# Contents

# Statistics 475.732:

# Quantitative Fisheries Science



## 1 Introduction

Required reading. Chapter 1 of Hilborn and Walters.

**Exercise 1.1** There is more to stock assessment than just fitting a curve. For example, on p. 10 HW say "*These are the wrong questions*". Explain why.

### 1.1 Basic terminology and notation

Biomass:

Cohort:

Recruitment:

Spawners:

Escapement:

Fecundity:

Carrying capacity:

Effort:

CPUE:

Selectivity/vulnerability:

Mortalities:

Otolith:

## 1.2 The “hard core”of QFS

Active areas of modeling include:

Stock-recruitment relationship: An extremely fundamental issue. Discussed at length in HW Chapter 7.

Stock assessment: That is, estimating the current status of the stock, possibly also population parameters, and “predicting” the effect of exploitation.

Estimation of biological parameters: Such as growth rate, maximum size, age of maturity etc.

Selectivity: That is, how the fishing gear selects fish of different size. In conjunction with knowledge of growth, yield-per-recruit analysis can be undertaken. Also, permits modeling of unaccounted fishing mortality.

Environmental effects: For example, has does water temperature affect recruitment, growth, migration, carrying capacity...

Multispecies models: For example, multivariate modeling of assemblages, modeling of tropic levels...

## 1.3 The quantitative tools

The fundamental tools include the following:

- experimental design
- modeling:
  - regression (often nonlinear)
  - generalized linear models
  - measurement error models
  - dynamic (time series) models

  - mixed models

  - spatial models.

- witchcraft, e.g., VPA?

The list of "hot" tools currently includes:

- mixed effects models
- errors-in-variables models
- state-space models
- hierarchical models
- meta-analysis
- Bayesian analysis
- ADMB

# 2 Stock-recruitment modeling

This is an *extremely* fundamental question (see the opening paragraph, Chapter 7 of Hilborn and Walters). At first glance the problem appears to be just the modeling of $x$ (stock) against $y$ (recruitment). However, this ignores the errors-in-variables problem (p. 287 HW) and the temporal structure (p. 290 HW), the modeling of which has captured a considerable amount of research effort.

## 2.1 Background

Here, recruitment is assumed to occur prior to fishing mortality (?!?!) and hence the stock-recruitment relationship is intended to summarize the *natural* processes and variabilities occuring between the egg stage and recruitment to the fishery.

It is commonly assumed that stock-recruitment curve will have a decreasing slope due to "compensation" - processes which lead to higher per-capita recruitment at low stock size. Compensation can arise from habitat limitations, which could include the extent of suitable habitat for spawning and subsequent survival of larvae and juveniles, and limits on food availability. The other possibility (aka., depensation, or the Allee effect) is that of an increasing slope over some part of the stock-recruitment curve. This could arise if predators remove a given number of pre-recruits.

Myers and Cadigan (1993) found strong evidence of compensation in several commercial species of marine demersal (ground) fish of Atlantic Canada, and Myers et al. (1995) found little evidence of depensation in a study which analysed data from 128 fish stocks.

If the species is semelparous (spawning only once) then recruitment to the spawning stock must exceed spawners over some range of the curve or else the stock will not be viable even in the absense of fishing.

**Exercise 2.1** Read Myers et al. (1995).

## 2.2 Beverton-Holt and Ricker curves for stock-recruitment

These are the two most commonly used models for stock-recruitment and they both assume compensation. These curves can also be given a biological interpretation, arising under different assumptions on the form of the compensation.

### 2.2.1 Beverton-Holt curve

The **Beverton-Holt** curve has an upper asymptote and can be expressed in the form

$$r = \frac{as}{b+s} \ , \tag{1}$$

where $s$ is the spawning stock size and $r$ is the resulting recruits. Note that $a$ is the upper asymptote and $b$ is the spawning stock corresponding to recruitment of $a/2$. The recruits-per-spawner at low stock size (i.e., the slope of the curve at the origin) is $a/b$.

**Biological derivation of Beverton-Holt curve**

A biological "rationale" for the Beverton-Holt curve is that it is obtained from the differential equation

$$\frac{dn_t}{dt} = -(q + pn_t)n_t \ , \qquad p > 0, q > 0 \tag{2}$$

where $n_t$ is the number of fish alive in the cohort at time $t$. Note that $p = 0$ corresponds to no compensation (survival rate independent of $n_t$) because then the rate of change in $n_t$ is proportional to $n_t$. Also, in equation (2) the amount of compensation decreases over time as $n_t$ decreases.

To keep us honest, we will work through the solution of equation (2). Begin by writing it as

$$dt = -\frac{1}{(q + pn_t)n_t} dn_t \tag{3}$$

and by noting that

$$\frac{1}{(q + pn_t)n_t} = \frac{1}{q}\left[\frac{1}{n_t} - \frac{p}{q + pn_t}\right] \ .$$

Integrating both sides of equation (3) then gives

$$\begin{aligned} t &= -\frac{1}{q}\left[\log(n_t) - \log(q+pn_t)\right] + c \\ &= -\frac{1}{q}\log\left(\frac{n_t}{q+pn_t}\right) + c\ . \end{aligned} \tag{4}$$

Now, if it is assumed that the number of eggs is proportional to the number of spawners then $n_0 = ks$, and so at time $t = 0$ equation (4) gives

$$0 = -\frac{1}{q}\log\left(\frac{ks}{q+pks}\right) + c$$

from which we obtain

$$c = \frac{1}{q}\log\left(\frac{ks}{q+pks}\right)$$

and so

$$\begin{aligned} t &= -\frac{1}{q}\log\left(\frac{n_t}{q+pn_t}\right) + \frac{1}{q}\log\left(\frac{ks}{q+pks}\right) \\ &= \frac{1}{q}\log\left(\frac{ks(q+pn_t)}{n_t(q+pks)}\right) \end{aligned}$$

It $t_r$ is the age of recruitment then we have $r = n_{t_r}$ and taking exponents on both sides of the above equation gives

$$e^{qt_r} = \frac{ks(q+pn_t)}{n_t(q+pks)}$$

which, by denoting $\gamma = e^{qt_r}$, can be rearranged to

$$\gamma n_t(q+pks) = ks(q+pn_t)$$

i.e.,

$$n_t(\gamma(q+pks) - pks) = qks$$

which gives (finally)

$$n_t = \frac{qks}{\gamma q + (\gamma-1)pks} = \frac{\frac{q}{(\gamma-1)p}s}{\frac{\gamma q}{(\gamma-1)pk} + s}$$

which is of the form in equation (1) with parameters $a$ and $b$ depending on $q$, $p$, fecundity, and time of recruitment. $\square$

### 2.2.2 Ricker curve

The **Ricker** curve is unimodal and falls back down to an asymptote of zero at high spawning stock size. It can be expressed in the form

$$r = ase^{-bs} \tag{5}$$

where $a$ is the recruits-per-spawner at low stock sizes and $b$ controls the amount of compensation.

The Ricker curve can also be obtained on biological grounds, in this case, via the differential equation

$$\frac{dn_t}{dt} = -(q + ps)n_t \ . \tag{6}$$

Note that, unlike the Beverton-Holt curve, the amount of compensation (given by the $ps$ term) depends on spawners and *does not* decrease over time. Hence the reduced recruitment at high spawning stock sizes.

Solving equation (6) proceeds as shown for the Beverton-Holt case, but here it is considerable simpler (Exercise).

### 2.2.3 Lognormal random variability

The lognormal distribution is the standard assumption for the random variability associated with the *observed* recruitment, $R$, and it is assumed to act multiplicatively. That is, the observed recruitment is

$$R \sim r\varepsilon = f(s)\varepsilon \tag{7}$$

where $\varepsilon$ is lognormal distributed and $f(s)$ is the functional form of the stock-recruitment curve.

The use of the multiplicative lognormal error structure can be justified for a variety of reasons, including

1. The response variable is restricted to positive values.

2. Empirical evidence suggests that variability in recruitment increases with its expected value.

3. Mechanistic justification, by considering the total survival from egg to recruit as a product of many smaller survival events (p. 264 HW[1]).

Recall that if $W$ is a normal random variable, then $e^W$ is a lognormal random variable. Hence, it is typical to see equation (7) expressed in the form

$$R \sim f(s)e^W \ ,$$

where it is usually assumed that $W$ has mean 0 and variance $\sigma_w^2$.

One complication of this error distribution is that the expected value of $e^W$ is *greater* than one. In general, it can be shown that if $W$ has mean $\mu_w$ and variance $\sigma_w^2$ then $e^W$ has mean $\nu = e^{\mu_w+\sigma_w^2/2}$ and variance $\nu^2(e^{\sigma_w^2}-1)$. In particular, if $W$ has mean 0, then the lognormal variable $e^W$ has expected value $E[e^W] = e^{\sigma_w^2/2}$. Consequently,

$$E[R] = f(s)e^{\sigma_w^2/2} \ . \tag{8}$$

It is, however, the case that

$$\text{Median}[R] = f(s) \ .$$

## 2.3 Fitting stock-recruitment curves

With a multiplicative log-normal error structure, the natural approach is to take logs to obtain the convenience of additive normal errors. For the Beverton-Holt curve this gives

$$\log(R_{y+t_r}) = \log\left(\frac{as_y}{b+s_y}\right) + W_y \ ,$$

[1]HW's mention of the Central Limit Theorem in this context is too glib to be meaningful. It actually requires application of the Liapunov or Lindeberg-Feller versions of the CLT (see Billingsley, P. *Probability and measure*). One requirment for application of these CLT's is that no single survival event dominates the others.

where $S_y$ denotes the spawners in year $y$ and $R_{y+t_r}$ is the observed number of fish recruiting at time $t_r$ later.

The Beverton-Holt curve is not linear in the parameters $a$ and $b$ and hence requires the use of nonlinear least squares to minimize

$$\sum_y \left(\log(R_{y+t_r}) - \log\left(\frac{as_y}{b+s_y}\right)\right)^2 .$$

For the Ricker curve,

$$\begin{aligned}
\log(R_{y+t_r}) &= \log(as_y e^{-bs_y} e_y^W) \\
&= \log(a) + \log(s_y) - bs_y + W_y \\
&= \alpha + \log(s_y) - bs_y + W_y \ ,
\end{aligned}$$

where $\alpha = \log(a)$. This is linear in $\alpha$ and $b$ and can be fitted using linear regression. Note that the $\log(s)$ term is simply a constant because, for now, we are regarding the spawning stock as a known explanatory variable. In the statistical jargon it is called an offset and in the case of linear regression it is most easily handled by subtracting it from the response variable. That is, we use linear regression to fit

$$\log\left(\frac{R_{y+t_r}}{s_y}\right) = \alpha - bs_y + W_y \ .$$

## 2.4 Inference

The Ricker curve is fitted by linear regression, hence (assuming the model is correct) the statistical properties of the estimators $\hat{\alpha}$, $\hat{b}$ and $\hat{\sigma}_w$ are known exactly. In the case of nonlinear regression, the nice statistical properties only hold approximately, that is, the properties are obtained asymptotically for large sample sizes. See Appendix 7.3.

### 2.4.1 Functions of parameters

A frequent aspect of fisheries modeling is that it is often some function of the parameters that is of primary interest. For example, the number of recruits-per-spawner at low stock size is a quantity that has recently received much attention because it essentially quantifies the ability of the stock to recover from overexploitation. Recall that this is the slope of the stock recruit curve at the origin and is given by $a/b$ if using the Beverton-Holt curve and by $a = e^{\alpha}$ if using the Ricker curve.

**Delta method**

If the quantity of interest is some function of the parameters, then that function of the parameter estimates is the natural choice for the estimate of the quantity of interest. For example, if $\hat{a}$ and $\hat{b}$ are the estimates of the Beverton-Holt parameters then the natural estimate of the slope at the origin will be $\hat{a}/\hat{b}$. If $\hat{\alpha}$ is the estimate of $\alpha$ in the Ricker curve then the slope would be estimated by $e^{\hat{\alpha}}$. However, what are the properties of these estimates, for example, how do you calculate its variance?

The delta-method is a method for obtaining the approximate distribution of functions of parameter estimates. The delta method is a result of the delta theorem which states that if the estimates are asymptotically normal (i.e., approximately normal for large sample sizes) then any differentiable function of the estimates is also asymptotically normal.

The delta theorem and method are obtained from Taylor series expansion of the function of parameters, as demonstrated in the following example for the function $e^{\alpha}$. The general version of the delta method is presented in Appendix 7.3

**Example 2.1** Suppose that a Ricker curve has been fitted to stock-recruit data using linear regression, which provides an estimate of $\alpha$ and its standard error, which we'll denote by $\hat{\alpha}$ and $s_{\alpha}^2$, respectively. What is the (approximate) standard error of $\hat{a} = e^{\hat{\alpha}}$? Consider the first-order Taylor's series expansion of $e^{\hat{\alpha}}$ around $e^{\alpha}$,

$$\hat{a} = e^{\hat{\alpha}} \approx e^{\alpha} + e^{\alpha}(\hat{\alpha} - \alpha) \ .$$

Taking the variance of both sides gives

$$\mathrm{Var}[\hat{a}] \approx (e^{\alpha})^2 \mathrm{Var}(\hat{\alpha}) \ ,$$

and hence the variance of $\hat{a}$ is estimated by $s_a^2 = e^{2\hat{\alpha}} s_{\alpha}^2$. □

**Re-parameterization**

Another way to obtain inferences about a function of the parameters is to re-parameterize the model so that the quantity of interest is expressed directly as a parameter. For example, letting $\delta = a/b$, the Beverton-Holt model can be rewritten as a nonlinear model involving parameters $\delta$ and $b$

$$\log(R) = \log\left(\frac{\delta s}{1+\frac{s}{b}}\right) + W \ .$$

Similarly, the Ricker model can be fitted as a nonlinear model involving parameters $a$ and $b$ in the form

$$\log\left(\frac{R}{s}\right) = \log(a) - bs + W \ .$$

Ultimately, one wishes to make inference about the unknown parameters and this could be done using the estimated standard errors (using the delta method where necessary) or via likelihood ratio methodology. In general, the likelihood ratio methodology is the more reliable of the two, but it also can be more work to implement because it may require re-parameterization and optimization over subspaces of the parameter space. However, the likelihood profile option in the ADMB software has the potential to alleviate this task. See Appendix 7.3.

## 2.5 Example

## 2.6 Deficiencies in the model

If one stops to think about how the stock and recruitment data are gathered then it quickly becomes obvious that at least two fundamental assumptions underlying the use of least squares are violated (p. 287 of HW).

### 2.6.1 Measurement error in spawning stock

The explanatory variable, spawning stock, is not something that can be easily measured (as we will see in the next chapters). That is, there is error in the explanatory variable. This is known as an *errors-in-variables* model or alternatively a *measurement error* model.

The effect of measurement error in the explanatory variables of a linear regression model are well known (e.g., Seber, 1977). Essentially, the measurement error tends to "spread" the range of the explanatory variables, thereby resulting in the magnitude of coefficients being underestimated.

Here, we briefly look at explicit modeling of errors in variables using maximum likelihood estimation (p. 586 of Casella and Berger). The idea is to consider both the spawners *and* recruits as data. In particular, suppose the observed spawners, $S$, are assumed to be

$$S = se^V$$

where $V$ are normal with mean 0 and variance $\sigma_v^2$. Then, if the recruitment measurement errors $W$ and spawners measurement errors $V$ are all assumed to be independent, the log-likelihood function for the data $(S_y, R_{y+t_r}), y = 1, ..., n$ is

$$-n\log(\sigma_v) - \sum_y \frac{(\log(S_y) - \log(s_y))^2}{2\sigma_v^2} - n\log(\sigma_w) - \sum_y \frac{(\log(R_{y+t_r}) - \log(f(s_y)))^2}{2\sigma_w^2}\ . \tag{9}$$

Note that in addition to $\sigma_v, \sigma_w$ and the parameters of $f(s)$, each of the $s_y$ values is also an unknown to be estimated.

It turns out that the log-likelihood (9) is not bounded! (Why?) The usual way to avoid this is to assume that the ratio $\lambda = \sigma_v^2/\sigma_w^2$ is known. It can then be shown that the estimates of the parameters of $f(s)$ are well behaved (i.e., consistent). However, the estimates of the variances are not! The reason for this odd behaviour is because the usual assumptions required for maximum likelihood to estimate all parameters consistently are not valid due to the fact that the number of parameters to be estimated increases with sample size - this is because there are $n$ values of $s_t$ to be estimated.

But wait, the next subsection shows that it gets worse...

### 2.6.2 Correlated errors

It should not be forgotten that that $S_y$ and $R_y$ are collected sequentially over time. Therefore, correlated errors could arise due to another explanatory variable, such as an environmental influence. Environmental effects (e.g., water temperature) can be highly auto-correlated and therefore could induce temporal autocorrelation in the measurement errors of both $S$ and $R$.

It can also be the case that both spawning stock and recruitment are estimated from the *same* survey. For example, for salmon, observed recruitment is often calculated as the sum of catch and number of spawners going up river. Hence, the measurement errors in $S_y$ and $R_y$ may not be independent.

### 2.6.3 Time-series bias

There is another, more subtle aspect of the time-series nature of the data that can cause problems. This arises because the response variable has the potential to influence subsequent values of the explanatory variable.

**Example:**

The time series bias is a consequence of small sample size and it disappears as sample size is increased. However, time series of stock-recruitment data are typically not very long and hence the time-series bias may be quite prevalent.

## 2.7 Yield-per-recruit analysis

So far, we have been looking at stock-recruitment relationships with the idea of ensuring that recruitment is not drastically reduced by fishing the stock too hard. In this section, we ask the basic question - given a certain number of recruits in a cohort, how should it be fished in order to maximize the yield. This is typically not a statistical question and hence is not covered in any detail here.

Yield-per-recruit analysis essentially just calculates/simulates the passage of the cohort. It requires certain biological information, specifically an average growth curve for the fish (to give weight at age), and an assumed natural mortality. Then, the yield to the fishery can be determined under any given harvest regime. Typically this simply consists of specifying a target fishing mortality and a minimum size limit.

In yield-per-recruit analysis it is common to make the simplifying assumption of knife edge recruitment to fishery. That is, the fishing gear is such that all fish below a certain size are "immune" to fishing, and all fish above the minimum size are equally vulnerable to fishing. The result of these unrealistic assumptions is that the optimum fishing strategy is to fish infinitely hard, using a gear with knife-edge selection occurring at a certain size!

More sophisticated yield-per-recruit analyses use realistic selection curves which reflect the fact that the fishing gear will catch small fish which may be discarded and killed. These analyses give much more sensible results.

The yield-per-recruit analysis is probably best regarded as a device for determining a reasonable minimum size limit and fishing gears to be deployed. It does not take into consideration how the dynamics of the stock change as it is fished down (i.e., ignores the stock-recruitment question).

## 2.8 Closing remarks

Estimation of stock-recruitment relationships is *extremely* problematic. Unfortunately, the difficulty of this task has been sufficient reason for many fisheries to be managed under the assumption of no relationship between stock and recruitment, and this is believed to have lead to the collapse of a number of fisheries worldwide.

More recently, meta-analysis is being employed, whereby stock-recruitment data on all sockeye salmon run (say) are simultaneously analysed. It has not helped that the documentation of stock-recruitment databases tends to be very sketchy, and hence one can seldom be sure of modeling the correct error structure.

## 2.9 References

Beverton, R. J. H., and S. J. Holt. 1957. On the dynamics of exploited fish populations. Fishery Investigation, Series 2 (19): 533 p. HMSO London.

Casella, G., and R. L. Berger. 1990. Statistical inference. Wadsworth, Pacific Grove, California. 650 pp.

Ludwig, D., and C. J. Walters. 1981. Measurement errors and uncertainty in parameter estimates for stock and recruitment. Can. J. Fish. Aquat. Sci. 38: 711-720.

Mertz, G., and R. A. Myers. 1995. Estimating the predictability of recruitment. Fish. Bull. 93: 657-665.

Myers, R. A. 1997. Comment and reanalysis: paradigms for recruitment studies. Can. J. Fish. Aquat. Sci. 54: 978-981.

Myers, R. A., N.J. Barrowman, J. A. Hutchings, and A. A. Rosenberg. 1995. Population dynamics of exploited fish stocks at low population levels. Science. 269: 1106-1108.

Myers, R. A., W. Blanchard, and K. R. Thompson. 1990. Summary of North Atlantic fish recruitment 1942-1987. Canadian Technical Report of Fisheries and Aquatic Sciences No. 1743.

Myers, R. A., and N. G. Cadigan. 1993. Density-dependent juvenile mortality in marine demersal fish. Can. J. Fish. Aquat. Sci. 5x: 1576-1590.

Myers, R. A., G. Mertz, and P. S. Fowlow. 1997. Maximum population growth rates and recovery times for Atlantic cod, *Gadus morhua*. Fish. Bull. 95: 762-772.

Seber, G. A. F. 1977. Linear regression analysis. Wiley.

Waters, J. R., and G. R. Huntsman. 1986. Incorporating mortality from catch and release into yield-per-recruit analyses of minimum size limits. Nth Amer. J. Fish. Mgmt. 6: 463-471.

# 3 Surplus production models

In this chapter we look at surplus production models for biomass (HW, Chap 8). (Hilborn and Walters give their Chapter 8 the title "Biomass dynamic models" because they view surplus production as a concept that is relevant to a wide class of models.) However, the terminology "surplus production model" is very well established in the fisheries arena and we shall stick with it here.

Surplus production models are conceptually very simple because they model only the total (recruited) biomass without explicit regard to other features such as the age-distribution, recruitment, natural mortality etc. When information on these other features is available (e.g., data on the age structure of the population) then fancier models can be used (see next Chapters!). HW emphasize that the fancier models may not necessarily always be better - the very simplicity of the surplus production models can be a virtue. Moreover, in some fisheries it is very difficult, too expensive, or practicably impossible to get additional information about the stock. Thus, the surplus production model will continue to be an indispensible tool of stock assessment.

## 3.1 Surplus production

If $b_y$ is the biomass of the stock at time (year) $y$ then

$$b_{y+1} = b_y + r_y + growth_y - c_y - d_y \tag{1}$$

where $r_y$ is recruitment, $c_y$ is catch, and $d_y$ is the biomass of fish lost due to natural mortality. The term $growth_y$ is due to the physical growth of fish.

Adding together the recruitment and growth terms gives the total "production" of the stock. Surplus production is the amount by which this exceeds the loss due to natural mortality. Denoting surplus production in year $y$ by $p_y$, we have

$$p_y = r_y + growth_y - d_y \ ,$$

and the biomass dynamics equation can then be written

$$b_{y+1} = b_y + p_y - c_y \ . \tag{2}$$

Note that surplus production is the amount by which the biomass will increase when there is no fishing. It is also the amount which can be caught without changing the biomass.

It could be said that some important components are missing from equation (2). For example, unaccounted fishing mortality is substantial in some fisheries and is not included in the documented catch.

### 3.1.1 Surplus production curves

Surplus production curves give the expected surplus production as a function of current biomass, that is $p_y \equiv p(b_y)$. In general, the curve must go through the origin and also drop down to zero again for large biomass (Why?).

**Schaefer curve**

This is the simplest and most widely used surplus production curve. It is simply a quadratic and is usually expressed in the form

$$p(b_y) = rb_y\left(1 - \frac{b_y}{K}\right) \ . \tag{3}$$

This quadratic takes the value zero when biomass is 0 or $K$. Parameter $K$ is commonly called "carrying capacity" because it is the biomass at which the stock is expected to maintain equilibrium when there is no fishing. The quadratic is symmetric, hence it is maximal at $b_y = K/2$, whence maximum surplus production is $rK/4$.

The Schaefer curve is appealing for its simplicity. It also can be justified on biological grounds because it is the discrete time version of the differential equation

$$\frac{db_y}{dy} = rb_y\left(1 - \frac{b_y}{K}\right)$$

which has the (ecologically popular) logistic model as its solution.

**Pella-Tomlinson curve**

This is a generalization of the simple quadratic and can be expressed in the form

$$p(b_y) = rb_y\left(1 - \left(\frac{b_y}{K}\right)^{m-1}\right) . \tag{4}$$

This curve is skewed to the left for $m < 2$ and to the right for $m > 2$.

## 3.2 The data

Before jumping into the fitting of surplus production curves, it is *very* important to get a feel for the origins of the data.

Catch data $C_y$ are essential. These data are relatively easy to obtain in a regulated and enforced fishery. The catch data could be obtained from the log-books of the fishing vessels or from the records of onshore processing plants. In addition to the catch data, it is necessary to have either data on the fishing effort *or* a relative biomass indice.

Ideally the relative biomass indice would be calculated as the catch rate obtained by fishing with standardized gear and operational procedures. This typically requires the use of dedicated research effort. The research fishing would typically occur over a short period of time and take a negligible fraction of the population. Then it may be reasonable to assume that the catch rate is directly proportional to the biomass.

$$E[I_y] = qb_y$$

or perhaps

$$E[\log(I_y)] = \log(q) + \log(b_y) \tag{5}$$

if one uses lognormal errors.

However, research surveys can be extremely expensive. (the operating cost of a 70 m. ocean-going research vessel can easily exceed \$1000 per hour) and it is very often the case that the only other data is a measure of fishing effort. Two general strategies for using catch and effort data are possible.

1. Calculate the observed catch-per-unit-effort, CPUE. This is just

$$\text{CPUE}_y = \frac{C_y}{E_y}$$

where $E_y$ is the fishing effort. The CPUE can then be used as a relative biomass indice. Research surveys indices can be considered a special case of CPUE data corresponding to a short-term fishery in which a standardized fishing effort is applied.

2. Model the catch as a function of *both* biomass and effort. The error term can arise from variability in the fishery and from measurement error involved in observing the catch. The distinction is subtle, but is relevant because one should note that the measured catch, $C_y$, will typically be used in place of the true catch $c_y$ in equation (1).

The drawback of the first approach lies in the validity of using CPUE as a relative biomass indice. The second approach avoids this difficulty, but could suffer from the fact that effort is a covariate which is very difficult to reliably measure, and hence the "errors-in-variables" problem may arise. (Indeed, this can be used as the motivation for the total least squares approach to sequential population analysis; e.g. HW p. 387) These issues are covered in more detail in the next section.

### 3.2.1 Inherent problems with CPUE data

Despite their widespread use, catch-per-unit-effort data are notoriously unreliable. In part, this is due to the difficulty of getting good measurements of both catch *and* effort.

It can be a considerable challenge to determine the effective amount of effort applied to the fishery. This is particularly problematic if the fishing fleet contains vessels of vastly different size and power, and the fishing gear is not common to all vessels. Even in a homogeneous fishery, some vessels outperform others. Also, when does fishing start? For example, should effort include time spent searching for fish?

It is not unusual for considerable modeling effort to be invested in obtaining some standardized measure of fishing effort. This might be via multiple linear regression (or GLM) using covariates such as size and power of the vessel, experience of crew, time of day, addition of fish-finding equipment.

The assumption that CPUE is proportional to biomass requires that the density of fish increases in proportion to biomass. However, behavioural features of the fish can render this assumption invalid. For example, if the fish aggregate in large schools upon which the fishery is directed, then the total stock abundance could decline substantially before CPUE shows an obvious decline. On the other hand, if the fish are spatially dis-aggregated then an increase in stock size would result in an increase in the range of the population without a corresponding increase in density.

Another weakness is that the presumption that CPUE is proportional to biomass is only strictly valid for small catches obtained from moderate fishing effort. To see this, note that CPUE $\propto b$ is equivalent to $C \propto bE$ which implies that catch can exceed biomass for sufficiently large effort! Defining $b$ to be the biomass midway through the season might be more reasonable.

Fishing effort is most appropriately regarded as a measure of the relative *rate* at which fish are being caught. For example, if daily fishing effort, $E$, is constant and occurs over a short fishing season then one could write

$$\frac{db_y}{dy} = -qE$$

which assumes that the fishing season is sufficiently short that natural mortality can be ignored. This differential equation has solution

$$b_{y+\Delta t} = b_y e^{-qE\Delta t}$$

where $\Delta t$ is the number of days in the fishing season. Thus, the catch is

$$c_y = b_y - b_{y+\Delta t} = b_y \left(1 - e^{-qE\Delta t}\right) \ . \tag{6}$$

**Exercise 3.1** . Deduce the appropriate equation for catch when natural mortality occurs at rate $m$. That is, when

$$\frac{db_y}{dy} = -qE - m \ .$$

## 3.3 Fitting surplus production curves

Here, we shall assume that the data to be modeled are a sequence of relative biomass indices. In principle, the methods herein could be modified to suit the alternative approach of modeling catch.

Hilborn and Walters cover three estimation methods. Here, we look at these three and also a further three which have recently been developed.

The first method, based on an equilibrium assumption, is not a true biomass dynamics model because it does not make use of any of the temporal information in the data. The other five models do utilize the temporal structure. They differ in how they apply error terms to the two equations defining the dynamic surplus production model. These two equations are the "process" equation (the biomass dynamics equation (2)) which gives the next biomass given the current biomass and catch, and the "observation" equation (e.g., equation (5)).

### 3.3.1 Equilibrium methods

These methods assume that the stock is always at equilibrium no matter what the catch is. That is, it is assumed that the catch, $c$, corresponds to surplus production. This enables one to write

$$c = qEb = rb\left(1 - \frac{b}{K}\right) \ . \tag{7}$$

Solving this for $b$ gives

$$b = K\left(1 - \frac{qE}{r}\right)$$

and substituting back into equation (7) gives

$$c = qEK\left(1 - \frac{qE}{r}\right)$$

or equivalently

$$\frac{c}{E} = qK\left(1 - \frac{qE}{r}\right) \ .$$

These last two equations enable one to estimate $qK$ and $q/r$ using a quadratic regression of catches on effort, or a linear regression of CPUE on effort, respectively. This enables the optimal effort and corresponding maximum surplus production to to estimated. Note that it is not possible to get individual estimates for $q$, $K$ and $r$.

**Exercise 3.2** Show that optimal effort is given by $r/2q$.

**Warning:** The equilibrium assumption used above will be absolutely absurd in almost all fisheries and is just about guaranteed to result in terrible, and probably very dangerous, estimates of optimal effort and maximum surplus production. This method is best regarded as part of the ancient history of stock assessment.

### 3.3.2 Process-error model

This model assumes that there is random error/variability in the process equation. This assumption is, of course, eminently sensible because we hardly believe that biological processes such as growth, recruitment and mortality will be deterministic. They will no doubt be highly dependent on environmental influences. Also, the process error could also include the effect of misreported catches and unaccounted mortalities.

The other assumption is that there is *no* error in the observation equation. That is, that the relative biomass indices are measured exactly without error! This assumption is made so that one can write

$$b_y = \frac{I_y}{q} \ . \tag{8}$$

Assuming normal error and the Pella-Tomlinson form of surplus production,

$$b_{y+1} = b_y + rb_y\left(1 - \left(\frac{b_y}{K}\right)^{m-1}\right) - C_y + \epsilon_y \ , \tag{9}$$

and substituting (8) into (9) gives

$$I_{y+1} = I_y + rI_y\left(1 - \left(\frac{(I_y/q)}{K}\right)^{m-1}\right) - C_y/q + \varepsilon_y \ .$$

One can then apply nonlinear least squares to obtain estimates of $r$, $q$ and $K$ (which corresponds to maximum likelihood under the assumption that $\varepsilon_y$ are iid Normal with mean zero). When the Schaefer surplus production curve is used ($m = 2$) then the estimates can be obtained by multiple linear regression.

### 3.3.3 Observation-error model

This model explicitly includes the observation error associated with the measurement of the relative biomass indice $I_y$. However, it assumes that the biomass dynamics of the stock are *perfectly* described by the process equation. Thus, for any given values of the inital biomass $b_0$ and parameters $r, K$ and $q$, the entire biomass trajectory $b_0, b_1, b_2, ...$ can be specified.

Under the assumption that the errors on the $I_y$ are multiplicative lognormal, one can write

$$\log(I_y) = \log(q) + \log(b_y) + \varepsilon$$

and use nonlinear least squares to obtain the maximum likelihood estimates of $b_0, r, K$ and $q$. In practice, the likelihood can be extremely flat in the neighbourhood of the solution. If the stock was not fished heavily prior to time $t = 0$ then it is common practice to set $b_0 = K$, that is, to assume that the stock was initially (at time $t = 0$) at its carrying capacity.

### 3.3.4 Penalized likelihood

The stochastic process and observation equations used in the previous two sections are

$$b_{y+1} = b_y + p(b_y) - C_y + \epsilon_y \qquad (10)$$

$$\log(I_y) = \log(q) + \log(b_y) + \varepsilon_y \ . \qquad (11)$$

Let all $\epsilon_y$'s and $\varepsilon_y$'s be independent with $\epsilon_y$'s distributed $\mathrm{N}(0, \sigma^2)$ and $\varepsilon_y$'s distributed $\mathrm{N}(0, \tau^2)$.

Since all the error terms are independent and normally distributed then it is very natural to consider minimizing the "total least-squares" obtained from considering the negative of the log-likelihoods from the process and observation equations. That is,

$$\begin{aligned} n\log(\sigma) \quad &+ \quad \sum_y \frac{(b_{y+1} - [b_y + p(b_y) - C_y])^2}{2\sigma^2} \\ &+ \quad n\log(\tau) + \sum_y \frac{(\log(I_y) - \log(q) - \log(b_y))^2}{2\tau^2} \ . \end{aligned} \tag{12}$$

In practice, it is not possible to reliably estimate both $\sigma$ and $\tau$ and hence it is usual to assume that the ratio $\lambda = \sigma^2/\tau^2$ is known. Then, the total least squares simplifies to the minimization of

$$-2n\log(\sigma) - \frac{\sum_y (b_{y+1} - [b_y + p(b_y) - C_y])^2 + \lambda \sum_y (\log(I_y) - \log(q) - \log(b_y))^2}{2\sigma^2} \ .$$

However, the total least-squares approach is not a conventional likelihood approach for the simple reason that the states are *not* observed, yet equation (12) treats them as though they were! This is, in fact, an example of penalized likelihood, whereby the "fictitious" likelihood term (the first line of equation (12)) is viewed as a penalty which penalizes the model for the states departing far from the biomass dynamics equation.

### 3.3.5 Classical state-space model

The state-space model is more rigorous than the penalized likelihood because it is based on obtaining the true likelihood for the observed data. That is, it treats the $b_y$ as unobserved states and the $I_y$ as data. The difficulty with state-space modeling is that obtaining the likelihood for the data can be very difficult because of the fact that $b_y$ are now also random.

It is necessary to consider the joint density function for $b_y$'s and $I_y$'s. This is given by

$$\begin{aligned} f_{b,I} &= f(b_1, b_1, b_2, ..., b_n, I_1, I_2, ..., I_n | C_1, C_2, ..., C_n) \\ &= f(b_1) f(I_1|b_1) f(b_2|b_1, C_1) \times ... \times f(b_n|b_{n-1}) f(I_n|b_n) \ . \end{aligned} \tag{13}$$

The conditional densities in (13) are obtained immediately from (11) and (10). The density function for the data, $I_y$, is then obtained by integrating over the unobserved $b_y$. That is,

$$f_I = f(I_1, I_2, ..., I_n | C_1, C_2, ..., C_n) = \int f_{b,I} \ db_1 db_2, ..., db_n \ . \tag{14}$$

In general, this integral is not tractable, with the exception that when the process and observation equations are both linear functions of the states, and the errors are all normally distributed, then it can be obtained via the Kalman filter. However, in the surplus production model above the process equation is nonlinear, whence an approximate Kalman filter can be used by working with a linear approximation to the process equation. This is known as the extended Kalman filter.

The likelihood is simply obtained by noting that the density function $f_I$ is a function of parameters $q, \sigma^2, \tau^2$ and the parameters of the surplus production curve $p(b_y)$.

### 3.3.6 Bayesian state-space model

The tractability problems facing the likelihood implementation of the state-space model are not relevant to the Bayesian implementation. Bayesian analyses can usually be carried out using general purpose computer intensive Markov Chain Monte Carlo (MCMC) techniques to sample from the posterior distribution of the unknowns. The use of these sampling techniques does require some care and the onus is on the modeler to ensure that the sample has converged and has been run for sufficiently long.

**Prior specification**

The Bayesian approach requires a prior distribution for the unknowns. For example, if the surplus production curve is of the Schaefer form with parameters $r$ and $K$, then the unknowns are $(r, K, q, \sigma^2, \tau^2, b_1, b_2, ..., b_n)$.

Specification of reasonable priors for $(r, K, q, \sigma^2, \tau^2)$ can be extremely challenging. For example, Meyer and Millar (1999) and Millar and Meyer (in press) used a non-informative prior on $q$ which assumed prior density proportional to $1/q$. This choice for $q$ is widely accepted amongst the Bayesian fisheries literature. Obtaining priors for the other parameters was more problematic and involved the use of prior information. In particular, the informative prior on $r$ was obtained by a meta-analysis of similar stocks, and the vaguely informative prior on $\sigma^2$ was based on auxilliary data indicating the variability in population processes of the species.

The prior for $(r, K, q, \sigma^2, \tau^2, b_1, b_2, ..., b_n)$ is induced from the prior on $(r, K, q, \sigma^2, \tau^2)$ by virtue of the fact that

$$\begin{aligned} & f(r, K, q, \sigma^2, \tau^2, b_1, b_2, ..., b_n) \\ & \qquad = f(r, K, q, \sigma^2, \tau^2) f(b_1|K) f(b_2|b_1, r, K, \sigma^2) \times ... \times f(b_n|b_{n-1}, r, K, \sigma^2) \ . \end{aligned}$$

Note the explicit conditioning on catches, $C_y$.

**Sampling from the posterior distribution**

The posterior distribution is

$$\begin{aligned} & f(r, K, q, \sigma^2, \tau^2, b_1, b_2, ..., b_n | I_1, I_2, ..., I_n) \\ & \quad \propto f(r, K, q, \sigma^2, \tau^2, b_1, b_2, ..., b_n, I_1, I_2, ..., I_n) \\ & \quad = f(I_1|b_1, q, \tau^2) f(I_2|b_2, q, \tau^2) \times ... \times f(I_n|b_n, q, \tau^2) f(r, K, q, \sigma^2, \tau^2, b_1, b_2, ..., b_n) \ . \end{aligned}$$

This posterior can be sampled using Gibbs sampling via the BUGS software. This implementation requires just a few lines of code and has the advantage that the modeler does not have to explicitly specify the posterior distribution. The ADMB software is another alternative (Exercise). It uses the Metropolis-Hastings algorithm and has the ability to acheive an efficient sampler by virture of the fact that it is able to calculate an approximation to the posterior distribution to be used as the proposal distribution - see Chapter 6.

## 3.4 Closing remarks

It is undeniably the case that both process and observation error are present. The general consensus is that, if only one source of error is to be modeled, then observation error models are preferable to process error models.

Penalized likelihood has a strong following in some parts of the stock assessment world, particularly the Pacific Northwest region. AD model builder is very well suited to fitting such models. However, penalized likelihood is known to have poor properties in many applications and this can be true even for large sample sizes.

State-space models have recently opened the door to modeling of both sources of error. However, as always, it remains to be seen whether the extra complexity of these models will lead to better inference.

## 3.5 References

Meinhold, R. J., and N. D. Singpurwalla. 1983. Understanding the Kalman filter. AmerStat. 37: 123-127.

Meyer, R. and R. B. Millar. 1999. BUGS in Bayesian stock assessment. Can. J. Fish. Aquat. Sci. 56: 1078-1087.

Millar, R. B. and R. Meyer. Nonlinear state-space modeling of fisheries biomass dynamics using the Gibbs sampler. Appl. Stat. In press.

Polacheck, T., R. Hilborn, and A. E. Punt. 1993. Fitting surplus production models: comparing methods and measuring uncertainty. Can. J. Fish. Aquat. Sci. 50: 2597-2607.

Prager, M. H. 1994. A suite of extensions to a nonequilibrium surplus-production model. Fish. Bull. 92:374-389.

# 4 Hybrid models

Here, the term "hybrid" is used to denote models that are midway between surplus production models and fully age-structured models (see next Chapter). That is, they dis-aggregate surplus production into the processes of growth, recruitment and natural mortality, but do not utilize fully age-structured data.

Even if age-structured catch data are not available, it could still be very advantageous to consider modeling some age-structure. For example, if a recruitment indice was available and it showed recruitment to be highly variable, then an age-structured model would utulize knowledge about recruitment strength. The age-structure also allows the growth process to specified if a growth curve for the species is available.

## 4.1 Using biological information

Let $n_{a,t}$ be the number of age $a$ fish at time $t$, and let $k$ be the age at which fish recruit to the fishery. Suppose that a weight-at-age curve is known for the species and let $w_a$ denote weight at age $a$. The total recruited biomass at time $t$ is then

$$b_t = \sum_{a \geq k} n_{a,t} w_a \ .$$

The most appropriate form of the dynamics equation will depend on the timing of the fishery. Following the HW (p. 332) example, if the fishery takes place over a short season at the start of the time period and recruitment occurs instantaneously at the start of the time period, then

$$\begin{aligned} n_{k,t} &= r_t && (1) \\ n_{a+1,t+1} &= n_{a,t}(1-h_t)e^{-m_t} \ , a \geq k && (2) \end{aligned}$$

where $h_t$ is the proportion of fish harvested and $m_t$ is the instantaneous rate of natural mortality. In many cases, it will simply be the case that

$$h_t = \frac{c_t}{b_t} \ .$$

though if reliable effort data, $E_t$ were available then one might assume $h_t = 1 - e^{-qE_t}$.

In principle, the form of the equation for population dynamics can be tailored to the nature of the fishery. For example, if fishing is spread evenly over the entire year then equation (2) would typically be replaced with

$$n_{a+1,t+1} = n_{a,t}(1 - h_t e^{m_t/2})e^{-m_t} \ . \tag{3}$$

**Exercise 4.1** Equation (3) can be given a heuristic justification by virtue of the fact that it is obtained if proportion $h_t$ is harvested instantaneously at the middle of the year. Verify this.

### 4.1.1 Parameter estimation

If a relative biomass indice is available then one approach would be to use a simple observation error model using equation (11) of the previous chapter. As with the surplus production model, one could alternatively use a different form of the observation equation. For example, one could use equation (3.6) in place of (3.11).

Fitting the observation error model requires a time series of predicted $b_t$ values, which in turn requires initializing the population by specification of numbers at age at time $t_0$. Thus, there are far too many unknown quantities and it is necessary to assume some are known (using auxilliary information).

It is typical to either assume constant unknown recruitment $r = r_t$ or to use auxilliary information on recruitment. The auxilliary information could be a recruitment indice obtained from juvenile research surveys or from some function relating relative recruitment to environmental data. This approach permits the numbers at age $t_0$ to be deduced by using the recruitment information for $A - k$ years prior to $t_0$ (where $A$ is the maximum age of fish in the model) provided information on harvest rates prior ot $t_0$ is also available.. Of course, recruitment could also be obtained from a stock-recruitment curve.

With all these assumptions made, the fit then requires estimation of $q$ (from equation (11)) and $r$ (in the constant recruitment case).

## 4.2 Delay-difference models

The delay-difference model (Deriso 1980) is a special case of the above model in which a particular shape of the weight-at-age curve is specified. This leads to a dynamics equation in which $b_{t+1}$ can be predicted from $b_{t-1}$ and $b_t$, the advantage of which is that it is no longer necessary to know the entire age-structure of the population.

The delay-difference model assumes the weight-at-age curve to be of the form

$$w_a = w_\infty \left(1 - e^{-c(a-a_0)}\right) \ . \tag{4}$$

This is the well known von-Bertalanffy growth curve. Parameter $a_0$ is the age at which the fish has zero weight, $w_\infty$ is the maximum theoretical weight, and $c$ is the growth rate.

**Remark:** The von-Bertalanffy is *not* the obvious choice for weight-at-age. In fact, one can use physiological arguments to obtain differential equations for growth, the solution of which leads to the von-Bertalanffy curve as a model for length-at-age! However, over the size range of recruited fish, the von-Bertalanffy may be reasonable to describe weight-at-age.

**Exercise 4.2** Show that equation (4) implies that

$$w_a = \alpha + \rho w_{a-1} \ , \qquad \alpha > 0, 0 < \rho < 1 \ . \tag{5}$$

The above exercise establishes that weight-at-age can be expressed as a linear first order difference equation. The weights at age of recruited fish can therefore be determined once $\alpha$, $\rho$ and $w_{k-1}$ (weight of a pre-recruit) are specified.

**Exercise 4.3** Show that equations (1), (2) and (5) lead to the second order difference equation

$$b_t = (1+\rho)s_{t-1}b_{t-1} - \rho s_{t-1}s_{t-2}b_{t-2} - \rho s_{t-1}w_{k-1}r_{t-1} + w_k \rho_t \ , \tag{6}$$

where $s_t = (1-h_t)e^{-m_t}$ denotes the overall survival rate.

### 4.2.1 Parameter estimation

In theory, parameters $\alpha, \rho$ and $w_{k-1}$ could be estimated, but in practice this leads to an overparameterized model and it is best they they be fixed. The same will likely be true of $m$.

This leaves parameter $q$ of equation (11), recruitment, and the initial biomass at times $t_0$ and $t_0 - 1$. It would be reasonable to assume that $b_{t_0}$ and $b_{t_0-1}$ are equal if the fishery has not been fished too hard initially.

As with the surplus production case, the observation error model would be the "default" method of matching the predicted biomass indice trajectory with the observed trajectory and it would likely be done using least squares (possibly on the log scale). State-space modeling has recently been applied to this model, from both the frequentist and Bayesian viewpoints.

## 4.3 Closing remarks

When first introduced, there was much initial enthusiasm about delay-difference models because they were regarded as a much desired compromise between the simplistic surplus production models and the complicated age-structured models. However, delay difference models do not appear to have been widely used in stock assessment. This may be because it can be difficult to obtain good biological knowledge about the species, and it should not be forgotten that growth of fish is known to exhibit considerable temporal variability.

Because these methods do not utilize catch-at-age data, it is an implicit assumption that the harvest rate is common across all ages. This assumption is seldom true - and this may well be another reason why these methods are not in widespread use.

Hilborn and Walters note that other data on the fishery can be utilized by these hybrid models. For example, the models permit calculation of the mean weight, $\bar{w}_t$, in each year. Higher order moments (e.g., variance and skewness) of the weight distribution can also be calculated. These can be equated to their observed values and added as an additional sums of squares term in the least squares (say) formula.

## 4.4 References


Deriso, R. B. 1980. Harvesting strategies and parameter estimation for an age-structured model. Can. J. Fish. Aquat. Sci. 37: 268-282.

Kimura, D. K., J. W. Balsiger, and D. H. Ito. 1996. Kalman filtering the delay difference equation: practical approaches and simulations. Fish. Bull. 94: 678-691.

Meyer, R. and R. B. Millar. 1999. Bayesian stock assessment using a state-space implementation of the delay difference model. Can. J. Fish. Aquat. Sci. 56: 37-52.

Schnute, J. T. 1985. A general theory for analysis of catch and effort data. Can. J. Fish. Aquat. Sci. 42: 414-429.

Schnute, J. T. 1987. A general fishery model for a size-structured fish population. Can. J. Fish. Aquat. Sci. 44: 924-940.

# 5 Sequential Population Analysis

The basic objective of sequential population analysis, SPA, is simple - to model the annual progression of cohorts (year-classes) from their time of recruitment through to either their "departure" from the fishery or through to the current year. This requires that the fish can be reliably aged so that catch-at-age data can be obtained. Note that here, catch-at-age is numbers at age (*not* weight).

Fish are usually aged by observing a hard body part such as an otolith or scale. Typically, on the order of $10^3$ fish are aged, and this information is then presented in the form of an age-length key which gives the age-distribution of a fish of any given length. Due to annual variability in growth, it is desirable that age-length keys be obtained annually. The age-length key enables the age-distribution of the entire catch to be inferred from the length-distribution of the catch.

Within the general framework of SPA, the implementation for any particular assessment depends critically on what are considered to be the "data" (i.e., response variable).

## 5.1 The data

If a relative index of numbers at age is available from research fishing then it makes sense that this index will be the response variable because it will almost certainly be more reliable than measurements obtained from the commercial fleet. In this case, catch will appear only in the cohort (process) equation.

In the absense of research fishing, one needs to utilize information from commercial fishing. This typically involves treating commercial catch as the response and treating effort (if available) as an explanatory variable.

Auxiliary data from other sources can also be used (Section 5.5.2) For example, these could be recruitment indices, CPUE data, estimates from tagging studies etc.

The first example below is an assessment where comprehensive research survey data are available. The second example assumes such data are unavailable and makes use of commercial catch and effort data. The third example gives an alternative model formulation to the second example whereby the error structure assumed for observed catch-at-age is probably more realistic and better reflects the nature of the data.

## 5.2 Example 1: Northern cod

The so-called "Northern cod" is the stock management group in NAFO (North Atlantic Fisheries Organization) regions 2J, 3K and 3L. An extensive research cruise of approximately 6 weeks duration was undertaken each autumn (Oct-Nov).

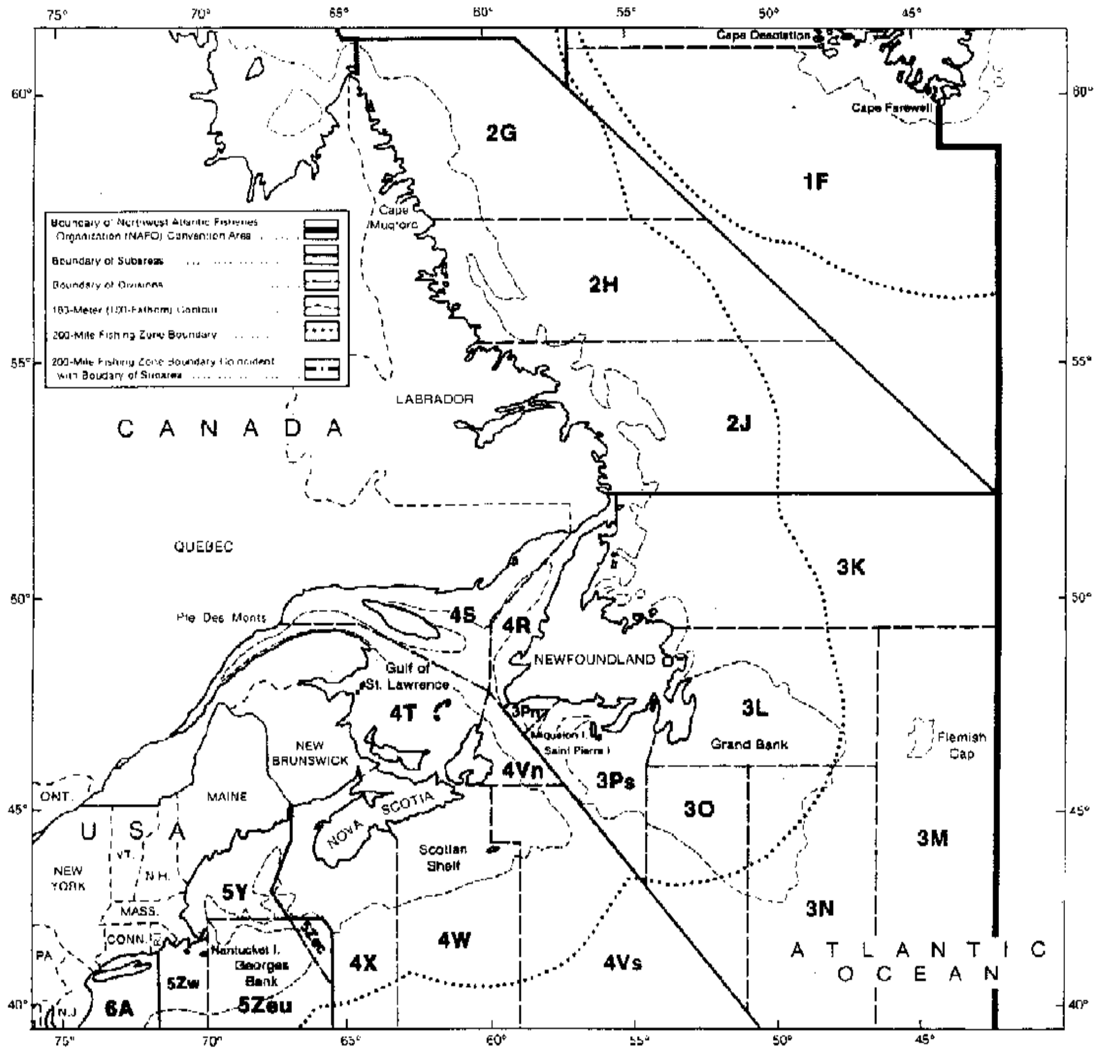



Fig. 3. NAFO Div. 2J, 3K, and 3L showing the Canadian 200-mile fishing zone.

Let $c_{a,y}$ denote the commercial catch of age $a$ fish taken during year $y$. The observed catch-at-age data are given below (from Myers and Cadigan 1995b)

**Table 5.1. Catch at age**

| | Age | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Year | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 1978 | 1323 | 17556 | 39206 | 20319 | 7711 | 3078 | 1530 | 1083 | 437 | 219 | 105 |
| 1979 | 1152 | 12361 | 37439 | 29202 | 10982 | 3460 | 1300 | 757 | 560 | 183 | 116 |
| 1980 | 2554 | 12025 | 28814 | 30016 | 18017 | 4830 | 1217 | 520 | 232 | 229 | 56 |
| 1981 | 2185 | 7172 | 13191 | 24800 | 22014 | 11848 | 3175 | 779 | 309 | 195 | 125 |
| 1982 | 1702 | 31286 | 19003 | 14297 | 25435 | 16930 | 11936 | 1923 | 338 | 156 | 90 |
| 1983 | 2585 | 13616 | 42602 | 19028 | 12044 | 14701 | 8934 | 6341 | 1018 | 248 | 90 |
| 1984 | 782 | 14871 | 31760 | 38624 | 12503 | 7246 | 8910 | 4227 | 2536 | 451 | 146 |
| 1985 | 650 | 14824 | 36614 | 33922 | 28006 | 7050 | 3836 | 5162 | 2905 | 1681 | 254 |
| 1986 | 831 | 15219 | 44168 | 45869 | 26025 | 14722 | 3104 | 2000 | 1977 | 1101 | 574 |
| 1987 | 2329 | 9217 | 32340 | 49061 | 28469 | 19505 | 5818 | 1346 | 676 | 873 | 391 |
| 1988 | 2779 | 14651 | 20184 | 47917 | 45725 | 18608 | 9026 | 4337 | 774 | 422 | 366 |
| 1989 | 1696 | 17639 | 21150 | 25212 | 38708 | 28499 | 8696 | 3640 | 1695 | 572 | 244 |
| 1990 | 7693 | 40557 | 36410 | 22695 | 16390 | 17940 | 9156 | 2865 | 1084 | 478 | 103 |
| 1991 | 3111 | 31654 | 53805 | 29553 | 9064 | 6164 | 4745 | 1696 | 641 | 250 | 88 |
| 1992 | 430 | 3860 | 14535 | 12211 | 4526 | 1372 | 376 | 199 | 104 | 18 | 9 |
| 1993 | 940 | 4993 | 3343 | 1940 | 700 | 147 | 21 | 0 | 0 | 0 | 0 |

The relative index of numbers at age from the autumn research cruises is the average number caught per tow:

**Table 5.2. Relative indices of numbers at age**

| | Age | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Year | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1978 | 5.39 | 11.51 | 13.95 | 5.51 | 1.62 | 0.63 | 0.47 | 0.33 | 0.12 | 0.09 |
| 1979 | 1.94 | 11.78 | 16.79 | 10.53 | 2.27 | 0.92 | 0.31 | 0.26 | 0.19 | 0.06 |
| 1980 | 2.48 | 3.83 | 13.23 | 13.31 | 4.99 | 1.19 | 0.37 | 0.23 | 0.11 | 0.16 |
| 1981 | 5.12 | 2.74 | 3.26 | 9.67 | 8.79 | 3.66 | 0.74 | 0.23 | 0.10 | 0.11 |
| 1982 | 5.87 | 5.92 | 3.83 | 2.79 | 5.82 | 5.31 | 2.59 | 0.57 | 0.16 | 0.09 |
| 1983 | 12.22 | 10.62 | 10.83 | 3.87 | 2.43 | 5.33 | 2.93 | 1.42 | 0.36 | 0.14 |
| 1984 | 10.79 | 15.23 | 11.34 | 9.59 | 2.30 | 1.37 | 2.09 | 1.30 | 0.54 | 0.28 |
| 1985 | 7.27 | 12.35 | 10.01 | 7.28 | 4.24 | 0.92 | 0.78 | 0.67 | 0.41 | 0.15 |
| 1986 | 4.77 | 20.70 | 31.29 | 21.28 | 10.14 | 5.26 | 1.37 | 0.58 | 0.68 | 0.42 |
| 1987 | 2.04 | 4.03 | 13.23 | 11.61 | 4.38 | 2.67 | 1.38 | 0.34 | 0.17 | 0.19 |
| 1988 | 3.93 | 3.20 | 5.29 | 10.57 | 10.13 | 2.58 | 1.55 | 0.79 | 0.15 | 0.11 |
| 1989 | 8.98 | 8.30 | 6.20 | 6.52 | 8.23 | 4.84 | 1.62 | 0.98 | 0.43 | 0.16 |
| 1990 | 10.93 | 12.95 | 8.61 | 5.64 | 3.90 | 3.98 | 1.68 | 0.55 | 0.23 | 0.12 |
| 1991 | 3.35 | 13.97 | 9.00 | 3.31 | 1.10 | 0.50 | 0.35 | 0.16 | 0.04 | 0.02 |
| 1992 | 1.78 | 2.30 | 2.72 | 1.42 | 0.35 | 0.04 | 0.02 | 0.01 | 0.00 | 0.01 |
| 1993 | 0.60 | 0.83 | 0.34 | 0.22 | 0.04 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |

The values in Table 5.2 are treated as the response variable and the commercial catches in Table 5.1 are treated as known constants (measured without error).

Let $n_{a,y}$ denote the number of age $a$ fish at the *beginning* of year $y$. The majority of fishing takes place over the few ice free months around the middle of the year and hence it is reasonable to use the dynamics equation

$$n_{a+1,y+1} = n_{a,y}e^{-m} - c_{a,y}e^{-m/2} \ . \tag{1}$$

The instantaneous natural mortality for this species has traditionally been taken to be constant for all ages and years, $m = 0.2$. (Recently, this assumption has been identified as a likely contributing cause to the infamous demise of Northern cod.)

The data in Table 5.2, denoted $I_{a,y}$ are modeled as lognormal. One slight complication is that the research surveys are conducted around November, rather than at the start of the year. For this reason, Myers and Cadigan (1995b) use the observation equation

$$\log(I_{a,y}) = \log(q_a) + \frac{\log(n_{a,y}) + 11\log(n_{a+1,y+1})}{12} + \epsilon_{a,y} \ , \tag{2}$$

where $\epsilon$'s are iid Normal. The above equation is effectively assuming

$$\log(n^*_{a,y}) = \frac{\log(n_{a,y}) + 11\log(n_{a+1,y+1})}{12} \ .$$

where $n^*_{a,y}$ is the number of age $a$ fish (at the beginning of year $y$) that are still alive at the time of the survey in November of year $y$.

The $q_a$ quantities are age-specific catchabilities of the research trawl. It is very apparent from Table 5.2 that these initially increase with age. It is much more difficult to know whether the catchabilities decline for the older ages. This could occur if the older fish are more able to avoid the research trawl, or if they migrate to more remote areas or areas not amenable to bottom trawling.

### 5.2.1 Parameter estimation

A sequential population analysis of these data would construct the cohort trajectories from estimates of the numbers at age in 1978, and recruitments in years 1979 through 1993. That is, it needs to estimate

$$n_{a,1978} \qquad a = 3, ..., 13 \tag{3}$$

$$n_{3,y} \qquad y = 1979, ..., 1993\ . \tag{4}$$

Note that these quantities correspond to the top row and left hand column of a table of numbers at age.

**Table 5.3. Numbers at age**

| | Age | | | | |
|---|---|---|---|---|---|
| Year | 3 | 4 | ... | 11 | 12 |
| 1978 | $n_{1978,3}$ | $n_{1978,4}$ | ... | $n_{1978,11}$ | $n_{1978,12}$ |
| 1979 | $n_{1979,3}$ | | | | |
| 1980 | $n_{1980,3}$ | | | | |
| . | . | | | | |
| . | . | | | | |
| . | . | | | | |
| 1992 | $n_{1992,3}$ | | | | |
| 1993 | $n_{1993,3}$ | | | | |

The dynamics equation (1) is then used to fill out the rest of this table.

Thus, nonlinear least squares can be used to minimize the objective function arising from the observation equation (2),

$$\sum_{a,y} \left( \log(I_{a,y}) - \log(q_a) - \frac{\log(n_{a,y}) + 11 \log(n_{a+1,y+1})}{12} \right)^2$$

The parameters to be estimated are the catchabilities, $q_a$, the initial population numbers in 1978, and recruits. One could reduce the number of parameters by incorporating structure or auxilliary knowledge about the shape of the age-specific catchabilities.

In contrast, the analysis of Myers and Cadigan proceeds as a VPA (virtual population analysis) by making assumptions about the terminal fishing mortalities to obtain the right-hand column and lower row of Table 5.3. The cohorts are then projected back in time by re-writing equation (1) as

$$n_{a,y} = n_{a+1,y+1} e^m + c_{a,y} e^{m/2}\ .$$

**Exercise 5.1** Fit a SPA model to the above Northern cod data using ADMB.

## 5.3 Example 2: Doubleday's Method

This is essentially the methodology described in section 11.6 of HW. Here, research survey data are not available and hence the actual commercial catch data are treated as response variables and fishing effort data (if available) are treated as an explanatory variable.

As with the previous example, natural mortality will be assumed constant over all years and ages. This is necessary to reduce the number of parameters in the model. However, it must be recognized that this is certainly not a desirable assumption.

The cohort dynamics equation takes the form

$$n_{a+1,y+1} = n_{a,y} e^{-(f_{a,y}+m)} \ , \tag{5}$$

where $f_{a,y}$ is (instantaneous) fishing mortality and $z_{a,y} = f_{a,y} + m$ is total (instantaneous) mortality.

Note that the number of fish removed from the cohort is

$$n_{a,y} - n_{a+1,y+1} = n_{a,y} \left[1 - e^{-(f_{a,y}+m)}\right] \ .$$

The number that are removed as catch is giving by scaling the removals by the proportion of total mortality that is due to fishing. This gives rise to the well known "Baranov catch equation"

$$c_{a,y} = n_{a,y} \frac{f_{a,y}}{f_{a,y}+m} \left[1 - e^{-(f_{a,y}+m)}\right] \ . \tag{6}$$

Assuming lognormal errors, the observation equation is then

$$log(C_{a,y}) = \log(n_{a,y}) + \log\left(\frac{f_{a,y}}{f_{a,y}+m}\left[1 - e^{-(f_{a,y}+m)}\right]\right) + \epsilon_{a,y} \ . \tag{7}$$

### 5.3.1 Parameter estimation

As with the previous example, nonlinear least squares can be used to minimize the objective function arising from the observation equation, which in this case is

$$\sum_{a,y}\left(\log(C_{a,y}) - \log(n_{a,y}) - \log\left(\frac{f_{a,y}}{f_{a,y}+m}\left[1 - e^{-(f_{a,y}+m)}\right]\right)\right)^2$$

The parameters to be estimated are the initial population numbers, recruits, and (instantaneous) fishing mortalities $f_{a,y}$. The model is therefore overparameterized and it is necessary to place some structure on the fishing mortalities.

It is probably reasonable to assume that the instantaneous fishing mortality on age $a$ fish in year $y$ is a product of the fishing effort in year $y$ and the age-specific selectivity. (This is known as the separable mortality assumption.) That is,

$$f_{a,y} = q_a E_y \ . \tag{8}$$

In principle, the fit could estimate the $q_a$ and $E_y$ quantities. However, if reliable effort data are available then they should be used. Similarly, in some cases, it may be assumed that the catchabilities, $q_a$, are "known" up to a constant, in which case one would specify

$$f_{a,y} = k q_a E_y \ .$$

## 5.4 Example 3: Fournier and Archibald's Method

Fournier and Archibald (1982) make the very pertinent point that the assumption of lognormally distributed catch-at-age (or research indice) may be quite inappropriate. In particular, catch-at-age data can take on extremely small, or even zero, values for older age groups (see Tables 5.1 and 5.2). Hence, the log of these data could be wildly sensitive to random variation.

The idea here is to use what we know about how the catch-at-age data are obtained. Essentially, if the aged fish are a random sample from the commercial catch, then the proportions at age obtained from this sample are multiplied by the total catch numbers to obtain catch-at-age numbers. Thus, it makes good sense to model the estimated proportions at age as from a binomial distribution, and to model the total catch numbers as lognormal (say).

If the numbers-at-age in the sample of aged fish are denoted $S_{a,y}$ and $p_{a,y}$ are the true proportions at age, then the log-likelihood term for the $S_{a,y}$ is

$$\sum_{a,y} S_{a,y} \log(p_{a,y}) \ .$$

Note that the true proportions at age are just

$$p_{a,y} = \frac{c_{a,y}}{c_y^+} \qquad \text{where} \qquad c_y^+ = \sum_a c_{a,y} \ .$$

The catches, $c_{a,y}$, are given by the Baranov catch equation (6).

### 5.4.1 Parameter estimation

Combining the log-likelihood term for the $S_{a,y}$ and a lognormal term for the total annual catch, $c_y^+$, gives the log-likelihood function

$$\sum_{a,y} S_{a,y} \log(p_{a,y}) + \frac{\sum_y (\log(C_y^+) - \log(c_y^+))^2}{\sigma^2} \tag{9}$$

As before, the quantities to be estimated are the inital population numbers, recruits, and $f_{a,y}$'s. It is likely that the $\sigma^2$ parameter in (9) would be specified rather than estimated.

## 5.5 Some variations

### 5.5.1 Parameter reduction

When reliable effort data are not available then it is necessary to estimate an effort parameter, $E_y$, for each year (see equation 8). Consequently, as more data are collected over time, the number of parameters also increases. This means that the nice optimality properties of maximum likelihood may not hold because these require the amount of data to increase while the number of parameters remains fixed.

One method of removing the need to estimate the $E_y$ in each year is to set the observed and expected total annual catches to be equal. That is

$$\sum_a C_{a,y} = \sum_a c_{a,y} = \sum_a n_{a,y} \frac{f_{a,y}}{f_{a,y} + m} \left[1 - e^{-(f_{a,y} + m)}\right] \ ,$$

where $f_{a,y} = q_a E_y$ and the equation on the right-hand side is Baranov's catch equation. Thus, given $n_{a,y}, a = 1, ..., A$, a value of $E_y$ can be found to satisfy the above equation. This can not be done explicitly and requires a numerical optimizer. One can then use $f_{a,y} = q_a E_y$ in the cohort dynamics equation (5) to determine the numbers at age in year $y+1$.

If total mortality is not too high then it may be reasonable to simplify the Baranov catch equation by using the fact that $f_{a,y} + m \approx 1 - e^{-(f_{a,y}+m)}$, and hence the above equation can be written

$$\sum_a C_{a,y} \approx \sum_a n_{a,y} f_{a,y} = E_y \sum_a n_{a,y} q_a \ .$$

That is,

$$E_y = \frac{\sum_a C_{a,y}}{\sum_a n_{a,y} q_a} \ .$$

Other approximations to the Baranov catch equation could be used. For example, the methodology employed by NIWA for the age-structured snapper model is based on the equation

$$\sum_a C_{a,y} \approx \sum_a n_{a,y} \left[ 1 - e^{-f_{a,y}} \right] \ .$$

### 5.5.2 Adding extra terms to the log-likelihood

The age-structured of these models permits many other types of observable data to be used. For example, because recruitment is modeled, it would be straightforward to include a sums-of-squares term to (9) if a recruitment indices were available. If reliable weight-at-age information were available then one could compare the average observed weight at age with the theoretical weight at age.

Another modification to the log-likelihood is given by assuming some annual variability in equation (8). For example

$$\log(f_{a,y}) = \log(k) + \log(q_a) + \log(E_y) + D_y \ ,$$

where $D_y$ is normally distributed. The term

$$\frac{\sum_y D_y^2}{\sigma_1^2}$$

is then added to the log-likelihood (with $\sigma_1^2$ probably specified) and the maximization is also with respect to the $D_y$. Note, though, that this extra term does not correspond in any way to a likelihood because there are no new data being used. It is another example of penalized likelihood, with a term that penalizes departures from the assumption of separable mortality. It could also be viewed as a kind of measurement error term, in that the $D_y$ could be interpretted as modeling the error in estimation of fishing effort.

### 5.5.3 Current research

**State-space models**

Gudmundsson (1994) and Schnute (1994) both added random variability to the cohort dynamics equation. Scnute's is a general methodology paper which covers both the theory and implementation of state-space and penalized-likelihood techniques. However, it does not actually apply these methodologies to data. (NB: Schnute refers to penalized-likelihood as "errors-in-variables")

Gudmundsson (1994) incorporates randomness in the dynamics by adding random terms to the separable mortality equation (8) and applies state-space methodology. This model was implemented using the extended Kalman filter and applied to Icelandic cod.

**Bayesian inference**

The models covered in this Chapter could also be fitted under the Bayesian paradigm. In particular, Millar and Meyer (1999) implemented a Bayesian state-space model that is somewhat in the spirit of Gudmundsson (1994) in the sense that it models random variability in mortality. This model was applied to the Northern cod data of Example 1. The use of MCMC avoided the computational difficulties associated with the non-Bayesian approaches.

## 5.6 Closing remarks

The complexity of age-structured models requires that a considerable effort be put into model implementation because it is unlikely to be achieved within a standard software procedure or function. This may result in little time left for the process of model building and checking.

There are many possible violations to the model assumptions, including ageing errors, within-year correlations in data, unaccounted mortality etc. Indeed, Richards and Schnute (1998) demonstrate an age-strucutured model which fits only to the mean age in each year and which should be less sensitive to ageing errors. Myers and Cadigan (1995a) present an analysis of the Northern cod data in which the VPA is extended to allow correlated errors among ages within a year for the research survey estimates of relative number at age.

## 5.7 References


Fournier, D. A., and C. P. Archibald. 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. 39: 1195-1207.

Gudmundsson, G. 1994. Time series analysis of catch-at-age observations. Appl. Statist. 43: 117-126.

Methot, R. D. 1990. Synthesis model: an adaptable framework for analysis of diverse stock assessment data. INPFC Bull. 50: 259-277.

Millar, R. B. and R. Meyer. 1999. Bayesian state-space modeling of age-structured data: fitting a model is just the beginning Can. J. Fish. Aquat. Sci. 56: in press.

Myers, R. A., and N. G. Cadigan. 1995a. Statistical analysis of catch-at-age data with correlated errors. Can. J. Fish. Aquat. Sci. 52: 1265-1273.

Myers, R. A., and N. G. Cadigan. 1995b. Was an increase in natural mortality responsible for the collapse of northern cod? Can. J. Fish. Aquat. Sci. 52: 1274-1285.

Richards, L. J., and Schnute, J. T. (1998) Model complexity and catch-age analysis. *Can. J. Fish. Aquat. Sci.* **55**, 949-957.

Schnute, J. T. 1994. A general framework for developing sequential fisheries models. Can. J. Fish. Aquat. Sci. 51: 1676-1688.

# 6 Bayesian inference

The fitting methodology we have employed so far is maximum likelihood based. Historically, this approach has been used in the vast majority of statistical analyses and inferences. However, Bayesian inference has made a strong re-appearance over the last few years.

The Bayesian approach to statistics was, in fact, the original way that was proposed for thinking about probability/evidence etc. It fell out of favour in the 1920's when Sir Ronald Fisher established the idea of repeatability. Fisher's approach did away with the nuisance of having to specify prior knowledge, and it also proved to be more tractable mathematically.

However, there have always been some statisticians who have maintained that the Bayesian approach is the only logically consistent way to utilize the information that is contained in data. Moreover, a new generation of powerful numerical techniques have been developed, with the consequence that it is now easier to implement complex Bayesian models than frequentist models.

In the fisheries context, it is also the case that some scientists view the use of prior information as a strength of the Bayesian approach. For example, Hilborn and Liermann (1998) argue that we are well advised to make use of the experiences and knowledge of fisheries scientists who have gone before us. These notions can be formalized through the use of meta-analysis to obtain "formal" priors from the analysis of existing data on similar species.

The reader should attempt section 6.1 and as much of 6.2 as is palatable. Sections 6.3 and 6.4 are more mathematical and are included only for completeness, and may be omitted if desired.

## 6.1 Questioning the frequentist approach

**Example 6.1** Let $X_1$ and $X_2$ be iid discrete random variables with density function

$$f(x;\theta) = P_\theta(X = x) = \begin{cases} 0.5, & x = \theta - 1 \\ 0.5, & x = \theta + 1 \end{cases}$$

where $-\infty < \theta < \infty$ is an unknown parameter to be estimated. A 75% confidence "interval" for $\theta$ is given by

$$[L(X_1, X_2), R(X_1, X_2)] = \begin{cases} \text{the point } \frac{X_1+X_2}{2}, & \text{if } X_1 \neq X_2 \\ \text{the point } X_1 - 1, & \text{if } X_1 = X_2 \end{cases}$$

because, if used under repetition of the experiment, it contains $\theta$ with probability 0.75.

However, if $X_1 \neq X_2$ then we can be certain that our 75% CI contains $\theta$ and it would seem more appropriate to say that we are 100% confident about the CI. Similarly, If $X_1 = X_2$ then we feel less than 75% "confident".

**Example 6.2** Consider the situation of a laboratory with two weighing machines. Both machines are unbiased, but the old machine has standard deviation of 1 gm and the new machine of 0.01 gm. The laboratory uses either machine with probability 0.5.

If our sample is weighed on the new machine, do we draw our inference on the basis of a standard deviation of 0.01 gm, or should we consider that under repetition of the experiment there is a 0.5 probability of using the old machine?

**Example 6.3** Consider the most powerful test of $H_0$ vs $H_a$ where the density functions for $X \in \{1, 2, 3\}$ under the null and alternative hypotheses are

| | $x = 1$ | $x = 2$ | $x = 3$ |
|---|---|---|---|
| $P_0(X = x)$ | .009 | .001 | .99 |
| $P_a(X = x)$ | .001 | .989 | .01 |

The test with critical region $\{1, 2\}$ is the most powerful size 0.01 test. However, it doesn't seem very sensible to reject $H_0$ when $x = 1$ because the outcome $x = 1$ is actually nine times more likely under $H_0$ than $H_a$?!?

**Example 6.4** Likelihood Principle: "All of the information contained in the data is contained in the likelihood function".

The likelihood principle (which is accepted as being sensible by many statisticians) is at odds with the frequentist approach.

## 6.2 The Bayesian model

Bayesian statistics uses probability to directly specify degrees of belief. This is achieved by adding one further ingredient to the model - a probability distribution on the parameter space, $\Omega$.

Prior to observing any data, the probability distribution on $\Omega$ is called the *prior* distribution, with density $f(\boldsymbol{\theta})$. It can be viewed as quantifying our knowledge of $\boldsymbol{\theta}$ prior to data collection. After observing the data, $\mathbf{x}$, the likelihood for the data is calculated and Bayes' theorem is used to update our knowledge/belief about $\boldsymbol{\theta}$ - this is quantified by the *posterior* density $f(\boldsymbol{\theta}|\mathbf{x})$.

We implicity use this approach to inference everyday.

**Example 6.5** The enthralled Otago student notices something small and white fall past the window. What was the object??? Some possibilities are:

1. Snow flake. The data is "observation of something small and white falling past the window", and certainly, a snow flake has a high likelihood of appearing small and white. Also, being Dunedin, a snow flake has a relatively high prior probability of being outside a window (at any time of year!).

   $\Rightarrow$ Reasonably high posterior probability that the object was a snow flake.

2. Cherry blossom. Also very likely to appear small and white. In spring a cherry blossom would also have high prior probability, but not otherwise.

   $\Rightarrow$ Reasonably high posterior probability that the object was a cherry blossom if it is spring, low otherwise.

3. Computer. Not very likely to appear small and white (unless the student just caught a glance of it as it whizzed by). Being on campus, there may be a modest prior probability of a computer being thrown from a window.
   $\Rightarrow$ Low posterior probability that object was a computer.

4. Elephant. Not very likely to appear small and white. Also, very small prior probability of an elephant being outside a window in Dunedin.
   $\Rightarrow$ Extremely low posterior probability that the object was an elephant.

### 6.2.1 The posterior distribution

In earlier chapters we denoted the density function of $\mathbf{x}$ as $f(\mathbf{x};\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the unknown parameter(s). Here, we shall denote it as $f(\mathbf{x}|\boldsymbol{\theta})$ to emphasize that $\boldsymbol{\theta}$ is now considered to be a random variable and that the density function of $\mathbf{x}$ is conditional on the value of the random variable $\boldsymbol{\theta}$.

The posterior density function is obtained from the prior density $f(\boldsymbol{\theta})$ and the likelihood $L(\boldsymbol{\theta}) \equiv f(\mathbf{x}|\boldsymbol{\theta})$ by Bayes' theorem (1763),

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})}{f(\mathbf{x})} = \frac{f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})}{\int_\Omega f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})d\boldsymbol{\theta}} \;. \tag{1}$$

**Example 6.6** Let $X|\mu$ be distributed $N(\mu,\sigma^2)$ where $\sigma$ is known and let the prior distribution for $\mu$ be $N(\nu,\tau^2)$. Then, $\mu|x$ is distributed $N(m,v)$ where

$$m = \frac{x/\sigma^2 + \nu/\tau^2}{1/\sigma^2 + 1/\tau^2} \quad \text{and} \quad v = (1/\sigma^2 + 1/\tau^2)^{-1} \;.$$

**Exercise 6.1** Let $X|p$ be binomial$(n,p)$ and let the prior for $p$ be a beta distribution with density function

$$f(p) = \frac{p^{(a-1)}(1-p)^{(b-1)}}{B(a,b)}, \quad 0 < p < 1,$$

for some $a, b > 0$. Show that $p|x$ also has a beta distribution and determine its parameters.

**Exercise 6.2** Let $X|\lambda$ be Poisson($\lambda$) and let the prior for $\lambda$ be $\Gamma(\alpha, \beta)$. Show that the $\lambda|x$ is also $\Gamma$ distributed and determine its parameters.

**Exercise 6.3** Let $X|\lambda$ be Exp($\lambda$) and let the prior for $\lambda$ be inverse gamma with parameters $\alpha$ and $\beta$. That is, $1/\lambda$ is distributed $\Gamma(\alpha, \beta)$.

(a). Determine the (inverse gamma) prior density function for $\lambda$.

(b). Show that $\lambda|x$ is also inverse gamma and with parameters $\alpha+1$ and $\beta/(x\beta+1)$.

The above example and exercises demonstrated the use of conjugate families.

**Definition 6.1 (Conjugate family)** *If the prior density $f(\boldsymbol{\theta})$ and the posterior density $f(\boldsymbol{\theta}|\mathbf{x}) \propto f(\boldsymbol{\theta})f(\mathbf{x}|\boldsymbol{\theta})$ always belong to the same family of distributions (regardless of $\mathbf{x}$) then that family of distributions is the* conjugate family *to the likelihood $f(\mathbf{x}|\boldsymbol{\theta})$.*

**Exercise 6.4** Let $X|\mu$ be distributed $N(\mu, \sigma^2)$ where $\sigma$ is known and let the prior distribution for $\mu$ be Exp($\lambda$). Calculate the density function of the posterior, including the constant of integration.

### 6.2.2 Interpretting the prior and posterior distributions

The frequentist development of probability is via repetition e.g. a probability is the long-run proportion of an event. In contrast, a Bayesian views probability as a subjective expression of knowledge about the unknown parameter. It is possible that two different Bayesian statisticians may use different priors and hence end up with different posteriors. This is the aspect of the Bayes method that frequentist statisticians find most unacceptable.

One possible solution is to use a *noninformative* prior as an expression of no prior knowledge about $\boldsymbol{\theta}$. For example, if $\theta$ is a location parameter (e.g. the mean of a normal distribution) with parameter space $\mathbb{R}$ then it is usual to use the noninformative prior $f(\theta) = 1, -\infty < \theta < \infty$. If $\theta$ is a scale parameter (e.g. the standard deviation of a normal distribution) then the usual noninformative prior is $f(\theta) = 1/\theta, \theta > 0$. These are examples of *improper* priors because they do not integrate to 1.

If the parameter space is finite with $N$ elements then a natural noninformative prior is to set $f(\theta) = 1/N$ for all $\theta \in \Omega$.

**Example 6.7** Dice throw. The obvious prior for a fair dice is $f(\theta) = 1/6, \theta \in \{1, 2, 3, 4, 5, 6\}$ and this specifies our (subjective) believe about the outcome of a coin toss. Suppose the coin is tossed and we are told that the outcome is an even number. What is our posterior believe now?

For $\theta = 2, 4$, or $6$,

$$f(\theta|\text{even}) = \frac{f(\theta)f(\text{even}|\theta)}{f(\text{even})} = \frac{1/6 \times 1}{1/2} = 1/3$$

and for $\theta = 1, 3$ or $5$, $f(\theta|\text{even}) = 0$ because then $f(\text{even}|\theta) = 0$.

**Exercise 6.5** Example 7.1 revisited. Using a noninformative prior on the location parameter $\theta$, determine the posterior density functions $f(\theta|x_1 \neq x_2)$ and $f(\theta|x_1 = x_2)$. Are they consistent with intuition??

**Exercise 6.6** Let $X|\lambda$ be $\text{Exp}(\lambda)$ and suppose we let $f(\lambda)$ be the improper noninformative prior for scale parameters, $f(\lambda) = 1/\lambda, \lambda > 0$. Show that $\lambda|x$ has an inverse gamma distribution, i.e., the distribution of $1/\lambda$, given $x$, is gamma.

It is sometimes possible to give the prior distribution a frequentist interpretation. For example, the scenario of Exercise 7.4 arises in estimation of earthquake magnitude. The (prior) distribution of earthquake magnitudes on the Richter scale is well modelled as exponential. The observation $X|\mu$ is the estimated magnitude (of an earthquake of true magnitude $\mu$) and can be assumed to be normally distributed with mean $\mu$. The posterior distribution $\mu|x$ then has a frequentist interpretation because we can talk about repeat experiments (from the population of earthquakes). This kind of approach is quite widely used and is known as *empirical Bayes* (or superpopulation modeling; or mixture modeling). It is not viewed as an acceptable approach by true Bayesian statisticians whom argue that probability *must* be viewed as a subjective statement of belief.

## 6.3 Bayesian inference

Inference is viewed as a decision problem. The decision could be specification of a single values of $\boldsymbol{\theta}$ (point estimation), or specification of an interval/region, or acceptance/rejection of a hypothesis.

Given observation of $\mathbf{x}$, let $L(d(\mathbf{x}), \boldsymbol{\theta})$ be the *loss function* associated with a decision $d(\mathbf{x})$ when $\boldsymbol{\theta}$ is the value of the parameter.

**Definition 6.2 Bayes decision.** *Given* $\mathbf{x}$*, the* Bayes decision *is the decision minimizing the posterior expected loss,* $E_{\mathbf{x}}(L(d(\mathbf{x}), \boldsymbol{\theta}))$*.*

The notation $E_{\mathbf{x}}$ is expectation over $\boldsymbol{\theta}$ given $\mathbf{x}$ and $E_{\mathbf{x}}(h(\mathbf{x}, \boldsymbol{\theta}))$ is equivalent to $E(h(\mathbf{X}, \boldsymbol{\theta})|\mathbf{X} = \mathbf{x})$. That is, the expectation is taken with respect to $f(\boldsymbol{\theta}|\mathbf{x})$.

### 6.3.1 Point estimation

If $\theta \in \Omega \in \mathbb{R}$ then a quadratic loss function would typically be used, $L(d, \theta) = (d - \theta)^2$. Then

$$E_{\mathbf{x}}([d(\mathbf{x}) - \theta]^2) = E_{\mathbf{x}}([d(\mathbf{x}) - E_{\mathbf{x}}(\theta)]^2) + E_{\mathbf{x}}([\theta - E_{\mathbf{x}}(\theta)]^2)$$

which is minimized by the Bayes decision estimator, $d(\mathbf{x}) = E_{\mathbf{x}}(\theta)$. This is just the expected value of $\boldsymbol{\theta}$ with respect to the posterior density $f(\boldsymbol{\theta}|\mathbf{x})$.

### 6.3.2 Interval or region estimation

Here, the set of possible decisions is the collection of all subsets of $\Omega$ and a possible loss function would be $1_{[\boldsymbol{\theta} \notin d(\mathbf{x})]}$ which penalises when $\boldsymbol{\theta}$ is not in the region $d(\mathbf{x})$. This is not a sensible loss function because it has expected posterior loss equal to $P(\boldsymbol{\theta} \notin d(\mathbf{x})|\mathbf{x})$ and therefore is always minimized by choosing $d(\mathbf{x}) = \Omega$.

It is necessary to penalize for width of $d(\mathbf{x})$ and a more reasonable loss function is given by

$$L(d,\boldsymbol{\theta}) = 1_{[\boldsymbol{\theta} \notin d]} + w(d)$$

where $w(d)$ is a loss due to the width of $d$. Then, the posterior expected loss is

$$E_{\mathbf{x}}(L(d,\boldsymbol{\theta})) = P(\boldsymbol{\theta} \notin d(\mathbf{x})|\mathbf{x}) + w(d) \ .$$

Note that for any fixed width, the posterior expected loss is obtained by a region of the form $\{\boldsymbol{\theta} : f(\boldsymbol{\theta}|\mathbf{x}) > a\}$, where $a$ depends on the width. Thus, the Bayes decision is a region of highest posterior density.

### 6.3.3 Hypothesis tests

Consider testing $H_0 : \boldsymbol{\theta} \in \Omega_0 \subset \Omega$ versus $H_1 : \boldsymbol{\theta} \in \Omega_1 = \Omega \setminus \Omega_0$. Only two decisions are available, $d_0$ (accept $H_0$) or $d_1$ (accept $H_1$). The loss function would typically be of the form

$$L(d_i,\boldsymbol{\theta}) = \begin{cases} 0, & \text{if } \boldsymbol{\theta} \in \Omega_i \\ c_i, & \text{if } \boldsymbol{\theta} \notin \Omega_i, \end{cases}$$

where $i = 0, 1$ and $c_i$ is the cost of incorrectly choosing hypothesis $H_i$.

**Exercise 6.7** Show that the Bayes decision is to reject $H_0$ (i.e., accept $H_1$) if $P(\boldsymbol{\theta} \in \Omega_0|\mathbf{x}) < c_0/(c_0 + c_1)$.

Remark: When $\boldsymbol{\theta}$ is a continuous random variable then the above test assumes that both hypotheses are composite.

## 6.4 Admissibility

In Chapter 4 we developed techniques for obtaining minimum variance unbiased estimators (if they existed) and we also looked at the notion of mean squared error (MSE). Here, we present the more general concept of risk and the criterion of admissibility. We see that Bayes decisions are admissible.

**Definition 6.3** *For a fixed $\boldsymbol{\theta}$, the* risk *of decision $d(\mathbf{x})$ is $R_{\boldsymbol{\theta}}(d) = E_{\boldsymbol{\theta}}(L(d(\mathbf{X}), \boldsymbol{\theta}))$ where the expectation is over values of $\mathbf{X}$.*

If quadratic loss is used for point estimation of $\theta \in \mathbb{R}$ then the risk of an estimator is its MSE,

$$\begin{aligned} R_\theta(d) = E_\theta([d(\mathbf{X}) - \theta]^2) &= E_\theta([d(\mathbf{X}) - E_\theta(d(\mathbf{X}))]^2) + E_\theta([E_\theta(d(\mathbf{X})) - \theta]^2) \\ &= Var_\theta(d(\mathbf{X})) + bias_\theta^2(d(\mathbf{X})) \ . \end{aligned}$$

**Definition 6.4** *A decision rule $d(\mathbf{x})$ is said to be* inadmissible *if there exists another decision rule $d(\mathbf{x})$ such that*

$$R_{\boldsymbol{\theta}}(d^*) \leq R_{\boldsymbol{\theta}}(d) \quad \forall \boldsymbol{\theta} \in \Omega$$

*with strict inequality for at least one value of $\boldsymbol{\theta}$.*

*A decision rule is said to be* admissible *if it is not inadmissible.*

It is clear that only admissible decisions are worthy of consideration.

**Theorem.** Bayes decisions minimize the expected (over $\boldsymbol{\theta}$) risk.

**Proof.** It can be shown (Exercise) that $E(h(X,Y)) = E(E(h(X,Y)|Y)) = E(E(h(X,Y)|X))$, or, using the notation of this chapter, $E(h(X,Y)) = E(E_Y(h(X,Y))) = E(E_X(h(X,Y)))$. Hence,

$$\begin{aligned} E(R(\boldsymbol{\theta})) &= E(E_{\boldsymbol{\theta}}(L(d(\mathbf{X}), \boldsymbol{\theta}))) \\ &= E(E_{\mathbf{X}}(L(d(\mathbf{X}), \boldsymbol{\theta}))) \ . \qquad (2) \end{aligned}$$

The Bayes decision minimizes $E_{\mathbf{X}}(L(d(\mathbf{X}), \boldsymbol{\theta}))$ for all values of $\mathbf{X}$, and hence minimizes (2).

**Corollary.** Bayes decisions are admissible.

**Proof.** Suppose not. Then there exists a decision $d^*(\mathbf{x})$ with lower risk for at least one value of $\boldsymbol{\theta}$. But, then $d^*(\mathbf{x})$ has lower expected risk, contradicting the above theorem.

[The proof is not complete because $d^*(\mathbf{x})$ could have lower risk at just one $\boldsymbol{\theta}$ value (say), and if $\boldsymbol{\theta}$ is a continuous random variable then $d(\mathbf{x})$ and $d^*(\mathbf{x})$ will still have the same expected risk. However - we don't really care about differences in decision risk on sets or probability zero. Moreover, risk functions will typically be continuous functions of $\boldsymbol{\theta}$ and values of $\boldsymbol{\theta}$ for which one risk function betters another will therefore constitute an interval or region.]

## 6.5 MCMC

The (joint) posterior density is given by equation (1) in section 6.2. In practice, $\boldsymbol{\theta}$ may be of high dimension, but one may just be interested in some quantities such as $E(\theta_1|\mathbf{x})$, calculation of which would then require a high dimensional integration. A similar problem is faced if one wishes to obtain the marginal density of any of the unknowns.

Markov Chain Monte Carlo (MCMC) is a general numerical technique for sampling from the joint posterior distribution. An estimate of $E(\theta_1|\mathbf{x})$ can then be obtained as the average of the $\theta_1$ values in this sample.

### 6.5.1 Metropolis-Hastings algorithm

Suppose we wish to generate a sample from a distribution with density function $g(\boldsymbol{\theta})$. Let a starting value $\boldsymbol{\theta}^{(0)}$ be given, and let $q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)})$ be the proposal density. The sample is generated as follows: the candidate value $\boldsymbol{\phi}$ is accepted with probability

$$\alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\phi}) = \min\left(1, \frac{g(\boldsymbol{\phi})q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\phi})}{g(\boldsymbol{\theta}^{(t)})q(\boldsymbol{\phi}|\boldsymbol{\theta}^{(t)})}\right) \ . \tag{3}$$

It accepted, the next value is $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\phi}$, otherwise there is no move, i.e., $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$.

**Exercise 6.8** From equation (3) show that

$$g(\boldsymbol{\theta}^{(t)})q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)})\alpha(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}) = g(\boldsymbol{\theta}^{(t+1})q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t+1)})\alpha(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)}) \ .$$

The above Exercise establishes that

$$g(\boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) = g(\boldsymbol{\theta}^{(t+1)})p(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t+1)}) \ . \tag{4}$$

Note that if $\boldsymbol{\theta}^{(t)}$ is distributed according to $g$, then the left-hand side of equation (4) is the joint density of $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)})$. Integrating the joint density with respect to $\boldsymbol{\theta}^{(t)}$ gives

$$\int g(\boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)})d\boldsymbol{\theta}^{(t)} = \int g(\boldsymbol{\theta}^{(t+1)})p(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t+1)})d\boldsymbol{\theta}^{(t)} = g(\boldsymbol{\theta}^{(t+1)}) \ .$$

That is, if $\boldsymbol{\theta}^{(t)}$ is distributed according to $g$, then the value $\boldsymbol{\theta}^{(t+1)}$ obtained by the Metropolis-Hastings algorithm is also distributed according to $g$.

### 6.5.2 Gibbs Sampler

The Gibbs sampler (Gelfand and Smith, 1990) is a numerical technique for sampling from the joint posterior distribution, $f(\theta_1, \theta_2, ..., \theta_n|\mathbf{x})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_n)$ are the unknowns and $\mathbf{x}$ denotes the observables. Given a starting vector $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \ldots, \theta_n^{(0)})$ the Gibbs sampler proceeds by sampling from the univariate full-conditional posteriors as follows

$$\begin{aligned}
\text{simulate } \theta_1^{(1)} &\sim f(\theta_1|\theta_2^{(0)}, \ldots, \theta_n^{(0)}, \mathbf{x}) \\
\text{simulate } \theta_2^{(1)} &\sim f(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_n^{(0)}, \mathbf{x}) \\
&\vdots \\
\text{simulate } \theta_n^{(1)} &\sim f(\theta_n|\theta_1^{(1)} \ldots, \theta_{n-1}^{(1)}, \mathbf{x})
\end{aligned}$$

and yields $\boldsymbol{\theta}^{(m)} = (\theta_1^{(m)}, \ldots, \theta_n^{(m)})$ after $m$ such cycles. This defines a Markov chain with transition kernel $k(\boldsymbol{\theta}^{(m+1)}, \boldsymbol{\theta}^{(m)}) = \prod_{i=1}^n f(\theta_i^{(m+1)}|\theta_1^{(m+1)}, \ldots, \theta_{i-1}^{(m+1)}, \theta_{i+1}^{(m)}, \ldots, \theta_n^{(m)}, \mathbf{x})$, that, under mild conditions, converges to the joint posterior as its equilibrium distribution (see Gilks et al., 1996). More generally, it is enough to just sample each full conditional using a Metropolis-Hastings step (Gilks, 1996) which is convenient if the full conditionals are of non-standard form. This technique is known as Metropolis-Hastings within Gibbs (MH-Gibbs) sampling, or alternatively, as single-component Metropolis-Hastings.

To see why the Gibbs sampler works, consider the bivariate case where we wish to sample from a bivariate density $f(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \theta_2)$. The equation

$$f(\theta_1, \theta_2) = f(\theta_1) f(\theta_2|\theta_1)$$

can be interpretted as saying that if $\theta_1$ is distributed according to $f(\theta_1)$ and, $\theta_2$ is then chosen conditional on $\theta_1$, then the pair $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is jointly distributed $f(\boldsymbol{\theta})$. In particular, $\theta_2$ is distributed according to $f(\theta_2)$ and so we can make use of the equation

$$f(\theta_1, \theta_2) = f(\theta_2) f(\theta_1|\theta_2)$$

to say that if we now sample $\theta_1$ conditional on $\theta_2$ then we have another observation from $f(\theta_1, \theta_2)$...., and so on, successively generating samples from $f(\boldsymbol{\theta})$ by alternating the sampling from the two conditional distributions.

## 6.6 References


Gelfand, A. E., and A. F. M. Smith. 1990. Sampling-based approaches to calculating marginal densities. J. Amer. Stat. Assoc. 85: 398-409.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. Markov Chain Monte Carlo in Practice. Chapman and Hall, London. 486 pp.

Hilborn, R. and M. Liermann. 1998. Standing on the shoulders of giants: learning from experience in fisheries. Rev. Fish. Bio. Fish. 8: 273-283.

Punt, A. E., and R. Hilborn. 1997. Fisheries stock assessment and decision analysis: the Bayesian approach. Rev. Fish. Bio. Fish. 7: 35-63.

# 7 Appendices

## 7.1 Automatic Differentiation Model Builder, ADMB

The ADMB package from Otter Research Ltd (http://otter-rsch.com/admodel.htm) has gained a strong following from many fisheries research groups worldwide and is being used by many leading fisheries scientists. It is particularly strong in the Pacific Northwest region where it was developed, but is also being used by Atlantic fisheries agencies, International Pacific Halibut Commission, CSIRO and in NZ by NIWA and the Sea Food Industry Council.

ADMB is notorious for being a challenge to use. The reasons for this include the need to be moderately conversant with C++, the limited documentation, and the sheer power of ADMB to fit models! It is the latter reason which has earned ADMB the support it currently has - it appears to be (by far) the most powerful piece of software for fitting complex models and it can be surprisingly fast. Moreover, at the "flick of a switch" it can perform Bayesian inference by doing Metropolis-Hastings sampling from the posterior.

### 7.1.1 What is automatic differentiation?

Essentially, automatic differentiation is just repeated application of the chain rule for differentiation. By way of example, suppose that one had catch and effort data and wished to use the model

$$c_t = b_t(1 - e^{-qE_t}) \ .$$

If multiplicative lognormal errors are assumed then the estimates of $q$ and the parameters determining $b_t$ are obtained by minimization of

$$f = \sum_t \Big(\log(C_t) - \log(b_t) - \log(1 - e^{-qE_t})\Big)^2 \ .$$

For demonstration purposes, it will suffice to consider the automatic differentiation of a single component of the above objective function, say

$$f_t = \Big(\log(C_t) - \log(b_t) - \log(1 - e^{-qE_t})\Big)^2 \ .$$

Then,

$$
\begin{aligned}
\frac{\partial f_t}{\partial q} &= \frac{\partial f_t}{\partial\left(\log(C_t)-\log(b_t)-\log(1-e^{-qE_t})\right)} \\
&\quad \times \frac{\partial\left(\log(C_t)-\log(b_t)-\log(1-e^{-qE_t})\right)}{\partial \log(1-e^{-qE_t})} \\
&\qquad \times \frac{\partial \log(1-e^{-qE_t})}{\partial(1-e^{-qE_t})}\frac{\partial(1-e^{-qE_t})}{\partial e^{-qE_t}}\frac{\partial e^{-qE_t}}{\partial(-qE_t)}\frac{\partial(-qE_t)}{\partial q}\frac{\partial q}{\partial q} \ .
\end{aligned}
$$

Note that in the above expansion of the derivative, the numerator and denominator differ by just a single unary or binary operation. Thus, if the operators are *defined* to not only calculate and return their value, but also to calculate and return their derivative, then derivatives can be calculated with just a moderate increase in computational effort.

The above example is rather simple because $q$ only appears once in the objective, but in principle the idea of exploiting the chain rule still applies. The following example is based on an example from the Autodiff manual of Otter Research Ltd,

$$f = \sin(\theta_1^2 + \theta_2^2) + e^{\theta_1 + 3\theta_2} \ .$$

Let $u = \theta_1^2 + \theta_2^2$ and $v = \theta_1 + 3\theta_2$ so that we can write $f = \sin(u) + e^v$. Now,

$$
\begin{aligned}
\frac{\partial f}{\partial \theta_1} &= \frac{\partial f}{\partial \sin(u)}\frac{\partial \sin(u)}{\partial \theta_1} + \frac{\partial f}{\partial e^v}\frac{\partial e^v}{\partial \theta_1} \\
&= \frac{\partial f}{\partial \sin(u)}\frac{\partial \sin(u)}{\partial u}\frac{\partial u}{\partial \theta_1^2}\frac{\partial \theta_1^2}{\partial \theta_1}\frac{\partial \theta_1}{\partial \theta_1} + \frac{\partial f}{\partial e^v}\frac{\partial e^v}{\partial v}\frac{\partial v}{\partial \theta_1}\frac{\partial \theta_1}{\partial \theta_1} \\
&= 1 \times \cos(u) \times 2\theta_1 \times 1 + 1 \times e^v \times 1 \times 1 \ .
\end{aligned}
$$

**Exercise 7.1** Repeat the above automatic differentiation for $\frac{\partial f}{\partial \theta_2}$. □

In ADMB, the automatic differentiation has been implemented for a wide range of operations, including matrix operations.

## 7.2 The Newton-Raphson algorithm

For simplicity, consider the situation where we want to find the minimum of $f(\theta)$ where $\theta$ is in some subset of $\mathbb{R}$. Assume also that $f$ is twice differentiable and has only one extrema, corresponding to a unique minimum at $\theta_0$. The idea of the Newton-Raphson algorithm is to find the value of $\theta$ for which the derivative of $f$ is zero (which is $\theta_0$).

Let $\theta^{(i)}$ be the current estimate of $\theta_0$. The N-R algorithm is based on a Taylor's series expansion of $f'(\theta_0)$ about $\theta^{(i)}$. That is,

$$0 = f'(\theta_0) \approx f'(\theta^{(i)}) + f''(\theta^{(i)})(\theta_0 - \theta^{(i)}) \ .$$

Solving this for $\theta_0$ gives

$$\theta_0 \approx \theta^{(i)} - \frac{f'(\theta^{(i)})}{f''(\theta^{(i)})}$$

which gives the next estimate,

$$\theta^{(i+1)} = \theta^{(i)} - \frac{f'(\theta^{(i)})}{f''(\theta^{(i)})} \ .$$

If $f$ is a quadratic function then the minimum is found in one iteration.

More generally, $\boldsymbol{\theta}$ will be a vector. The derivation of the N-R algorithm is unchanged, except that the Taylor's series expansion is now expressed in terms of derivative vectors and Hessian matrices (matrices containing the second derivatives of $f$ with respect to $\boldsymbol{\theta}$). The N-R algorithm is then

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - f''(\boldsymbol{\theta}^{(i)})^{-1} f'(\boldsymbol{\theta}^{(i)}) \ .$$

One can regard $f''(\boldsymbol{\theta}^{(i)})^{-1} f'(\boldsymbol{\theta}^{(i)})$ as the direction in which to move, but sometimes it is better not to move the whole step and instead to consider

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - \lambda_i f''(\boldsymbol{\theta}^{(i)})^{-1} f'(\boldsymbol{\theta}^{(i)}) \ ,$$

where a reasonable value of $\lambda_i$ can be determined by a quick one-dimensional search.

## 7.3 Inference from maximum likelihood

This appendix briefly looks at how to estimate (approximate) standard errors and confidence intervals for the maximum likelihood estimates or for functions of the MLE's. It is assumed that the parameter vector is of size $s$, i.e., $\boldsymbol{\theta} \in \mathbb{R}^s$.

The variance matrix of the estimates can be estimated from the Hessian matrix of the log-likelihood, which is immediately available if the N-R algorithm was used to find the MLE by maximimizing the log-likelihood. The negative of the Hessian matrix is known as the (observed) information matrix, $I(\hat{\boldsymbol{\theta}})$. The estimated variance matrix is simply the inverse of the information matrix. This variance matrix is asymptotically correct, in the sense that the MLE's converge in distribution (at a $n^{1/2}$ rate) to a normal distribution having that variance matrix. Thus, it is approximate for finite sample sizes.

Confidence intervals and hypothesis tests can be performed using the estimated variance matrix (known as the Wald test), however in general it is usually safer to perform a likelihood ratio test instead.

We consider two cases

Simple hypothesis: $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. That is, we wish to test that $\boldsymbol{\theta}$ is a particular value.

Composite hypothesis: $H_0 : \boldsymbol{\theta} \in \Theta_0 \subset \Theta$ where $\Theta_0$ is an $s - r$ dimensional subset of $\Theta$. We shall assume (reparameterizing if necessary) that $\boldsymbol{\theta} = \begin{pmatrix} \psi \\ \lambda \end{pmatrix}$ where $\psi \in \mathbb{R}^r$, $\lambda \in \mathbb{R}^{s-r}$, and $\boldsymbol{\theta} \in \Theta_0$ are all points in $\Theta$ for which $\psi$ equals some specified value $\psi_0$.

### Simple hypothesis

Under appropriate conditions: under $H_0$

$$\text{LR}: \qquad 2[l(\hat{\boldsymbol{\theta}}_n; \mathbf{X}) - l(\boldsymbol{\theta}_0; \mathbf{X})] \to_D \chi^2_s$$

$$\text{Wald}: \qquad (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T I(\hat{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \to_D \chi^2_s$$

**Composite hypothesis**

If $\hat{\boldsymbol{\theta}}_n$ is denoted $(\hat{\psi}_n, \hat{\lambda}_n)$ and $\hat{\boldsymbol{\theta}}_{0n} = (\psi_0, \hat{\lambda}_{0n})$ is the ML estimator in $\Theta_0$ then (under appropriate conditions),

$$2[l(\hat{\boldsymbol{\theta}}_n; \mathbf{X}) - l(\hat{\boldsymbol{\theta}}_{0n}; \mathbf{X})] \to_D \chi^2_r$$

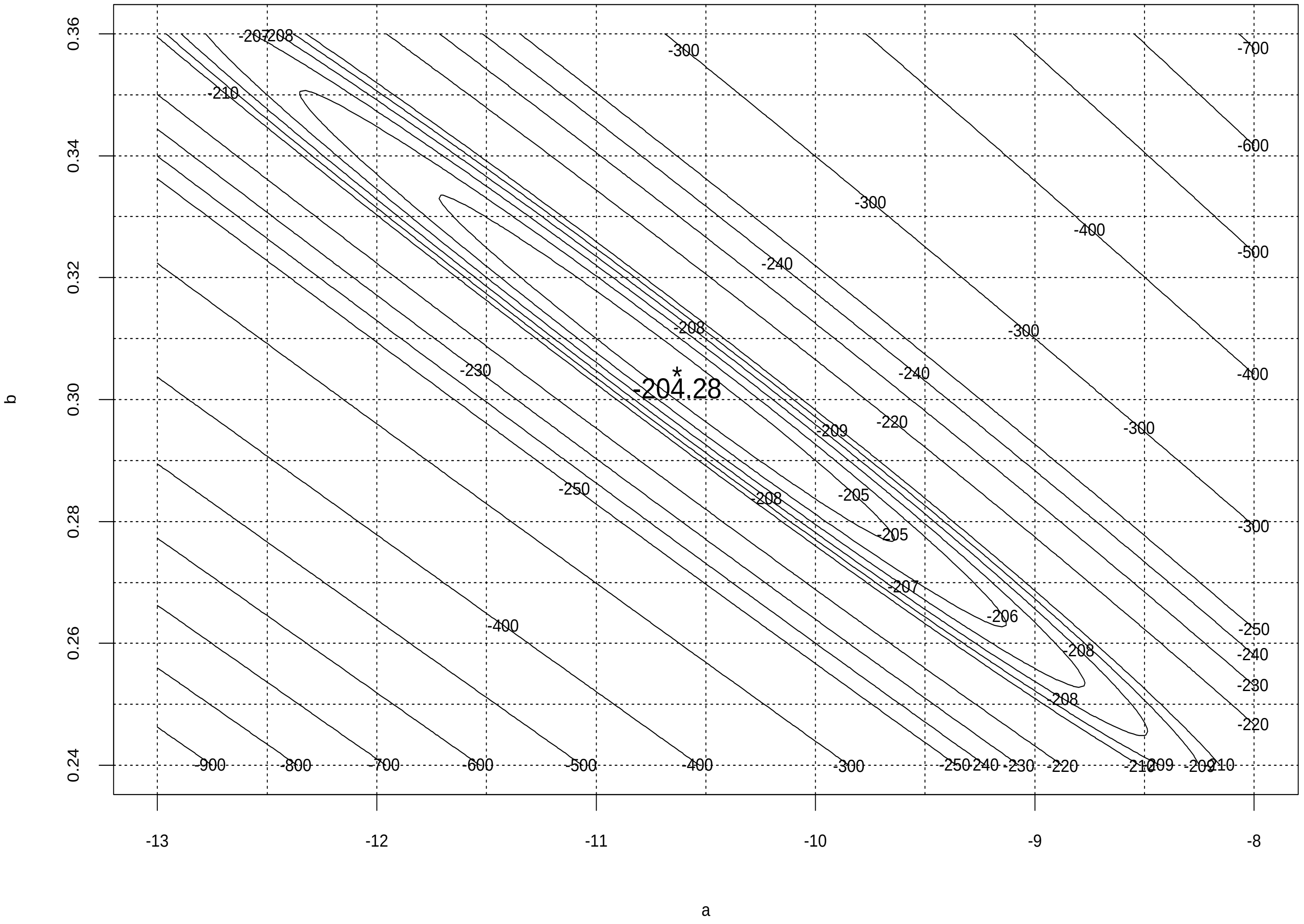
when $H_0$ is true.

**Example 7.1** Modeling catches from a covered cod-end selectivity experiment. A logistic regression was fitted to binomial data corresponding to the retention proportions of haddock of differing size retained in a cod-end with 113 mm mesh. A contour plot for the two parameters $a$ and $b$ is shown below.

Using the likelihood ratio statistic:

1. Test $H_0 : (a, b) = (-10, 0.3)$ using $\alpha = 0.05$.
2. Determine an approximate 95% CI for parameter $a$.

# Log-likelihood contours for 113 mm mesh codend



For the composite hypothesis the Wald test statistic becomes

$$(\hat{\psi}_n - \psi_0)^T [[I^{-1}(\hat{\boldsymbol{\theta}}_n)]_{\psi\psi}]^{-1} (\hat{\psi}_n - \psi_0) \to_D \chi^2_r$$

where $[I^{-1}(\hat{\boldsymbol{\theta}}_n)]_{\psi\psi}]$ is the upper $r \times r$ submatrix of $I^{-1}(\hat{\boldsymbol{\theta}}_n)$ (i.e., the asymptotic covariance matrix of $\hat{\psi}_n$) and $[I^{-1}(\hat{\boldsymbol{\theta}}_n)]_{\psi\psi}]^{-1}$ is called the Fisher information for $\psi$, evaluated at $\hat{\boldsymbol{\theta}}_n$, in the presence of the "nuisance" parameter $\lambda$. It can be shown that $[I^{-1}(\hat{\boldsymbol{\theta}}_n)]_{\psi\psi}]^{-1} = I_{\psi\psi} - I_{\psi\lambda} I^{-1}_{\lambda\lambda} I_{\lambda\psi}$.

### 7.3.1 Likelihood profiles in ADMB

Note that performing a test or obtaining a confidence interval for a single parameter corresponds to the In the composite hypothesis case. This requires that the likelihood be maximized over a subspace of the parameter space, and hence typically requires additional programming effort. A nice feature of ADMB is that it can do this at the "flick of a switch".

### 7.3.2 Delta Theorem

Suppose $g : \mathbb{R}^s \to \mathbb{R}^p$ where

$$g(\boldsymbol{\theta}) = \begin{pmatrix} g_1(\boldsymbol{\theta}) \\ . \\ . \\ . \\ g_p(\boldsymbol{\theta}) \end{pmatrix}$$

and that each co-ordinate $g_i : \mathbb{R}^s \to \mathbb{R}$ has derivative $g_i' = (\frac{\partial g_i}{\partial \theta_1}, ..., \frac{\partial g_i}{\partial \theta_s})^T$ that is continuous at $\boldsymbol{\theta}_0$.

Then, the approximate variance matrix for $g(\boldsymbol{\theta})$ is

$$G(\boldsymbol{\theta}_0) I_1^{-1}(\boldsymbol{\theta}_0) G(\boldsymbol{\theta}_0)^T ,$$

where

$$G(\boldsymbol{\theta}_0) = \begin{pmatrix} g_1^{'}(\boldsymbol{\theta}_0)^T \\ . \\ . \\ . \\ g_p^{'}(\boldsymbol{\theta}_0)^T \end{pmatrix},$$

and $I^{-1}(\hat{\boldsymbol{\theta}})$ is the approximate variance matrix for $\boldsymbol{\theta}$.

## 7.4 Discrete versus instantaneous rates of mortality

Let $n_t$ denote the number of fish in a particular cohort at the start of year $t$. Then, one might write

$$n_{t+1} = s_t n_t$$

where $s_t$ is the proportion of the cohort surviving to the end of year $t$. If fishing occurs in a pulse (over a very short time interval) at the start of the year then one might partition $s_t$ as the product of the proportion surviving harvest and the proportion of these surviving through to the end of the year, $s_t = (1 - h_t)p_t$, say.

However, in many fisheries it is often more useful to consider mortality as a process that is spread throughout the year and this can be expressed via the differential equation

$$\frac{dn_t}{dt} = -z_t n_t \ . \qquad (1)$$

The mortality rate $z_t$ is known as the total instantaneous mortality and it may be partitioned into natural mortality and mortality caused by fishing, $z_t = m_t + f_t$, say. The latter is often assume to be proportional to the fishing effort.

Rearranging equation (1) and integrating over year $t$ gives

$$\int_{n_t}^{n_{t+1}} \frac{1}{n_t} dn_t = -\int_t^{t+1} z_t dt$$

giving

$$\log(n_{t+1}) - \log(n_t) = -z_t \ ,$$

which results in

$$n_{t+1} = n_t e^{-z_t} \ .$$

**Exercise 7.2** If total instantaneous mortality, $z$, is constant over time, show that $n_t = n_0 e^{-zt}$, where $n_0$ is the size of the cohort at time 0.