FISH 604
Franz Mueter

Homework 2

TOTAL: 50 points

Due Thursday September 23, 2021

Chatham Strait sablefish: data exploration and fitting a simple growth model in R

I have provided almost all of the code you need (which is replicated in the accompanying R script file for your convenience). You do not need to show me everything you do, but you should answer specific questions, indicated by → and show or summarize the main results as appropriate. Please summarize your results with figures for support and answers to questions (Word doc or pdf). Feel free to use Rmarkdown but render your output as pdf, html or Word.

Please save your homework as 'Hwk1_FirstName_LastName.*' and submit it via Canvas.

Feel free to work together with someone else or in small groups - help each other out if you get stuck or don't understand something. Discuss the concepts and approaches. However, write up the results independently and submit your own work.

Problem 1: (5 points) Data import & exploration: The file *sablefish.csv* contains length (in mm), weights (in kg), and ages (in years) of male and female Chatham Strait sablefish (courtesy of Sherri Dressel, ADF&G, Juneau) and the year they were sampled. First, import the data into R and do some basic data exploration:

a. Import sablefish data into R:
   i. Import the sablefish data file into your current R working directory and read data into R, for example:
      ```
      sable <- read.csv("sablefish.csv", as.is=T)
      ```
      This will create a new data object (a "data frame") in your R workspace called `'sable'`.
   ii. Explore the object you created, using some basic tools:
      ```
      head(sable)           # first few rows
      hist(sable$Age)       # age structure
      table(sable$Age)      # age structure in Table form
      ```
      You can get some basic summary stats for each variable by typing:
      ```
      summary(sable)
      ```
      Note that Sex should be stored as a 'factor':
      ```
      is.factor(sable$Sex)
      ```
      If not, make sure to convert the variable to a factor, for example:
      ```
      sable$Sex <- factor(sable$Sex)
      ```

→ Explore length distribution by plotting histograms or density plots of the length of sablefish overall and by sex.

→ Are there obvious differences in the length distribution between males and females?

Problem 2: (10 points) Data exploration: Plot length (on y-axis) versus age, weight vs. age, and weight vs. length of all fishes using different symbols or colors for males and females.

For example, you can use one of the following to plot length-at-age by sex:
a) Using the `plot()` function and subsets:

```
plot(Length ~ Age, data=sable, subset = Sex == "Female", col=2,
   xlab = "Age", ylab = "Length (mm)")
points(Length ~ Age, data = sable, subset = Sex == "Male", col=4)
```

(Note the double equal sign, which is a relational operator used to assess equality – The expression '`Sex == "Female"`' creates a logical vector that has the same length as the variable 'Sex' and which is TRUE wherever Sex is equal to "Female" and FALSE otherwise! You can see what this expression returns by typing `sable$Sex == "Female"` at the command prompt)

Note that many symbols are plotted on top of each other because lengths and ages are measured in 10 mm and 1 year increments, respectively. You can "jitter" the symbols by a small amount to get a better indication of the number of data points:

```
plot(jitter(Length) ~ jitter(Age), data=sable, subset = Sex == "Female",
      col=2, xlab = "Age", ylab = "Length (mm)")
points(jitter(Length) ~ jitter(Age+0.5), data = sable,
      subset = Sex == "Male", col=4)
```

(Note that I also offset the age for males by 0.5, so they don't overlap with females)

Or, use 'ggplot' or the 'lattice' package ('`xyplot()`') (see script file)

Similarly, examine length-at-age by year. To do so, it is helpful to also convert Year (by default stored as integers) to a factor variable:

```
sable$Year <- factor(sable$Year)
```

To get an idea of the approximate shape of the length-age and weight-length relationships, you could use the handy function '`scatter.smooth()`' to add a LOESS smooth to the scatterplots by sex. To do so, construct an index (for subscripting) that selects females only (repeat for males):

```
sub <- sable$Sex == "Female"
scatter.smooth(jitter(sable$Age[sub]), jitter(sable$Length[sub]), col=2,
   xlab = "Age", ylab = "Length (mm)")
```

See the script file for exploring plots of length-at-age with loess smooths by year and sex or using ggplot as an alternative (***Note that we will delve into the algorithm used by loess and other smoothers once we cover Generalized Additive Models***)

Repeat a similar exploration of weight-at-age (i.e. substitute 'weight' for 'length' above).

→ Show a few illustrative plots by saving graphical output or copying and pasting them into your results file. To paste into Word, the Windows metafile format usually works well! Add a figure caption that clearly states what the figures show.

→ Questions: Does there appear to be an asymptotic length and/or weight for males and females? Does there appear to be a difference in growth between males and females? Are there apparent differences in the length-weight relationship among years?

Problem 3: (10 points) Fit a Ludwig van Bertalanffy (LVB) growth model to the sablefish data for both sexes combined. This exercise is intended to illustrate the general model fitting approach that we will use throughout class with a relatively simple non-linear model. We will cover some of the theory and the practical details of fitting models in R in future lectures. **Therefore, don't worry if you don't understand the details of the model fitting procedure at this point! I provided all the code you need.**

The LVB growth model is a non-linear model that can be written as:

$$L(a) = L_\infty\left(1 - e^{(-k(a-a_0)}\right)$$

where *L(a)* is the length of an individual fish at age *a* (in years) and $L_{\infty,}$ $k_,$ and a_0 are parameters to be estimated. They reflect the maximum asymptotic length, the growth coefficient, and an x-axis intercept (= age at which length would be zero), respectively.

To use and then fit the model in R, it is convenient to first write a function to compute the predicted values given the age, *a*, and values for each of the three parameters. These are provided as "arguments" to the LvB function:

```
LvB <- function(a, k, L.inf, a0) {
      L.inf*(1-exp(-k*(a-a0)))
}
```

[*For those new to programming in R, after running the code above, you should have an object named LvB in your workspace (check by typing* `ls()` *to see if LvB is listed and / or type* `LvB` *to see the function that you created). You can now use the function as needed but it will only be available in your current R session or if you save the output and re-load it later*]

The model is fit to the data similar to the way any standard linear regression model is fit to data: by finding the parameter combination that minimizes the sum of the squared differences between the observed and predicted values (= residual sum of squares). Thus, it is an example of an (Ordinary) Least-Squares regression. Unlike in linear models, there is no analytical solution for the parameter estimates and we need to solve the problem numerically. To do so, we need to have starting values for the parameters.

→ Based on the previous graphical analyses (disregarding sexes, i.e. we fit the model to both sexes combined) and the interpretation of the parameters in the model, choose reasonable starting values for L_∞ and a_0 and justify your choice! Use 0.05 for the growth parameter k.

Create a named vector of starting values (names are required for the fitting function):

```
ST <- c(k = 0.05, L.inf = …, a0 = …)
```

(**fill in the blanks (…) with your chosen starting values**. The name ‘ST’ is arbitrary)

Check whether the starting values are reasonable by plotting the data and adding “predicted” values for ages 1-80, given your starting values:

```
plot(jitter(Length) ~ jitter(Age), data=sable, col=2)
lines(1:80, LvB(1:80, k = ST[1], L.inf = ST[2], a0 = ST[3]), lwd=3)
```

The line should at least go through the data. If not, try different starting values!

Now you’re ready to fit the model. The function `nls()` (for “non-linear least-squares)’ is one option to use for fitting non-linear models using a least-squares approach.

Save the results to a new object, for example ‘`fit`’:

```
fit <- nls(Length ~ LvB(Age, k, L.inf, a0), data = sable, start = ST)
```

Look at a brief summary of the results using the `summary()` function:

```
summary(fit)
```

which shows (among other things) the fitted parameter values along with their estimated standard errors and (approximate) t-tests that, for each parameter, test the null hypothesis that the parameter is equal to 0. Note that the output looks very similar to the familiar output from a linear model, **but the standard errors and t-tests in this case are only approximate and cannot always be trusted** (depending on how “non-linear” the model is).

We will examine output from a variety of different models in more detail later. For the time being, focus on the parameter estimates. You can extract the estimates using:

```
coef(fit)
```

Create a scatterplot of the data so we can add a fitted line, based on the estimated parameter values:

```
plot(jitter(Length) ~ jitter(Age), data=sable, col=2)
```

Extract estimated parameter values (=coefficients) from the model

```
cf <- coef(fit)
```

→ Use these coefficients to compute predicted values over a range of length values and add a fitted line to the scatterplot.

→ Construct a 95% confidence interval for the overall growth parameter and for the estimated mean length at infinity as $\hat{\theta} \pm 2 \cdot se(\hat{\theta})$ where $\hat{\theta}$ is the estimate for the parameter of interest and $se(\hat{\theta})$ is its standard error. The standard errors are in the summary output and you can extract them in one of two ways:

```
summary(fit)$coef  # Table of coefficients, stored as a matrix
summary(fit)$coef[1,2]  # Standard error for k (in row 1, column 2 of the output)
```

Or you can extract the covariance matrix of the model parameters (works with most model objects that we will encounter in class):

```
vcov(fit)   # Variance-covariance matrix
```

Remember that this matrix has the variances on the diagonal, hence the standard errors are:

```
sqrt(diag(vcov(fit)))
```

Problem 4 (10 points): Assessing the model fit / model diagnostics!

To get an unbiased estimate of the LvB parameters for sablefish, we “only” have to assume that the observations are iid, which means “independent” (i.e. random samples from the sablefish population where each fish has the same probability of being sampled) and “identically distributed”, i.e. that each observation has the same (unspecified) probability distribution.

If we wish to estimate standard errors and confidence intervals, we also have to assume that the observations follow not just any probability distribution but are normally distributed with a mean specified by the model (The estimated values for *L(a)* in our example) and some variance σ^2. In other words, we assume that the <u>residuals</u> have a normal distribution with mean 0 and variance σ^2. In shorthand notation, this assumption is typically written as: $\varepsilon \sim N(0, \sigma^2)$.

Thus, when making probabilistic statements about parameters or testing parameters for significance, we assume that observations are “iid normal”!

Examine these assumptions for the model that we fit above **(a single curve fit to males + females combined)** graphically by plotting the residual distribution overall (e.g. barplot, histogram), by year (i.e. separate histograms/barplots by year), and by sex.

First, extract the residuals from the fitted model object:

```
r <- resid(fit)      # where fit is the fitted model object (see above)
```

which are simply the observations minus the fitted values. You can confirm this by computing:

```
r2 <- sable$Length - LvB(sable$Age,k=cf[1],L.inf=cf[2], a0=cf[3])
```

and comparing r to r2. Here, ‘cf’ are the parameter estimates from the model fit to all data (a vector of length 3 with the three parameter estimates).

→ Plot the distribution of the residuals as histograms and boxplots (overall, by year, by sex).

For example:

```
boxplot(r ~ Year, data = sable)
abline(h=0, col=2)  # as reference line
```

→ Examine these plots and <u>based on a graphical examination only</u>, briefly discuss the assumptions that residuals (<u>for the combined model</u>) are normally distributed and that they are identically distributed (with age, across years and between sexes)! Show one or two representative plots to illustrate your conclusion(s).

→ If you find that any of the regression assumptions are violated, what might be a good approach to dealing with it in this case?

Problem 5 (15 points): Repeat the analysis of growth separately for males and for females for a single year of your choice (avoid 2006, as it may be problematic).

→ Show plots of male and female length-at-age with the fitted lines (you can plot them separately or in a single panel using any tools in your toolkit)

→ Summarize the parameter estimates and their standard errors for males and females and compute approximate 95% confidence intervals for the growth parameters by sex.

→ Statistically test for a difference in the growth parameter k between males (k_m) and females (k_f) using a Wald test by computing the Wald statistic for the difference ($k_f - k_m$) and test whether it is larger than expected by chance under the null hypothesis that the difference is zero.

The Wald statistic for a parameter estimate $\hat{\theta}$ is:

$$W = \frac{\left(\hat{\theta} - \theta_0\right)^2}{var\left(\hat{\theta}\right)}$$

Because we want to test for a difference between growth paramters, in this case $\hat{\theta} = k_f - k_m$ and the null hypothesis is that $\theta_0 = 0$. Under the null hypothesis, this statistic has an asymptotic χ^2 distribution with one degree of freedom (*Note that in linear models with normal errors, the square root of W has a t-distribution, which is the basis for the t-tests for each of the parameter estimates in the summary output from linear models*). For computing $var\left(\hat{\theta}\right)$, assume that the estimates of k_f and k_m are independent.

To test the null hypothesis, you need to evaluate the probability that a χ^2 - distributed random variable with one degree of freedom is larger than or equal to the observed W. Use ‘`pchisq()`’ to find the probability.

→ Repeat the test to compare the L_∞ parameter between male and female sablefish estimated from a single year.

(Note: you could easily compute W values and their p-values for all parameters simultaneously using vectors).

→ Briefly summarize your conclusions about growth of male and female sablefish.