Data analysis project

Jess Qualley and Rebecca Hansen

March 27, 2020

##Clear code (**delete before submitting**\*)

rm(list = ls(all = TRUE))

#Preparation ##Loading and citing packages

#Load faraway package  
library(faraway)  
  
#Load plyr package  
library(plyr)

##   
## Attaching package: 'plyr'

## The following object is masked from 'package:faraway':  
##   
## ozone

#Load tidyverse package  
library(tidyverse)

## ── Attaching packages ───────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✔ ggplot2 3.2.1 ✔ purrr 0.3.3  
## ✔ tibble 2.1.3 ✔ dplyr 0.8.3  
## ✔ tidyr 1.0.2 ✔ stringr 1.4.0  
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ── Conflicts ──────────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::arrange() masks plyr::arrange()  
## ✖ purrr::compact() masks plyr::compact()  
## ✖ dplyr::count() masks plyr::count()  
## ✖ dplyr::failwith() masks plyr::failwith()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::id() masks plyr::id()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::mutate() masks plyr::mutate()  
## ✖ dplyr::rename() masks plyr::rename()  
## ✖ dplyr::summarise() masks plyr::summarise()  
## ✖ dplyr::summarize() masks plyr::summarize()

#Load lmer package  
library(lme4)

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

#Cite packages  
citation("faraway")

##   
## To cite package 'faraway' in publications use:  
##   
## Julian Faraway (2016). faraway: Functions and Datasets for Books  
## by Julian Faraway. R package version 1.0.7.  
## https://CRAN.R-project.org/package=faraway  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Manual{,  
## title = {faraway: Functions and Datasets for Books by Julian Faraway},  
## author = {Julian Faraway},  
## year = {2016},  
## note = {R package version 1.0.7},  
## url = {https://CRAN.R-project.org/package=faraway},  
## }  
##   
## ATTENTION: This citation information has been auto-generated from  
## the package DESCRIPTION file and may need manual editing, see  
## 'help("citation")'.

citation("plyr")

##   
## To cite plyr in publications use:  
##   
## Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data  
## Analysis. Journal of Statistical Software, 40(1), 1-29. URL  
## http://www.jstatsoft.org/v40/i01/.  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Article{,  
## title = {The Split-Apply-Combine Strategy for Data Analysis},  
## author = {Hadley Wickham},  
## journal = {Journal of Statistical Software},  
## year = {2011},  
## volume = {40},  
## number = {1},  
## pages = {1--29},  
## url = {http://www.jstatsoft.org/v40/i01/},  
## }

citation("tidyverse")

##   
## Wickham et al., (2019). Welcome to the tidyverse. Journal of  
## Open Source Software, 4(43), 1686,  
## https://doi.org/10.21105/joss.01686  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Article{,  
## title = {Welcome to the {tidyverse}},  
## author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostino McGowan and Romain François and Garrett Grolemund and Alex Hayes and Lionel Henry and Jim Hester and Max Kuhn and Thomas Lin Pedersen and Evan Miller and Stephan Milton Bache and Kirill Müller and Jeroen Ooms and David Robinson and Dana Paige Seidel and Vitalie Spinu and Kohske Takahashi and Davis Vaughan and Claus Wilke and Kara Woo and Hiroaki Yutani},  
## year = {2019},  
## journal = {Journal of Open Source Software},  
## volume = {4},  
## number = {43},  
## pages = {1686},  
## doi = {10.21105/joss.01686},  
## }

##Read in and organize data

#Read from csv file  
herring <- read.csv("HerringOtolithDatabase 15 Jan 2020\_JQ.csv", header=TRUE, fileEncoding="UTF-8-BOM", stringsAsFactors = FALSE)  
  
#Create a data frame  
herring\_data <- data.frame(oto.width = herring$AverageWidth,  
 sal.id = herring$FishCode,  
 sal.length = herring$SalmonLength,  
 coll.doy = herring$CollectionDayofYear,  
 lat = herring$Latitude,  
 long = herring$Longitude,  
   
 sal.sp = herring$SalmonSpecies,  
 coll.year = herring$CollectionYear,  
 stat.area = as.factor(herring$StatArea),  
   
 digestion = herring$Digestion,  
 sal.sex = herring$SalmonSex,  
 coll.month = herring$CollectionMonth)  
  
#check data frame structure\* do we need to change data types for any variables? i.e. coll.doy, sal.id?  
str(herring\_data)

## 'data.frame': 883 obs. of 12 variables:  
## $ oto.width : num 1.65 1.49 1.6 1.39 1.76 ...  
## $ sal.id : int 17047 17050 17054 17060 17065 17096 17110 17111 17112 17112 ...  
## $ sal.length: num 68.5 66 65 62 68.5 ...  
## $ coll.doy : int 57 22 339 356 48 126 111 175 175 175 ...  
## $ lat : num 49.6 49.4 49.4 49.4 49.6 ...  
## $ long : num -123 -123 -123 -123 -123 ...  
## $ sal.sp : Factor w/ 4 levels "","ch","co","pink": 2 2 2 2 2 2 2 2 2 2 ...  
## $ coll.year : int 2017 2017 2016 2016 2017 2017 2017 2017 2017 2017 ...  
## $ stat.area : Factor w/ 15 levels "1","13","14",..: 12 12 12 12 12 8 13 4 4 4 ...  
## $ digestion : int 4 1 2 2 2 1 4 1 2 3 ...  
## $ sal.sex : Factor w/ 3 levels "","female","male": 3 1 3 3 2 2 1 1 1 1 ...  
## $ coll.month: int 2 1 12 12 2 5 4 6 6 6 ...

#figure out which stat areas to remove  
levels(as.factor(herring$StatArea))

## [1] "1" "13" "14" "15" "16" "17" "18" "19" "20" "23" "25"   
## [12] "28" "29" "101" "125"

#filter for 2018 data, chinook salmon only, omit northern BC areas (Haid Gwaii) and west coast Vancouver Island  
herring\_data <- herring\_data %>% filter(herring\_data$sal.sp == "ch" & herring\_data$coll.year == "2018" & herring\_data$stat.area != "1" & herring\_data$stat.area != "101" & herring\_data$stat.area != "125" & herring\_data$stat.area != "23" & herring\_data$stat.area != "25")

#Section 1: Introduction, Question, Goals and Hypotheses (400 words)

##A) Pacific Salmon in British Columbia are valued for their social, cultural and economic benefits to many communities on our coast. Despite their important role in the coastal marine ecosystem (Schindler et al. 2003), diet studies from this region are outdated and limited to summer sampling. The adult Chinook and Coho Salmon Diet Program (Juanes Lab) seeks to address these knowledge gaps by sourcing stomach samples from the recreational fishing community to gather regional, seasonal and interannual diet data to monitor ecosystem response to environmental change.

We have found Pacific Herring are the dominant fish consumed throughout regions of the Salish Sea, year-round. For this project we will investigate factors that may be important to the size of herring consumed by Chinook Salmon. Previous work suggests a positive, linear relationship between otolith width (mm) and length (mm) of herring from this dataset. Otolith width (mm) measurements averaged from left and right otoliths will be used as a proxy for herring size (Stevenson and Campana, 1992) in this analysis. We will also use salmon catch data recorded by each angler and submitted with the associated stomach sample.

##B) How does Chinook size, catch location and time of year in the Salish Sea affect the size of herring prey consumed?

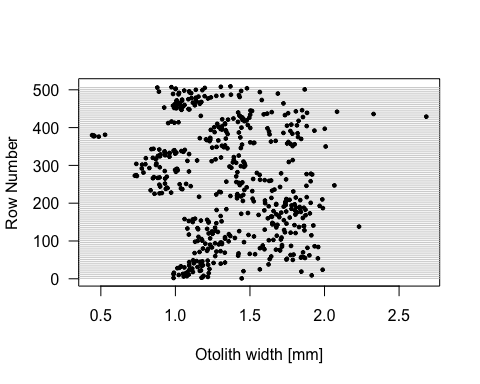
##C) This analysis is inferential with the goal of understanding factors that explain variation in herring size consumed by Chinook in the Salish Sea.

##D) We hypothesize that otolith width will be positively linear then asymptotic at maximum herring size as salmon length increases. Otolith width and day of year will show a quadratic, s-shaped relationship with the vertex midway through the calendar year. We expect a moderately positive, linear relationship with increasing latitude and a quadratic, concave up relationship with longitude with maximum herring size occurring in mid-channel waters. We expect that collinearity between latitude and longitude will cause in interaction effect on otolith width. An interaction may also exist between salmon length and latitude and between salmon length and day of year.

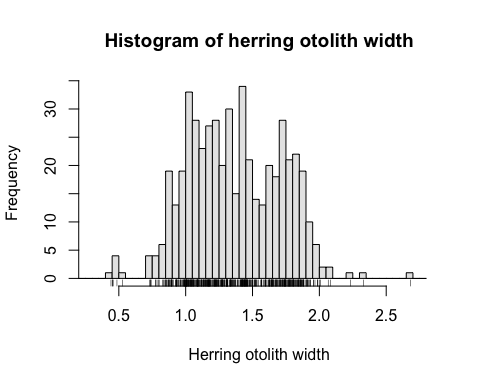
##E) We will use AIC fit with maximum likelihood estimation because it is flexible when comparing complex, non-nested models with parameters estimated from ecological field studies. AIC is commonly used in ecosystem studies where parameter estimates are relatively imprecise and we accept that the optimal model may be selected over the true, given the current sample size and our ecological understanding of the system.

#Section 2: The Response Variable (150 words) ##A)

#Add column for number of rows  
n <- nrow(herring\_data)  
herring\_data$row <- 1:n  
  
#Dotchart of otolith width (response variable)  
with(herring\_data, plot(oto.width, row, las = 1, type = "n", xlab = "Otolith width [mm]", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(oto.width, row, las = 1, type = "p", pch = 19, cex = 0.5))



#Histogram of response variable with rugmarks  
hist(herring\_data$oto.width, xlim = c(0.3, 2.8), breaks = seq(0.2, 2.8, by = 0.05), col = grey(0.9), main = "Histogram of herring otolith width", xlab = "Herring otolith width")  
rug(herring\_data$oto.width)



#Checking for zeroes and NAs in otilith width  
table(herring\_data$oto.width > 0, useNA = "always")

##   
## TRUE <NA>   
## 508 1

#exclude rows with NAs for continuous variables  
herring\_data <- herring\_data %>% na.omit()

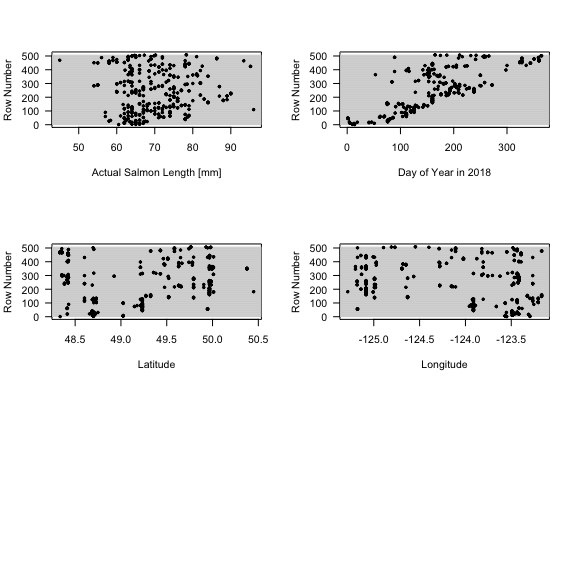
##B) The response variable (otolith width) is continuous and cannot take on zero or negative values. The range of possible response values in this dataset is 0.007 mm to 2.68 mm.

##C) Outliers at < 0.6 mm and > 2.2 mm are reasonable measurements representing a few individuals in the extreme age classes. There are no zero values and one missing otolith measurement making up 0.11 % of the data. We anticipate no difficulties in removing the single NA value from subsequent analyses. Non-independence is expected from repeated measurements of multiple herring that occurred within the same salmon stomach. Herring consumed at similar times of year may have similar growth trajectories that contribute to temporal autocorrelation. Spatial autocorrelation may occur if herring consumed in similar areas have similar prey fields and grow rates.

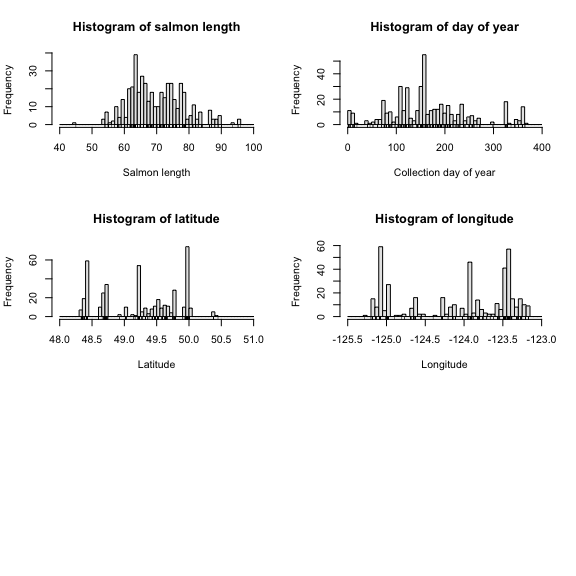
##D) We will consider the normal distribution for modelling.

#Section 3: The Explanatory Variables (150 words) ##A) ###Continuous explanatory variables

#Multipanel of dotcharts for continuous explanatory variables  
par(mfrow = c(3, 2))  
  
#Salmon length  
with(herring\_data, plot(sal.length, row, las = 1, type = "n", xlab = "Actual Salmon Length [mm]", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(sal.length, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#Collection day of year  
with(herring\_data, plot(coll.doy, row, las = 1, type = "n", xlab = "Day of Year in 2018", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(coll.doy, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#Latitude  
with(herring\_data, plot(lat, row, las = 1, type = "n", xlab = "Latitude", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(lat, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#Longitude  
with(herring\_data, plot(long, row, las = 1, type = "n", xlab = "Longitude", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(long, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#Multipanel of histograms for continuous explanatory variables  
par(mfrow = c(3, 2))



#Salmon length  
hist(herring\_data$sal.length, col = grey(0.9), breaks = seq(40, 100, by = 1), main = "Histogram of salmon length", xlab = "Salmon length")  
rug(herring\_data$sal.length)  
  
#Collection day of year  
hist(herring\_data$coll.doy, col = grey(0.9), breaks = seq(0, 400, by = 7), main = "Histogram of day of year", xlab = "Collection day of year")  
rug(herring\_data$coll.doy)  
  
#Latitude \*something is up with this y axis  
hist(herring\_data$lat, col = grey(0.9), breaks = seq(48.0, 51.0, by = 0.05), main = "Histogram of latitude", xlab = "Latitude")  
rug(herring\_data$lat)  
  
#Longitude  
hist(herring\_data$long, col = grey(0.9), breaks = seq(-125.5, -123, by = 0.05), main = "Histogram of longitude", xlab = "Longitude")  
rug(herring\_data$long)



###Categorical explanatory variables

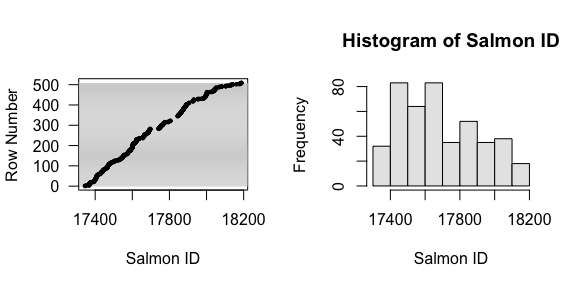
#Multipanel of dotcharts for categorical explanatory variables  
par(mfrow = c(1, 2))  
  
#Salmon id \*is this useful? Shows that ID increases as we would expect.  
with(herring\_data, plot(sal.id, row, las = 1, type = "n", xlab = "Salmon ID", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(sal.id, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#Make a histogram of herring frequency per salmon \*bins are off unsure why  
hist(herring\_data$sal.id, col = grey(0.9), bin = seq(17300, 18200, by = 1), main = "Histogram of Salmon ID", xlab = "Salmon ID")

## Warning in plot.window(xlim, ylim, "", ...): "bin" is not a graphical  
## parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):  
## "bin" is not a graphical parameter

## Warning in axis(1, ...): "bin" is not a graphical parameter

## Warning in axis(2, ...): "bin" is not a graphical parameter



#Make a table of salmon sex, digestion, collection month  
with(herring\_data, table(sal.sex, digestion)) #is this useful? Just shows digestion of herring vs. sex of salmon - perhaps omit. How important is it to reference digestion at all if we are not including it in our model?

## digestion  
## sal.sex 0 1 2 3 4  
## 0 3 13 33 34  
## female 1 5 37 90 85  
## male 0 2 19 73 45

with(herring\_data, table(digestion, coll.month)) #useful to show herring per month

## coll.month  
## digestion 1 2 3 4 5 6 7 8 9 10 11 12  
## 0 1 0 0 0 0 0 0 0 0 0 0 0  
## 1 1 0 2 0 2 2 0 1 0 0 2 0  
## 2 2 1 11 10 4 21 8 5 1 0 5 1  
## 3 10 2 13 25 16 53 31 11 10 1 8 16  
## 4 7 3 18 23 27 31 19 16 10 1 3 6

with(herring\_data, table(sal.sex, coll.month)) #useful to show salmon per month

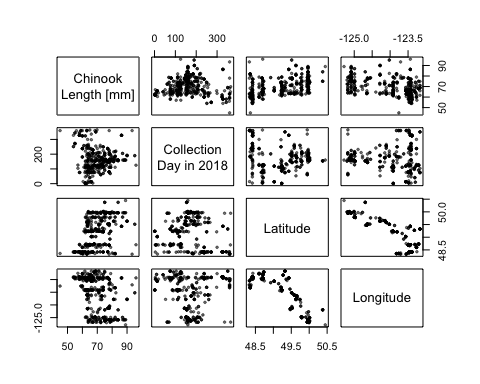
## coll.month  
## sal.sex 1 2 3 4 5 6 7 8 9 10 11 12  
## 6 1 9 11 15 13 4 8 5 1 0 10  
## female 14 2 20 34 28 59 32 12 3 1 9 4  
## male 1 3 15 13 6 35 22 13 13 0 9 9

## B)

One outlier is observed at 45cm and three outliers at 90 - 100cm. However, these are reasonable sizes for recreationally caught chinook salmon. No transformations are necessary as no obvious skewedness is observed.

#Section 4: Collinearity, Balance, and Variance Inﬂation Factors (200 words) ##A)

#create data frame for pairwise scatter with only continuous explanatory variables - salmon lenght, weight, collection day of year, latitude and longitude  
herring\_collinearity <- data.frame("sal.length" = herring\_data$sal.length, "coll.doy" = herring\_data$coll.doy, "lat" = herring\_data$lat, "long" = herring\_data$long)  
  
#Pairwise scatterplot of explanatory variables (salmon length, weight, day of year, latitude and longitude)  
plot(herring\_collinearity[1:4], cex = 0.5, pch = 19, col = rgb(0, 0, 0, 0.5),   
 labels = c("Chinook\nLength [mm]", "Collection\nDay in 2018", "Latitude", "Longitude"))



#Pairwise correlation coefficients for explanatory variables  
print(cor(na.omit(herring\_data[ ,c("sal.length", "coll.doy", "lat", "long")])), digits = 2)

## sal.length coll.doy lat long  
## sal.length 1.000 -0.0619 0.45 -0.3991  
## coll.doy -0.062 1.0000 -0.13 -0.0033  
## lat 0.449 -0.1284 1.00 -0.8721  
## long -0.399 -0.0033 -0.87 1.0000

#Variance Inflation Factors  
print(vif(na.omit(herring\_data[ ,c("sal.length", "coll.doy", "lat", "long")])))

## sal.length coll.doy lat long   
## 1.253474 1.077615 4.721313 4.428887

Tables in section 3 indicate more herring and salmon were collected between March and July (316) compared to other months (124). More females (218) were caught compared to males (139) and 83 fish were left unsexed. We should be cautious about drawing conclusions from models that include sex or collection month as any estimates may be driven by unbalanced variables.

Pairwise scatterplots suggest a concave up relationship between salmon length and collection day of year with a weak correlation coefficient of -0.062, perhaps due to seasonality of salmon caught during the spring and summer months. Salmon length is moderately, positively correlated (0.449) with latitude but moderately, negatively correlated longitude (-0.3991). There is a weak, slightly negative correlation (-0.1284) between collection day of year and latitude but a strong, negative correlation (-0.8721) between latitude and longitude. VIF values of latitude (4.721313) and longitude (4.428887) are greater than 3 and strongly colinear, which is expected due to the geographical constraints of the Strait of Georgia and Juan de Fuca. We will run models with either latitude or longitude to avoid the confounding effects of including both variables in a single model.

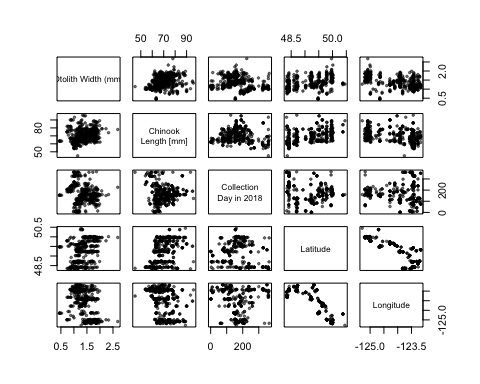
*address doy issue in discussion* maybe include totals in tables

#Section 5: Statistical Methods and Model Fitting (200 words)

We will be using linear mixed-effects models using the R Package lme4 (Bates et al., 2015). We are considering the normal distribution of otolith width (mm) modelled as function of salmon length (mm), day of year in 2018, latitude and longitude. We will include a random effect of multiple herring that occurred within individual salmon because similar sized fish are likely to travel in schools targeted by an individual and there are > 10 levels associated with this variable. We are excluding salmon sex and collection month due to unbalancing effects and excluding digestion level, which is the degree of herring degradation on a scale of 1 to 4, as it is not relevant to our research question. In our first round of modelling fitting all salmon lengths will be included followed by a second round of model fitting to salmon 62 cm and above as this is the regulation size restrictions for recreationally caught salmon in the Strait of Georgia. This will check the influence of removing under size regulation fish from the dataset.

## Exploratory plotting

#create data frame for pairwise scatter with all variables  
herring\_collinearity <- data.frame("oto.width" = herring\_data$oto.width, "sal.length" = herring\_data$sal.length, "coll.doy" = herring\_data$coll.doy, "lat" = herring\_data$lat, "long" = herring\_data$long)  
  
plot(herring\_collinearity[1:5], cex = 0.5, pch = 19, col = rgb(0, 0, 0, 0.5),   
 labels = c("Otolith Width (mm)", "Chinook\nLength [mm]", "Collection\nDay in 2018", "Latitude", "Longitude"))



## Model Structure

oto.width ~ coll.doy + lat + sal.length + (1 | sal.id) oto.width ~ coll.doy + long + sal.length + (1 | sal.id)

oto.width ~ coll.doy + lat + coll.doy*lat + (1 | sal.id) oto.width ~ coll.doy + sal.length + coll.doy*sal.length + (1 | sal.id) oto.width ~ lat + sal.length + lat*sal.length + (1 | sal.id) oto.width ~ coll.doy + long + coll.doy*long + (1 | sal.id) oto.width ~ long + sal.length + long\*sal.length + (1 | sal.id)

## JQ modelling

#scale data into new data frame  
  
model1 <- lmer(oto.width ~ coll.doy + lat + coll.doy\*lat + (1 | sal.id), data = herring\_data, REML = FALSE)

## Warning: Some predictor variables are on very different scales: consider  
## rescaling

model1

## Linear mixed model fit by maximum likelihood ['lmerMod']  
## Formula: oto.width ~ coll.doy + lat + coll.doy \* lat + (1 | sal.id)  
## Data: herring\_data  
## AIC BIC logLik deviance df.resid   
## 2.0741 26.5947 4.9630 -9.9259 434   
## Random effects:  
## Groups Name Std.Dev.  
## sal.id (Intercept) 0.2455   
## Residual 0.1545   
## Number of obs: 440, groups: sal.id, 239  
## Fixed Effects:  
## (Intercept) coll.doy lat coll.doy:lat   
## -1.165e+01 1.044e-02 2.676e-01 -2.264e-04   
## fit warnings:  
## Some predictor variables are on very different scales: consider rescaling

##Untransformed models

#herring\_data <- herring\_data %>% filter(coll.month == "3" | coll.month == "4" | coll.month == "5" | coll.month == "6" | coll.month == "7")  
  
#Fit model 1  
#model\_1 <- lmer(oto.width ~ coll.month + long + sal.length + digestion + (1|sal.id), data = herring\_data)

#Section 6: Model Checking ##All models summarized (delete before submitting?)

#Summary of model 1  
#summary\_1 <- summary(model\_1)$coef

##All models plotted against data (delete before submitting?)

#Plot model 1 residuals  
#plot(model\_1, which = 1)

##Chosen models

#Check the assumptions of at least one model  
#- plotting and interpreting the residuals of the model  
#- check for correlation in the residuals due to grouping variables

Summary of the plausibility of the model - Potential outliers

#Section 7: Model Summary, Conﬁdence Intervals and Model Comparison

#Model comparison and/or model selection and/or hypothesis testing  
#- AIC (or AICc) values  
#- AIC (or AICc) weights  
#- R-squared (or adjusted R-squared) values  
#- p-values.   
#Summary for at least one ﬁtted model  
#- coeﬃcient estimates and their standard errors  
#- conﬁdence intervals for coeﬃcient estimate

#Section 8: Plotting a Model with the Data

#Plot at least one model with data  
#- use method that shows strengths and weaknesses  
#- include measurement of uncertainty

#Section 9: Discussion (400 words) ##A) Assessment of how well the model seems to ﬁt the data - model checking - visualization of the model with the data - ##B) Qualitative interpretation of the model - interpretation of results in relation to your research questions and hypotheses - general interpretation of conﬁdence intervals or standard errors for coeﬃcient estimates - ##C) Limitations of the model(s) that you have ﬁt - Limitations - Possible options for improvement - Possible approaches for further data analysis

##D) Results into context - citie at least two relevant scientiﬁc papers

## References

Schindler, D. E., M. D. Scheuerell, J. W. Moore, S. M. Gende, T. B. Francis, and W. J. Palen. 2003. Pacific Salmon and the Ecology of Coastal Ecosystems. Frontiers in Ecology and the Environment 1:31.

Stevenson, D. K., and S. E. Campana. 1992. Otolith microstructure examination and analysis. Can. Spec. Publ. Fish. Aquat. Sci. 117: 126 p. 1. Page Otolith Microstructure Examination and Analysis.

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67 (1), 1-48.