Chapter 8

Model Assessment and Selection

We have defined various smoothers and nonparametric estimation techniques. In classical statistical theory we usually assume that the underlying model generating the data is in the family of models we are considering. For nonparametrics this assumptions is relaxed and asymptotic and finite sample bias and variance estimates are not always easy to find in closed form. In this Chapter we discuss some resampling methods that are commonly used to get approximations of bias, variance, confidence intervals, etc...

In particular we will look at the problem of choosing smoothing parameters. Remember how most of the smoothers we have defined have some parameter that controls the smoothness of the curve estimate. For kernel smoothers we defined the scale parameter, for local regression we defined the span or bandwidth, and for smoothing splines we had the penalty term. We will call all of these *the smoothing parameter* and denote it with λ. It should be clear from the context which of the specific smoothing parameters we are referring to.

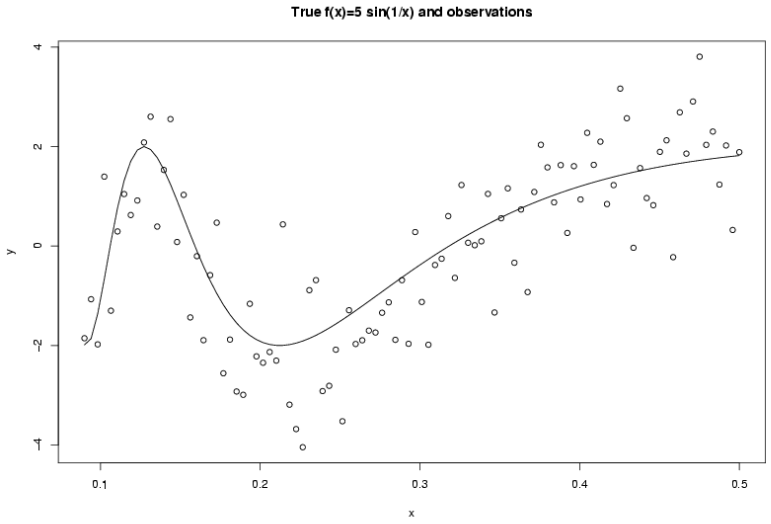


Figure 8.1: Outcomes of model with $f(x) = 5\sin(1/x)$ and IID normal errors with $\sigma^2 = 1$

8.1 The bias-variance trade-off

In smoothing in general there is a fundamental trade-off between the bias and variance of the estimate, and this trade-off is governed by the smoothing parameter.

Through out this section we will be using an artifical example defined by

$$y_i = 5\sin(1/x) + \epsilon_i, i = 1, \ldots, n \tag{8.1}$$

with the ϵ_i IID $N(0, 1)$ or t_3.

The trade-off is most easily seen in the case of the running mean smoother. The fitted running-mean smooth can be written as

$$\hat{f}_k(x_0) = \frac{1}{2k+1} \sum_{i \in N_k^S(x_0)} y_i$$

Under model (2.3). The variance is easy to compute. What is it?

The bias is

$$\mathrm{E}[\hat{f}_k(x_0)] - f(x_0) = \frac{1}{2k+1} \sum_{i \in N_k^S(x_0)} \{f(x_i) - f(x_0)\}$$

Notice that as k, in this case the smoothing parameter, grows the variances decreases. However, the bigger the k the more $f(x_i)$'s get into the bias.

We have no idea of what $\sum_{i \in N_k^S(x_0)} f(x_i)$ is because we don't know f! Let's see this in a more precise (not much more) way.

Say we think that f is smooth enough for us to assume that its second derivative $f''(x_0)$ is bounded. Taylor's theorem says we can write

$$f(x_i) = f(x_0) + f'(x_0)(x_i - x_0) + \frac{1}{2} f''(x_0)(x_i - x_0)^2 + o(|x_i - x_0|^2).$$

Because $\frac{1}{2} f''(x_0)(x_i - x_0)^2$ is $O(|x_i - x_0|^2)$ we stop being precise and write

$$f(x_i) \approx f(x_0) + f'(x_0)(x_i - x_0) + \frac{1}{2} f''(x_0)(x_i - x_0)^2.$$

Implicit here is the assumption that $|x_i - x_0|$ is small. This is the way these asymptotics work. We assume that the kernel size goes to 0 as n gets big.

Why did we only go up to the second derivative?

To makes things simple, let's assume that the covariates x are *equally spaced* and let $\Delta = x_{j+1} - x_j$ we can write

$$(2k+1)^{-1} \sum_{i \in N_k^S(x_0)} f(x_i) \approx f(x_0) + (2k+1)^{-1} \frac{k(k+1)}{6} f''(x_0) \Delta^2$$

So now we see that the bias increases with k^2 and the second derivative of the "true" function f. This agrees with our intuition.

Now that we have

$$\mathrm{E}\{\hat{f}_k(x_0) - f(x_0)\}^2 \approx \frac{\sigma^2}{2k+1} + \frac{k(k+1)}{6} f''(x_0) \Delta^2$$

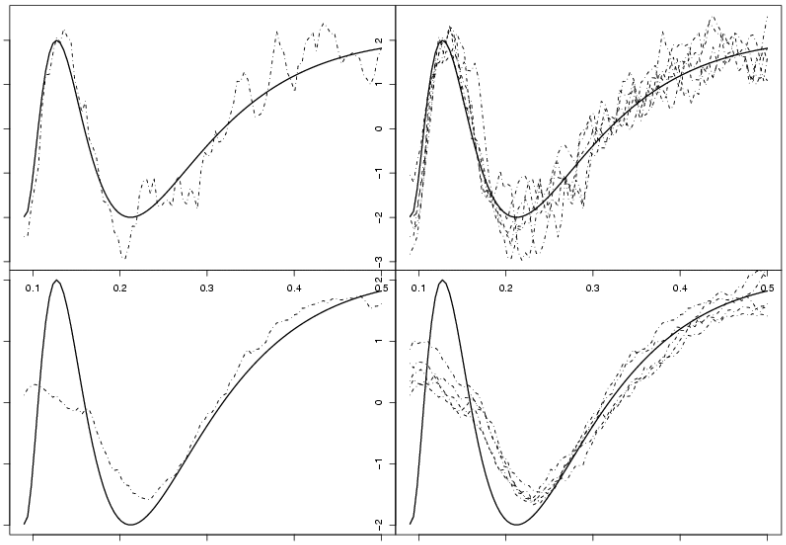


Figure 8.2: Smooths using running-mean smoother with bandwidths of .01 and 0.1. To the right are the smooths 25 replicates

we can actually find an optimal k

$$k_{opt} = \left\{ \frac{9\sigma^2}{2\Delta^4\{f''(x_i)\}^2} \right\}$$

Usually this is not useful in practice because we have no idea of what $f''(x)$ is like. So how do we chose smoothing parameters?

In Figure 8.2 we show the smooths obtained with a running mean smoother with bandwidths of 0.01 and 0.1 on 25 replicates defined by (8.1). The bias-variance trade-off can be clearly seen.

8.1.1 Bias-variance trade-off for linear smoothers

Define $\mathbf{S}_\lambda$ as the hat matrix for a particular smoother when the smoothing parameter λ is used. The "smooth" will be written as $\hat{\mathbf{f}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$.

Define

$$\mathbf{v}_\lambda = \mathbf{f} - \mathrm{E}(\mathbf{S}_\lambda \mathbf{y})$$

as the *bias* vector.

Define $\mathrm{ave}(\mathbf{x}^2) = n^{-1}\sum_{i=1}^n x_i^2$ for any vector $\mathbf{x}$. We can derive the following formulas:

$$\begin{aligned}
\mathrm{MSE}(\lambda) &= n^{-1}\sum_{i=1}^n \mathrm{var}\{\hat{f}_\lambda(x_i)\} + \mathrm{ave}(\mathbf{v}_\lambda^2) \\
&= n^{-1}\mathrm{tr}(\mathbf{S}_\lambda\mathbf{S}_\lambda')\sigma^2 + n^{-1}\mathbf{v}_\lambda'\mathbf{v}_\lambda \\
\mathrm{PSE}(\lambda) &= \{1 + n^{-1}\mathrm{tr}(\mathbf{S}_\lambda\mathbf{S}_\lambda')\}\sigma^2 + n^{-1}\mathbf{v}_\lambda'\mathbf{v}_\lambda.
\end{aligned}$$

Notice for least-squares regression $\mathbf{S}_\lambda$ is idempotent so that $\mathrm{tr}(\mathbf{S}_\lambda\mathbf{S}_\lambda') = \mathrm{tr}(\mathbf{S}_\lambda) = \mathrm{rank}(\mathbf{S}_\lambda)$ which is usually the number of parameters in the model. This is why we will sometimes refer to $\mathrm{tr}(\mathbf{S}_\lambda\mathbf{S}_\lambda')$ as the *equivalent number of parameters* or degrees of freedom of our smoother.

8.2 Cross Validation: Choosing smoothness parameters

In the section, and the rest of the class, we will denote with $\hat{f}_\lambda$ the estimate obtained using smoothing parameter λ. Notice that usually what we really have is the smooth $\hat{\mathbf{f}}_\lambda$.

We will use the model defined by (8.1). Figure 8.3 shows one outcome of this model with normal and t-distributed errors.

We are trying to find the λ that minimizes

$$\mathrm{MSE}(\lambda) = n^{-1}\sum_{i=1}^n \mathrm{E}[\hat{f}_\lambda(x_i) - f(x_i)]^2$$

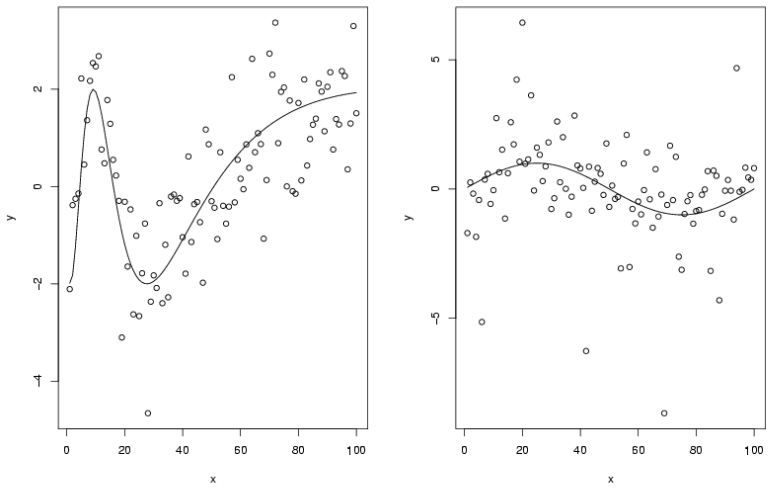


Figure 8.3: Outcomes of model (8.1)

Problem is we don't now f.

What if we could get a new set of data $y_1^*, \ldots, y_n^*$ from the same model producing the $y_1, \ldots, y_n$? This would be quite helpful because the *predictive squared error*

$$\text{PSE}(\lambda) = \text{E}[y_i^* - \hat{f}_\lambda(x_i)]^2 = \text{E}[\{y_i^* - f(x_i)\} - \{\hat{f}_\lambda(x_i) - f(x_i)\}] = \text{MSE}(\lambda) + \sigma^2.$$

says that $n^{-1}\sum_{i=1}^n [y_i^* - \hat{f}_\lambda(x_i)]^2$ is an average having expected value the MSE plus a constant. We could view this quantity as an estimate of MSE$(\lambda) + \sigma^2$. Since σ^2 doesn't depend on λ we could find the λ that minimizes it and think that we are close to the λ that minimizes the MSE.

Notice that the above calculation can be done because the y_i^*s are independent of the estimates $\hat{f}_\lambda(x_i)$s, the same can't be said about the y_is.

In practice it is not common to have a new set of data $y_i^*, i = 1, \ldots, n$. Cross-validation tries to imitate this by leaving out points (x_i, y_i) one at a time and estimating the smooth at x_i based on the remaining $n - 1$ points. The cross-

validation sum of squares is

$$\text{CV}(\lambda) = n^{-1} \sum_{i=1}^{n} \{y_i - \hat{f}_\lambda^{-i}(x_i)\}^2$$

where $\hat{f}_\lambda^{-i}(x_i)$ indicates the fit at x_i computed by leaving out the $i-th$ point.

We can now use CV to choose λ by considering a wide span of values of λ, computing CV(λ) for each one, and choosing the λ that minimizes it. Plots of CV(λ) vs. λ may be useful.

Why do we think this is good? First notice that

$$\begin{aligned} \text{E}\{y_i - \hat{f}_\lambda^{-i}(x_i)\}^2 &= \text{E}\{y_i - f(x_i) + f(x_i) - \hat{f}_\lambda^{-i}(x_i)\}^2 \\ &= \sigma^2 + \text{E}\{\hat{f}_\lambda^{-i}(x_i) - f(x_i)\}^2. \end{aligned}$$

Using the assumption that $\hat{f}_\lambda^{-i}(x_i) \approx \hat{f}_\lambda(x_i)$ we see that

$$\text{E}\{\text{CV}(\lambda)\} \approx \text{PSE}(\lambda)$$

However, what we really want is

$$\min_\lambda \text{E}\{\text{CV}(\lambda)\} \approx \min_\lambda \text{PSE}(\lambda)$$

but the law of large numbers says the above will do.

Why not simply use the averaged squared residuals

$$\text{ASR}(\lambda) = n^{-1} \sum_{i=1}^{n} \{y_i - \hat{f}_\lambda(x_i)\}^2?$$

It turns out this under-estimates the PSE. Notice in particular that the estimate $\hat{f}(x_i) = y_i$ always has ASR equal to 0! We will see how we can adjust the ASR to form "good" estimates of the MSE.

8.2.1 CV for linear smoothers

Now we will see some of the practical advantages of linear smoothers.

For linear smoothers in general it is not obvious what is meant by $\hat{f}_\lambda^{-i}(x_i)$. Let's give a definition...

Notice that any reasonable smoother will smooth constants into constants, i.e. $\mathbf{S1} = \mathbf{1}$. If we think of the rows $\mathbf{S}_{i\cdot}$ of $\mathbf{S}$ as weights of a kernels, this condition is requiring that all the n weights in each of the n kernels add up to 1. We can define $\hat{f}_\lambda^{-i}(x_i)$ as the "weighted average"

$$\mathbf{S}_{i\cdot}\mathbf{y} = \sum_{j=1}^{n} S_{ij} y_j$$

but giving zero weight to the ith entry, i.e.

$$\hat{f}_\lambda^{-i}(x_i) = \frac{1}{1 - S_{ii}} \sum_{j \neq i} S_{ij} y_j.$$

From this definition we can find CV without actually making all the computations again. Lets see how:

Notice that

$$\hat{f}_\lambda^{-i}(x_i) = \sum_{j \neq i} S_{ij} y_j + S_{ii} \hat{f}_\lambda^{-i}(x_i).$$

The quantities we add up to obtain CV are the squares of

$$y_i - \hat{f}_\lambda^{-i}(x_i) = y_i - \sum_{j \neq i} S_{ij} y_j - S_{ii} \hat{f}_\lambda^{-i}(x_i).$$

Adding and subtracting $S_{ii} y_i$ we get

$$y_i - \hat{f}_\lambda^{-i}(x_i) = y_i - \hat{f}_\lambda(x_i) + S_{ii}(y_i - \hat{f}_\lambda^{-i}(x_i))$$

which implies

$$y_i - \hat{f}_\lambda^{-i}(x_i) = \frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_{ii}}$$

and we can write

$$\mathrm{CV}(\lambda) = n^{-1} \sum_{i=1}^{n} \left\{ \frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_{ii}} \right\}^2$$

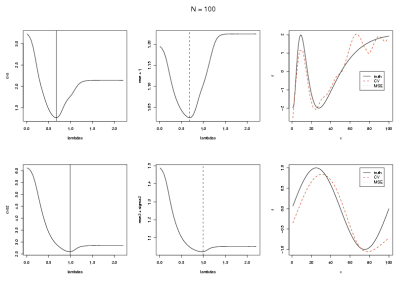


Figure 8.4: CV, MSE, and fits obtained for the normal and t models.

so we don't have to compute $\hat{f}_\lambda^{-i}(x_i)$!

Lets see how this definition of CV may be useful in finding the MSE.

Notice that the above defined CV is similar to the ASR except for the division by $1 - S_{ii}$. To see what this is doing we notice that in many situations $S_{ii} \approx [\mathbf{S}_\lambda \mathbf{S}_\lambda]_{ii}$ and $1/(1 - S_{ii})^2 \approx 1 + 2S_{ii}$ which implies

$$\text{E}[\text{CV}(\lambda)] \approx \text{PSE}(\lambda) + 2\text{ave}[\text{diag}(\mathbf{S}_\lambda)\mathbf{v}^2].$$

Thus CV adjusts ASR so that in expectation the variance term is correct but in doing so induces an error of $2S_{ii}$ into each of the bias components.

In Figure 8.4 we see the CV and MSE for $n = 100$ and $n = 500$ observatios

8.3 Model Selection

Suppose we observe a realization of a random variable Y, with distribution defined by a parameter $\boldsymbol{\beta}$

$$\prod_{\mathbf{x}_i \in N_0} f(y_i; \mathbf{x}_i, \boldsymbol{\beta}) \equiv f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) \tag{8.2}$$

where $\mathbf{y}$ is the observed response associated with the covariates $\mathbf{X}$ and $\boldsymbol{\beta} \in \mathbb{R}^P$ is a $P \times 1$ parameter vector.

We are interested in estimating $\boldsymbol{\beta}$. Suppose that before doing so, we need to choose from amongst P competing models, generated by simply restricting the general parameter space R^P in which $\boldsymbol{\beta}$ lies.

In terms of the parameters, we represent *the full model* with P parameters as:

$$\text{Model(P): } f_{\mathbf{Y}}(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}_P), \boldsymbol{\beta}_P = (\beta_1, \ldots, \beta_p, \beta_{p+1}, \ldots, \beta_P)'.$$

We denote the "true value" of the parameter vector $\boldsymbol{\beta}$ with $\boldsymbol{\beta}^*$.

Akaike (1977) formulates the problem of statistical model identification as one of selecting a model $f_{\mathbf{Y}}(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}_p)$ based on the observations from that distribution, where the particular restricted model is defined by the constraint $\beta_{p+1} = \beta_{p+2} = \ldots = \beta_P = 0$, so that

$$\text{Model(p): } f_{\mathbf{Y}}(\mathbf{y}; \mathbf{x}, \boldsymbol{\beta}_p), \boldsymbol{\beta}_p = (\beta_1, \ldots, \beta_p, 0, \ldots, 0)' \tag{8.3}$$

We will refer to p as the *number of parameters* and to Ω_p as the sub-space of $\mathbb{R}^P$ defined by restriction (8.2). For each $p = 1, \ldots, P$, we may assume model(p) to estimate the non-zero components of the vector $\boldsymbol{\beta}^*$. We are interested in a criterion that helps us chose amongst these P competing estimates.

In this Chapter we consider 3 methods for model selection.

8.3.1 Mallow's C_p

Mallow's C_p is a technique for model selection in regression (Mallows 1973). The C_p statistic is defined as a criteria to assess fits when models with different numbers of parameters are being compared. It is given by

$$C_p = \frac{\text{RSS}(p)}{\sigma^2} - N + 2p \tag{8.4}$$

If model(p) is correct then C_p will tend to be close to or smaller than p. Therefore a simple plot of C_p versus p can be used to decide amongst models.

In the case of ordinary linear regression, Mallow's method is based on estimating the mean squared error (MSE) of the estimator $\hat{\boldsymbol{\beta}}_p = (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{Y}$,

$$\mathrm{E}[\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}]^2$$

via a quantity based on the residual sum of squares (RSS)

$$\begin{aligned}\mathrm{RSS}(p) &= \sum_{n=1}^{N}(y_n - \mathbf{x}_n\hat{\boldsymbol{\beta}}_p)^2 \\ &= (\mathbf{Y} - \mathbf{X}_p\hat{\boldsymbol{\beta}}_p)'(\mathbf{Y} - \mathbf{X}_p\hat{\boldsymbol{\beta}}_p) \\ &= \mathbf{Y}'(\mathbf{I}_N - \mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p')\mathbf{Y}\end{aligned}$$

Here $\mathbf{I}_N$ is an $N \times N$ identity matrix. By using a result for quadratic forms, presented for example as Theorem 1.17 in Seber's book, page 13, namely

$$\mathrm{E}[\mathbf{Y}'\mathbf{A}\mathbf{Y}] = \mathrm{E}[\mathbf{Y}']\mathbf{A}\mathrm{E}[\mathbf{Y}] + \mathrm{tr}[\boldsymbol{\Sigma}\mathbf{A}]$$

$\boldsymbol{\Sigma}$ being the variance matrix of $\mathbf{Y}$, we find that

$$\begin{aligned}\mathrm{E}[\mathrm{RSS}(p)] &= \mathrm{E}[\mathbf{Y}'(\mathbf{I}_N - \mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p')\mathbf{Y}] \\ &= \mathrm{E}[\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}]^2 + \mathrm{tr}\left[\mathbf{I}_N - \mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\right]\sigma^2 \\ &= \mathrm{E}[\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}]^2 + \sigma^2\left(N - \mathrm{tr}\left[(\mathbf{X}_p'\mathbf{X}_p)(\mathbf{X}_p'\mathbf{X}_p)^{-1}\right]\right) \\ &= \mathrm{E}[\hat{\boldsymbol{\beta}}_p - \boldsymbol{\beta}]^2 + \sigma^2(N - p)\end{aligned}$$

where N is the number of observations and p is the number of parameters. Notice that when the true model has p parameters $\mathrm{E}[\mathrm{C}_p] = p$. This shows why, if model(p) is correct, C_p will tend to be close to p.

One problem with the C_p criterion is that we have to find an appropriate estimate of σ^2 to use for all values of p.

C_p for smoothers

A more direct way of constructing an estimate of PSE is to correct the ASR. It is easy to show that

$$\mathrm{E}\{\mathrm{ASR}(\lambda)\} = \left\{1 - n^{-1}\mathrm{tr}(2\mathbf{S}_\lambda - \mathbf{S}_\lambda\mathbf{S}_\lambda')\right\}\sigma^2 + n^{-1}\mathbf{v}_\lambda'\mathbf{v}_\lambda$$

notice that

$$\text{PSE}(\lambda) - \text{E}\{\text{ASR}(\lambda)\} = n^{-1}2\text{tr}(\mathbf{S}_\lambda)\sigma^2$$

This means that if we knew σ^2 we could find a "corrected" ASR

$$\text{ASR}(\lambda) + 2\text{tr}(\mathbf{S}_\lambda)\sigma^2$$

with the right expected value.

For linear regression $\text{tr}(\mathbf{S}_\lambda)$ is the number of parameters so we could think of $2\text{tr}(\mathbf{S}_\lambda)\sigma^2$ as a penalty for large number of parameters or for un-smooth estimates.

How do we obtain an estimate for σ^2? If we had a λ^* for which the bias is 0, then the usual unbiased estimate is

$$\frac{\sum_{i=1}^n \{y_i - f_{\lambda^*}(x_i)\}^2}{n - \text{tr}(2\mathbf{S}_{\lambda^*} - \mathbf{S}_{\lambda^*}\mathbf{S}'_{\lambda^*})}$$

The usual trick is to chose one a λ^* that does little smoothing and consider the above estimate. Another estimate that has been proposed it the first order difference estimate

$$\frac{1}{2(n-1)}\sum_{i=1}^{n-1}(y_{i+1} - y_i)^2$$

Once we have an estimate $\hat{\sigma}^2$ then we can define

$$C_p = \text{ASR}(\lambda) + n^{-1}2\text{tr}(S_\lambda)\hat{\sigma}^2$$

Notice that the p usually means number of parameters so it should be C_λ.

Notice this motivates a definition for degrees of freedoms.

8.3.2 Information Criteria

In this section we review the concepts behind Akaike's Information Criterion (AIC).

Akaike's original work is for IID data, however it is extended to a regression type setting in a straight forward way. Suppose that the conditional distribution of Y given $\mathbf{x}$ is know except for a P-dimensional parameter $\boldsymbol{\beta}$. In this case, the probability density function of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ can be written as

$$f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) \equiv \prod_{i=1}^{n} f(y_i; \mathbf{x}_i, \boldsymbol{\beta}) \tag{8.5}$$

with $\mathbf{X}$ the design matrix with rows $\mathbf{x}_i$.

Assume that there exists a true parameter vector $\boldsymbol{\beta}^*$ defining a true probability density denoted by $f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*)$. Given these assumptions, we wish to select $\boldsymbol{\beta}$, from one of the models defined as in (8.2), "nearest" to the true parameter $\boldsymbol{\beta}^*$ based on the observed data $\mathbf{y}$. The principle behind Akaike's criterion is to define "nearest" as the model that minimizes the Kullback-Leibler Information Quantity

$$\Delta(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{\beta}) = \int \{\log f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*) - \log f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta})\} \, f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*) \, d\mathbf{y}. \tag{8.6}$$

The analytical properties of the Kullback-Leibler Information Quantity are discussed in detail by Kullback (1959) . Two important properties for Akaike's criterion are

1. $\Delta(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{\beta}) > 0$ if $f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*) \neq f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta})$
2. $\Delta(\boldsymbol{\beta}^*; \mathbf{X}, \boldsymbol{\beta}) = 0$ if and only if $f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*) = f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta})$

almost everywhere on the range of $\mathbf{Y}$. The properties mentioned suggest that finding the model that minimizes the Kullback-Leibler Information Quantity is an appropriate way to choose the "nearest" model.

Since the first term on the right hand side of (8.5) is constant over all models we consider, we may instead maximize

$$H(\boldsymbol{\beta}) \;\; = \;\; \int \log f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}) f_{\mathbf{Y}}(\mathbf{y}; \mathbf{X}, \boldsymbol{\beta}^*) \, d\mathbf{y}$$

$$= \sum_{i=1}^{n} \int \log f(y_i; \mathbf{X}, \boldsymbol{\beta})\, f(y_i; \mathbf{x}_i, \boldsymbol{\beta}^*)\, dy_i. \tag{8.7}$$

Let $\hat{\boldsymbol{\beta}}_p$ be the maximum likelihood estimate under Model(p). Akaike's procedure for model selection is based on choosing the model which produces the estimate that maximizes $\mathrm{E}_{\boldsymbol{\beta}^*}\left[H(\hat{\boldsymbol{\beta}}_p)\right]$ amongst all competing models. Akaike then derives a criterion by constructing an asymptotically unbiased estimate of $\mathrm{E}_{\boldsymbol{\beta}^*}\left[H(\hat{\boldsymbol{\beta}}_p)\right]$ based on the observed data.

Notice that $H(\hat{\boldsymbol{\beta}}_p)$ is a function, defined by (8.6), of the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_p$, which is a random variable obtained from the observed data. A natural estimator of its expected value (under the true distribution of the data) is obtained by substituting the empirical distribution of the data into (8.6) resulting in the log likelihood equation evaluated at the maximum likelihood estimate under model(p)

$$l(\hat{\boldsymbol{\beta}}_p) = \sum_{i=1}^{n} \log f(y_i; \mathbf{x}_i, \hat{\boldsymbol{\beta}}_p).$$

Akaike noticed that in general $l(\hat{\boldsymbol{\beta}}_p)$ will overestimate $\mathrm{E}_{\boldsymbol{\beta}^*}\left[H(\hat{\boldsymbol{\beta}})\right]$. In particular Akaike found that under some regularity conditions

$$E_{\boldsymbol{\beta}^*}\left[l(\hat{\boldsymbol{\beta}}_p) - H(\hat{\boldsymbol{\beta}}_p)\right] \approx p.$$

This suggests that larger values of p will result in smaller values of $l(\hat{\boldsymbol{\beta}}_p)$, which may be incorrectly interpreted as a "better" fit, regardless of the true model. We need to "penalize" for larger values of p in order to obtain an unbiased estimate of the "closeness" of the model. This fact leads to the Akaike Information Criteria which is a bias-corrected estimate given by

$$\mathrm{AIC}(p) = -2l(\hat{\boldsymbol{\beta}}_p) + 2p. \tag{8.8}$$

See, for example, Akaike (1973) and Bozdogan (1987) for the details.

8.3.3 Posterior Probability Criteria

Objections have been raised that minimizing Akaike's criterion does not produce asymptotically consistent estimates of the correct model Notice that if we consider Model(p^*) as the correct model then we have for any $p > p*$

$$\Pr\left[AIC(p) < AIC(p^*)\right] = \Pr\left[2\{l(\hat{\boldsymbol{\beta}}_p) - l(\hat{\boldsymbol{\beta}}_{p^*})\} > 2(p - p^*)\right]. \qquad (8.9)$$

Notice that, in this case, the random variable $2\{l(\hat{\boldsymbol{\beta}}_p) - l(\hat{\boldsymbol{\beta}}_{p^*})\}$ is the logarithm of the likelihood ratio of two competing models which, under certain regularity conditions, is known to converge in distribution to $\chi^2_{p-p^*}$, and thus it follows that the probability in Equation (8.8) is not 0 asymptotically. Some have suggested multiplying the penalty term in the AIC by some increasing function of n, say $a(n)$, that makes the probability

$$\Pr\left[2\{l(\hat{\boldsymbol{\beta}}_p) - l(\hat{\boldsymbol{\beta}}_{p^*})\} > 2a(n)(p - p^*)\right]$$

asymptotically equal to 0. There are many choices of $a(n)$ that would work in this context. However, some of the choices made in the literature seem arbitrary.

Schwarz (1978) and Kashyap (1982) suggest using a Bayesian approach to the problem of model selection which, in the IID case, results in a criterion that is similar to AIC in that it is based on a penalized log-likelihood function evaluated at the maximum likelihood estimate for the model in question. The penalty term in the Bayesian Information Criteria (BIC) obtained by Schwarz (1978) is the AIC penalty term p multiplied by the function $a(n) = \frac{1}{2}\log(N)$.

The Bayesian approach to model selection is based on maximizing the posterior probabilities of the alternative models, given the observations. To do this we must define a strictly positive prior probability $\pi_p = \Pr[\text{Model}(p)]$ for each model and a conditional prior $d\mu_p(\boldsymbol{\beta})$ for the parameter given it is in Ω_p, the subspace defined by Model(p). Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be the response variable and define the distribution given $\boldsymbol{\beta}$ following (8.4)

$$f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) \equiv \prod_{i=1}^{n} f(y_i; \mathbf{x}_i, \boldsymbol{\beta})$$

The posterior probability that we look to maximize is

$$\Pr\left[\text{Model}(p)|\mathbf{Y}=\mathbf{y}\right] = \frac{\int_{\Omega_p} \pi_p f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X},\boldsymbol{\beta}) d\mu_p(\boldsymbol{\beta})}{\sum_{q=1}^{P} \int_{\Omega_q} \pi_q f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X},\boldsymbol{\beta}) d\mu_q(\boldsymbol{\beta})}$$

Notice that the denominator depends neither on the model nor the data, so we need only to maximize the numerator when choosing models.

Schwarz (1978) and Kashyap (1982) suggest criteria derived by taking a Taylor expansion of the log posterior probabilities of the alternative models. Schwarz (1978) presents the following approximation for the IID case

$$\log \int_{\Omega_p} \pi_p f_{\mathbf{Y}}(\mathbf{y}|\mathbf{X},\boldsymbol{\beta}) d\mu_p(\boldsymbol{\beta}) \approx l(\hat{\boldsymbol{\beta}}_p) - \frac{1}{2} p \log n$$

with $\hat{\boldsymbol{\beta}}_p$ the maximum likelihood estimate obtained under Model(p).

This fact leads to the Bayesian Information Criteria (BIC) which is

$$\text{BIC}(p) = -2l(\hat{\boldsymbol{\beta}}_p) + p \log n \tag{8.10}$$

Kyphosis Example

The AIC and BIC obtained for the gam are:

```
AIC(Age) =  83              BIC(Age) = 90
AIC(Age,Start) = 64         BIC(Age,Start) = 78
AIC(Age,Number) = 73        BIC(Age,Number) = 86
AIC(Age,Start,Number) = 60  BIC(Age,Start,Number) = 81
```

8.4 Bootstrap Standard Errors and Confidence Sets

Statistical science is the science of learning from experience. Efron and Tibshirani (1993) say “Most people are not natural-born statisticians. Left to our own devices

we are not very good at picking out patterns from a sea of noisy data. To put it another way, we are all too good at picking out non existing patterns that happen to suit our purposes."

Suppose we find ourselves in the following common data-analytic situation: a random sample $\mathbf{x} = (x_1, \ldots, x_n)$ from an unknown probability distribution F has been observed and we wish to estimate a parameter of interest $\theta = t(F)$ on the basis of $\mathbf{x}$. For this purpose, we calculate an estimate $\hat{\theta} = s(\mathbf{x})$ from $\mathbf{x}$.

A common estimate is the *plug-in* estimate $t(\hat{F})$ where $\hat{F}$ is the empirical distribution defined by

$$F(x) = \frac{\text{number of values in } \mathbf{x} \text{ equal to } x}{n}$$

Can you think of a plug-in estimate that is commonly used?

The bootstrap was introduced by Efron (1979) as a computer based method to estimate the standard deviation of $\hat{\theta}$.

What are the advantages:

- It is completely automatic
- Requires no theoretical calculations
- Not based on asymptotic results
- Available no matter how complicated the estimator $\hat{\theta}$ is.

A bootstrap sample is defined to be a random sample of size n drawn from $\hat{F}$, say $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$.

For each bootstrap sample $\mathbf{x}^*$ there is a bootstrap replicate of $\hat{\theta}$,

$$\hat{\theta}^* = s(\mathbf{x}^*).$$

The bootstrap estimate of $\mathbf{se}_F(\hat{\theta})$ is defined by

$$\mathbf{se}_{\hat{F}}(\hat{\theta}^*). \tag{8.11}$$

This is called the *ideal bootstrap estimate* of the standard error of $s(\mathbf{x})$.

Notice that for the case where θ is the expected value or mean of $\mathbf{x}_1$ we have

$$\mathbf{se}_{\hat{F}}(\bar{x}^*) = \mathbf{se}_{\hat{F}}(x_1^*)/\sqrt{n} = \sqrt{n^{-1}\sum_{i=1}^{n}(x_i - \hat{x})^2}/\sqrt{n}$$

and the ideal bootstrap estimate is the estimate we are used to. However, for any other estimator other than the mean obtaining (8.10) there is no neat formula that enables us to compute a numerical value in practice.

The bootstrap algorithm is a computational way of obtaining a good approximation to the numerical value of (8.10).

8.4.1 The bootstrap algorithm

The bootstrap algorithm works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the standard error of $\hat{\theta}$ by the empirical standard error, denoted by $\hat{\mathbf{se}}_B$, where B is the number of bootstrap samples used.

1. Select B independent bootstrap samples $\mathbf{x}_1^*, \ldots, \mathbf{x}_B^*$, each consisting of n data values drawing with replacement from $\mathbf{x}$.
2. Evaluate the bootstrap replication corresponding to each bootstrap sample
$$\hat{\theta}^*(b) = s(\mathbf{x}_b^*), b = 1, \ldots, B$$
3. Estimate the standard error $\mathbf{se}_F(\hat{\theta})$ by the sample standard error of the B replicates
$$\hat{\mathbf{se}}_B = \left[\frac{1}{B-1}\sum_{b=1}^{B}\{\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)\}^2\right]$$

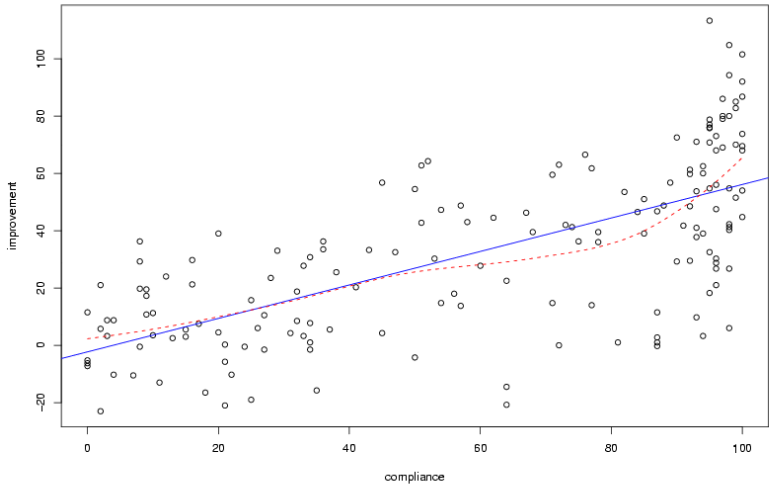


Figure 8.5: Estimated regression curves of Improvement on Compliance.

with

$$\hat{\theta}^*(\cdot) = B^{-1} \sum_{b=1}^{B} \hat{\theta}^*(b)$$

The limit of $\hat{\mathbf{se}}_B$ as B goes to infinity is the ideal bootstrap estimate of (8.10). But how close is (8.10) to $\mathbf{se}_F(\hat{\theta})$? See Efron and Tibshirani (1993) for more details.

8.4.2 Example: Curve fitting

In this example we will be estimating regression functions in two ways, by a standard least-squares line and by loess.

A total of 164 mean took part in an experiment to see if the drug cholostyramine lowered blood cholesterol levels. The men were supposed to take six packets of cholostyramine per day, but many of them actually took much less. Figure 8.5

shows compliance plotted against percentage of the intended dose actually taken. We also show a fitted line and a loess fit (using span=2/3). Notice the curves similar from 0 to 60, a little different from 60 to 80 and quite different from 80 to 100.

Assume the points a regression model

$$y_i = f(x_i) + \varepsilon_i, i = 1, \ldots, n$$

with the ε_i IID.

Say we are interested in the difference in rate of change of $f(x)$ in the 60–80 and 80–100 sections. We could define as the parameter to describe this. How can we do this?

Notice that finding a standard error for this estimate is not straight-forward. We can use the bootstrap.

Table 8.1: Estimates and bootstrap standard errors of $f(60)$, $f(80)$, and $f(100)$.

	$\hat{f}_{\text{line}}(60)$	$\hat{f}_{\text{line}}(80)$	$\hat{f}_{\text{line}}(100)$	$\hat{f}_{\text{loess}}(60)$	$\hat{f}_{\text{loess}}(80)$	$\hat{f}_{\text{loess}}(100)$
value:	33	44	56	28	35	66
$\hat{\text{se}}_{50}$:	2	2	3	5	4	4

As seen in Figure 8.6. Even when there is no parameter of interest, the bootstrap estimates of f give us an idea of what a confidence set is for the nonparametric estimates. We will see more of this in Chapter 7 and 8.

8.4.3 Confidence "intervals" for linear smoothers

It is easy to show that the variance-covariance matrix of the vector of fitted values $\hat{\mathbf{f}} = \mathbf{S}\mathbf{y}$ is

$$\mathbf{cov}(\hat{\mathbf{f}}) = \mathbf{S}\mathbf{S}'\sigma^2$$

and given an estimate of σ^2 this can be used to give point-wise standard errors, mainly by looking at $\text{diag}(\mathbf{S}\mathbf{S}')\sigma^2$.

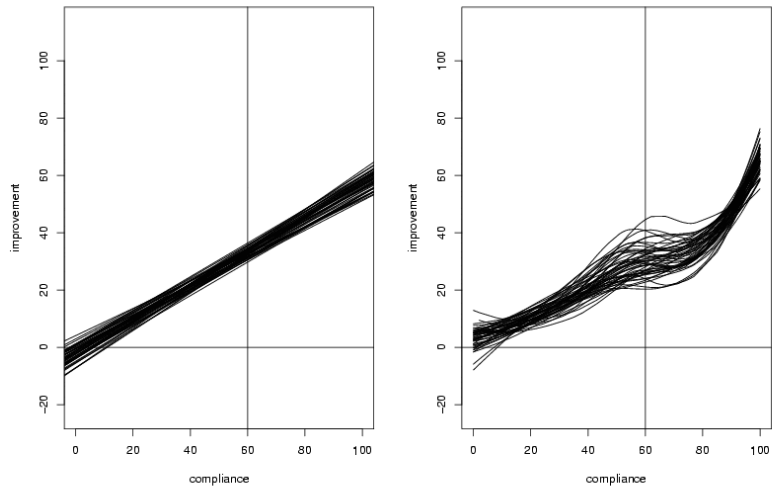


Figure 8.6: 50 boostrap curves for each estimation techinique.

Can we construct confidence intervals? What do we need?

First of all we need to know the distribution (at least approximately) of $\hat{\mathbf{f}}$. If the errors are normal we know that $\mathbf{f}$ is normally distributed. Why?

In the normal case, what are the confidence intervals for?

Remember that our estimates are usually biased, $\mathrm{E}(\hat{\mathbf{f}}) = \mathbf{Sf} \neq \mathbf{f}$. If our null hypothesis is $\mathbf{Sf} = \mathbf{f}$ (in the case of splines this is equivalent to assuming $f \in \mathcal{G}$) then our confidence intervals are for $\mathbf{f}$ otherwise it is much more convinient to compute them for $\mathbf{Sf}$. We will start using the notation $\tilde{\mathbf{f}} = \mathbf{Sf}$. We can think of $\tilde{\mathbf{f}}$ as the best possilble approximation to "the truth" $\mathbf{f}$ when using the $\mathbf{S}$ as a smoother.

To see how point-wise estimates can be useful, notice that we can get an idea of how variable $\hat{\mathbf{f}}(x_0)$ is. However, it isn't very helpful when we want to see how variable $\hat{\mathbf{f}}$ is as a whole.

What if we want to know if a certain function, say a line, is in our "confidence interval"? Point-wise intervals don't really help us with this.

8.4.4 Global confidence bands

Remember that $\hat{\mathbf{f}} \in \mathbb{R}^n$. This means that talking about confidence intervals doesn't make much sense. We need to consider confidence sets.

For example if the errors are normal we know that

$$\chi(\tilde{\mathbf{f}}) = (\hat{\mathbf{f}} - \tilde{\mathbf{f}})'(\mathbf{S}\mathbf{S}'\sigma^2)^{-1}(\hat{\mathbf{f}} - \tilde{\mathbf{f}})$$

is χ^2_n distributed. This permits us to construct confidence sets (which you can think of as random n-dimensional balls) for $\tilde{\mathbf{f}}$ of probability α

$$C_\alpha = \{\mathbf{g} \in \mathbb{R}^b; \chi(\mathbf{g}) \leq \chi_{1-\alpha}\} = \{\mathbf{g} \in \mathbb{R}^b; (\hat{\mathbf{f}} - \mathbf{g})'(\mathbf{S}\mathbf{S}'\sigma^2)^{-1}(\hat{\mathbf{f}} - \mathbf{g}) \leq \chi_{1-\alpha}\}.$$

Notice that the probability that the random ball doesn't fall on the approximate truth $\tilde{\mathbf{f}}$ is α:

$$\Pr(\tilde{\mathbf{f}} \notin C_\alpha) = \Pr\left[(\hat{\mathbf{f}} - \tilde{\mathbf{f}})'(\mathbf{S}\mathbf{S}'\sigma^2)^{-1}(\hat{\mathbf{f}} - \tilde{\mathbf{f}}) > \chi_{1-\alpha}\right] = \alpha.$$

This is only the case if we know σ^2.

Usually we construct an estimate

$$\hat{\sigma}^2 = (\mathbf{y} - \hat{\mathbf{f}})'(\mathbf{y} - \hat{\mathbf{f}})/\{n - \text{tr}(2\mathbf{S} - \mathbf{S}\mathbf{S}')\}$$

and define confidence sets

$$C(\tilde{\mathbf{f}}) = \{\mathbf{g} \in \mathbb{R}^b; \nu(\mathbf{g}) \leq G_{1-\alpha}\}$$

based on

$$\nu(\tilde{\mathbf{f}}) = (\hat{\mathbf{f}} - \tilde{\mathbf{f}})'(\mathbf{S}\mathbf{S}'\hat{\sigma}^2)^{-1}(\hat{\mathbf{f}} - \tilde{\mathbf{f}}).$$

Here $G_{1-\alpha}$ is the $(1-\alpha)$th quantile of the distribution of $\nu(\tilde{\mathbf{f}})$.

Do we know G? Not necessarily.

In the case of linear regression, where the gaussian model is correct and $\mathbf{S}$ is a p-dimensional projection, $\nu(\tilde{\mathbf{f}}) = \nu(\mathbf{f})$ has distribution $(n-p) + pF_{p,n-p}$.

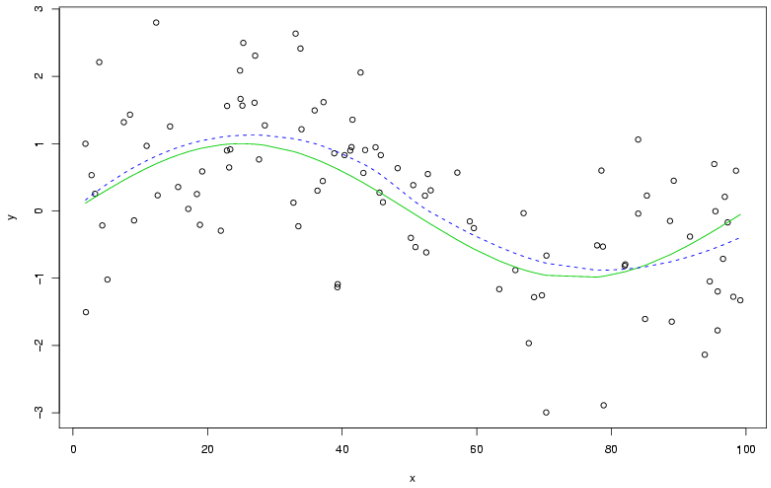
When this is not the case we can argue that the distribution is approximately

$$\{n - \text{tr}(2\mathbf{S} - \mathbf{S}\mathbf{S}')\} + \text{tr}(\mathbf{S}\mathbf{S}')F_{\text{tr}(\mathbf{S}\mathbf{S}'),n-\text{tr}(2\mathbf{S}-\mathbf{S}\mathbf{S}')}$$

If we are not sure of the normality assumption or that $\tilde{\mathbf{f}} \approx \mathbf{f}$ we can use the bootstrap to construct an approximate distribution $\hat{G}$ of G.

How do we do it?

Figure 8.7: The regression curve and an outcome with $n = 100$ and $\sigma^2 = 1$.



8.4.5 Bootstrap estimate of $G_{1-\alpha}$

A bootstrap sample is generated in the following way

- For some data $\mathbf{y}$ use some procedure (a linear smoother for example) to obtain an estimate $\hat{\mathbf{f}}$ of some estimand (in this case the regression function $\mathbf{f}$).
- Obtain residuals $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{f}}$.
- Take a simple random sample of size B from the residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n$. Notice that this makes them IID just like the εs.
- Construct a "new" data set

$$\mathbf{y}^* = \hat{\mathbf{f}} + \hat{\boldsymbol{\varepsilon}}^*$$

  with $\hat{\boldsymbol{\varepsilon}}^*$ the vector of resampled residuals.
- From the new data form a new estimate $\hat{\mathbf{f}}^*$.
- Finally we obtain the value of

$$\nu^* = (\hat{\mathbf{f}}^* - \hat{\mathbf{f}})'(\mathbf{S}\mathbf{S}'\hat{\sigma}^{*2})^{-1}(\hat{\mathbf{f}}^* - \hat{\mathbf{f}})$$

- We repeat this procedure many times and form an approximate distribution $\hat{G}$ with the values of ν^*. We may use the $(1-\alpha)$th quantile of $\hat{G}$ as an estimate of $G_{1-\alpha}$.

Let's consider the model $y_i = f(x_i) + \varepsilon_i, i = 1, \ldots, n$ with ε_i IID normal. In Figure 8.8 we see qqplots of the true G, the bootstrap G and the F-distribution approximation.

8.4.6 Displaying the confidence sets

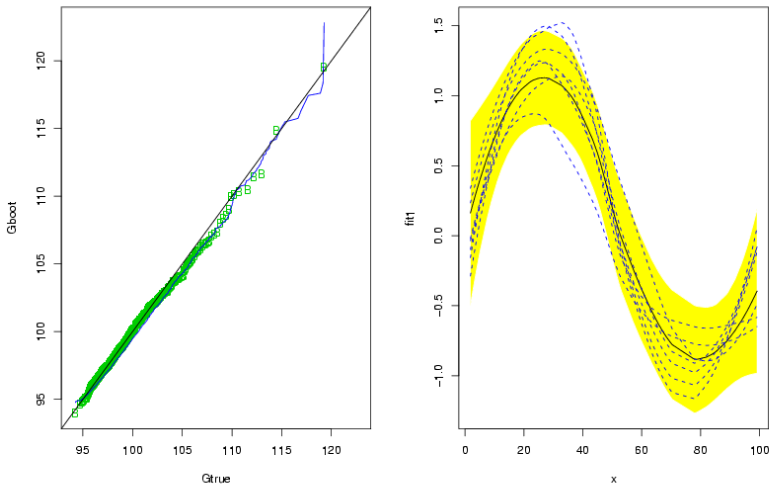
Displaying an $n - dimensional$ ball is not easy.

Global confidence bands usually show the projections of the confidence set onto each of the component sub-spaces. Notice that a function (now I'm using function and n-dimensional vector interchangeably) in this set would actually be in a confidence cube as opposed to a ball! So a vector within the confidence bands isn't

necessarily in the confidence ball. However its true that being in the ball implies being within the band.

Another popular approach is selecting a few functions at random from $N(\hat{\mathbf{f}}, \mathbf{SS}'\hat{\sigma}^2)$ and checking to see if they are in the confidence set. If they are, we plot them. This enables us to see what kind of "shape" functions in the confidence set have. Maybe they all have a bump, maybe a large amount of them are close to being constant lines, etc...

Figure 8.8: QQ-plot of bootstrap vs. true G and the F-distribution approximation. We also see point-wise confidence intervals and curves in (blue) and out (green) of the bootstrap confidence set.



8.4.7 Approximate F-test

Using the F-distribution approximations we may construct F-tests for testing various hypotheses.

The p-value given by the S-Plus function `gam()` is usually testing for linearity and using an F-distribution approximation.

Suppose we wish to compare 2 smoothers $\hat{\mathbf{f}}_1 = \mathbf{S}_1\mathbf{y}$ and $\hat{\mathbf{f}}_2 = \mathbf{S}_2\mathbf{y}$. For example, $\hat{\mathbf{f}}_1$ may be linear regression and $\hat{\mathbf{f}}_2$ may be a "rougher" smoother.

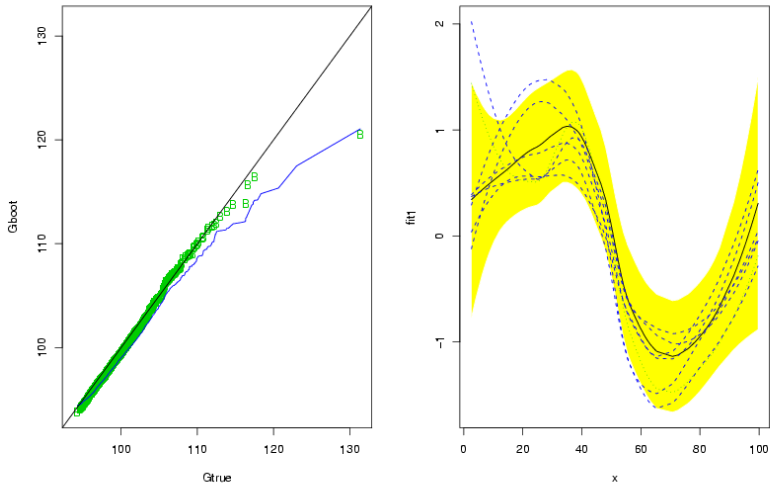
Let RSS_1 and RSS_2 be the residual sum of squares obtained for each smoother. Which one do you expect to be bigger?

and γ_1 and γ_2 be the degrees of freedom of each smoother, $\text{tr}(2\mathbf{S}_j - \mathbf{S}_j\mathbf{S}_j'), j = 1, 2$. An approximation that may be useful for this comparison is

$$\frac{(RSS_1 - RSS_2)/(\gamma_2 - \gamma_1)}{RSS2/(n - \gamma_2)} \sim F_{\gamma_2-\gamma_1, n-\gamma_2}$$

There are moment corrections that can make this a better approximation (see H&T).

Figure 8.9: Same as previos figure but with t-distributed errors



Bibliography

[1] Akaike, H. (1973), "Information theory and an extension of the maximum likelihood principle," in Petrov, B. and Csaki, B., editors, *Second International Symposium on Information Theory*, pp. 267–281, Budapest: Academiai Kiado.

[2] Bozdogan, H. (1987), "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions," *Psychometrica*, 52, 345–370.

[3] Bozdogan, H. (1994), "Mixture-model cluster analysis using a new informational complexity and model selection criteria," in Bozdogan, H., editor, *Multivariate Statistical Modeling*, volume 2, pp. 69–113, The Netherlands: Dordrecht.

[4] Kullback, S. (1959), *Information Theory and Statistics*, New York: John Wiley & Sons.

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

[6] Shibata, R. (1989), "Statistical aspects of model selection," in Williems, J. C., editor, *From Data to Model*, pp. 215–240, New York: Springer-Verlag.

[7] Efron B. and Tibshirani, R.J (1993), *An Introduction of the Bootstrap*. Chapman and Hall/CRC: New York.

[8] Efron, B. (1979). Bootstrap Methods: Another Look At the Jackknife, *The Annals of Statistics* 7, 1–26.