum_{i=1}^n \frac{\big|X_{(i)} - \widehat{\xi} - \widehat{\sigma}\Phi^{-1}(u_i)\big|}{\widehat{\sigma}(\Phi^{-1})'(u_i)\, u_i(1-u_i)}.$$

Invariant. .50, .60, .70, .80, .90, .95, .99 points: .268, .283, .301, .323, .356, .388, .469; based on 1000 simulations. (b) $\tau = E|X - \mu|$, the MAD. Limit distibution. (c) Testing equality of two such. Refer to Hjort (1988, SJS) about choice of kappahat.

53: Dr. *Livingstone I. Presume*. Estimate of

$$\theta = \theta(F_1, F_2) = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1\sigma_2}}.$$

Nonsmooth. Six categories. Bootstrapping: from top.