Data Analysis Project

Jess Qualley and Rebecca Hansen

April 06, 2020

## 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,  
 coll.month = herring$CollectionMonth,  
   
 sal.sp = herring$SalmonSpecies,  
 coll.year = herring$CollectionYear,  
 stat.area = as.factor(herring$StatArea))  
  
#Filter for 2018 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

## 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Despite their important role in the coastal marine ecosystem (Schindler et al. 2003), diet studies from this region are outdated and limited to summer sampling.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.

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 found 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 use salmon catch data recorded by each angler and linked to each 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 descriptive, with the goal of showing how explanatory variables describe patterns in herring size consumed by Chinook.

## D)

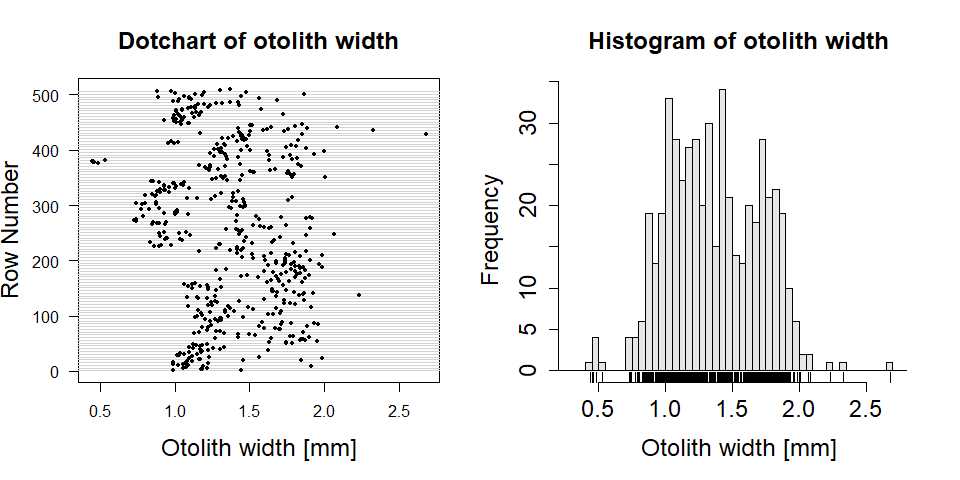
We hypothesize that otolith width will be positively linear with salmon length and collection day of year. We expect a moderately positive, linear relationship with increasing latitude and negative linear relationship with longitude if larger herring occur along western Strait of Georgia. A non-linear relationship with longitude is possible if maximum herring size occurrs in mid-channel waters. We expect an interaction between salmon length and latitude, because of higher productivity in northern regions. An interaction between salmon length and longitude is possible if mainland and Vancouver Island stocks differ in size. We expect an interaction between salmon length and day of year as each age class grows.

## 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 model, given the current sample size and our ecological understanding of the system. (396)

# Section 2: The Response Variable (150 words)

## A)



#Check 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)

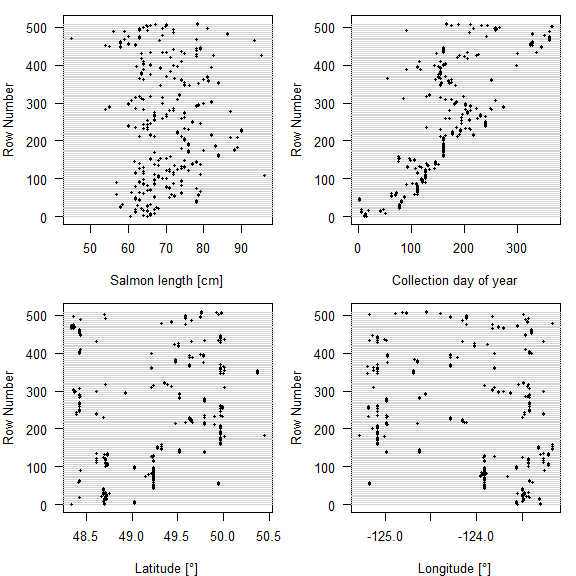
Outliers at < 0.6 mm and > 2.2 mm are reasonable measurements representing a few individuals in the extreme size classes. There are no zero values and one missing otolith measurement making up 0.11 % of the data. We anticipate no difficulties in removing the single NA value from subsequent analyses. 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.

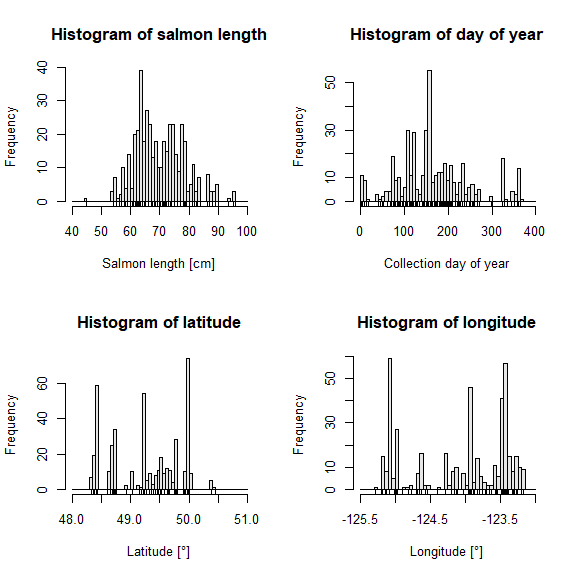
## D)

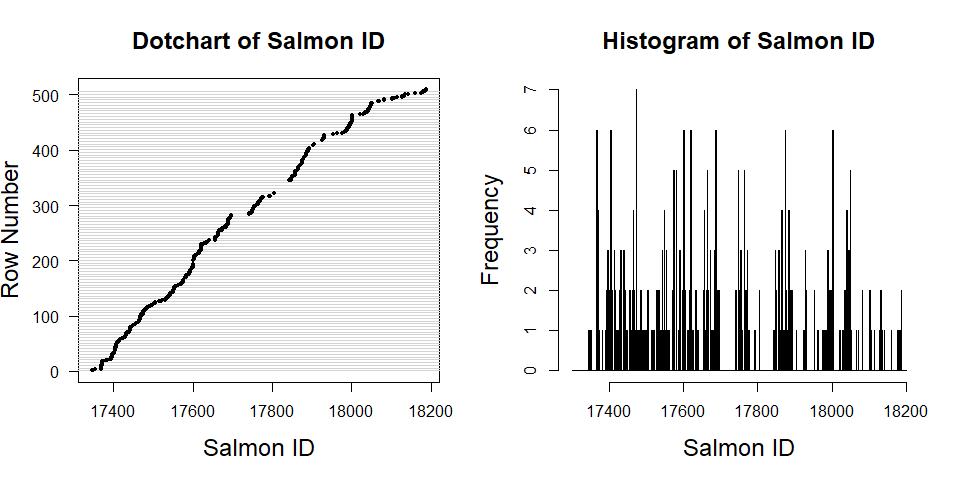
We will use the normal distribution for modelling. (140)

# Section 3: The Explanatory Variables

## A)





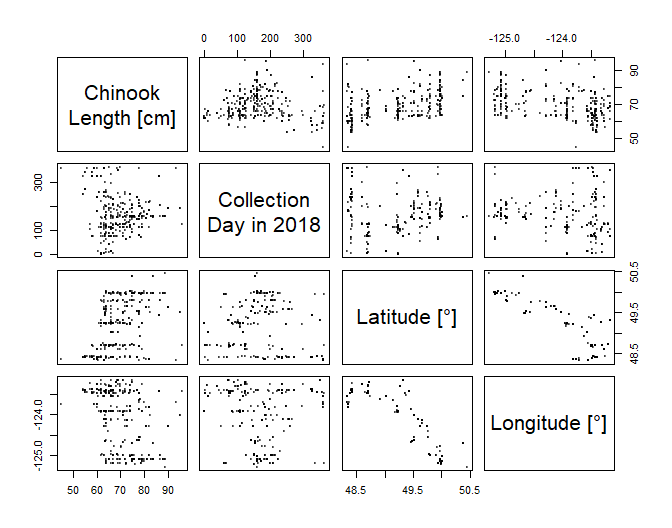


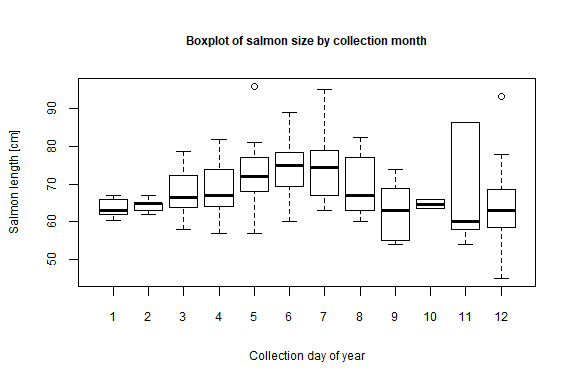
## B)

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 substantial skewedness is observed. (32)

# Section 4: Collinearity, Balance, and Variance Inﬂation Factors

## A)





#Calculate and view 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

#Calulate and view 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

The only categorical variable is salmon id and the range in the number of herring per salmon was 1 to 7, which will not confound any other categorical variables. We will model salmon ID as a random effect to account for unbalancing effects associated with salmon ID on the continuous variables.

The boxplot of mean salmon length per month shows greater variance and greater summer sample sizes where lengths overlap with those observed in earlier months of the year. This unbalanced sampling effort may be contributing to the concave up shape observed in the pairwise scatterplot. The remaining relationships observed in the pairwise scatterplots were consistent with the pairwise correlation coefficients. Salmon length was moderately, positively correlated (0.449) with latitude but moderately, negatively correlated longitude (-0.3991). There was a strong, negative correlation (-0.8721) between latitude and longitude. VIF values of latitude (4.721313) and longitude (4.428887) were 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. (200)

# Section 5: Statistical Methods and Model Fitting

We will be using linear mixed-effects models using the R Package lme4 (Bates et al. 2015). We are assuming that otolith width (mm), our response variable, is normally distributed as suggested by a histogram of otolith width. We will model otolith width (mm) as function of salmon length (cm), day of year in 2018, latitude and longitude. Dotplots and histograms of these predictor variables show few missing values or outliers so the data from these variables can be used for modelling. There wasn’t strong colinearity between predictor variables except for longitude and latitude, which we will not include in the same model. We have chosen to standardize our predictor variables as they were measured on vastly different scales. We will include a random intercept corresponding to individual salmon (salmon ID) to avoid autocorrelation from non-independence of multiple herring that may occur in a single salmon or experimental unit. Ecologically, it is likely that multiple herring in one stomach have similar sizes if they travel in the same school targeted by an individual salmon. We have chosen a random intercept for salmon ID because the number of levels is greater than 10 (Zuur et al. 2007). (196)

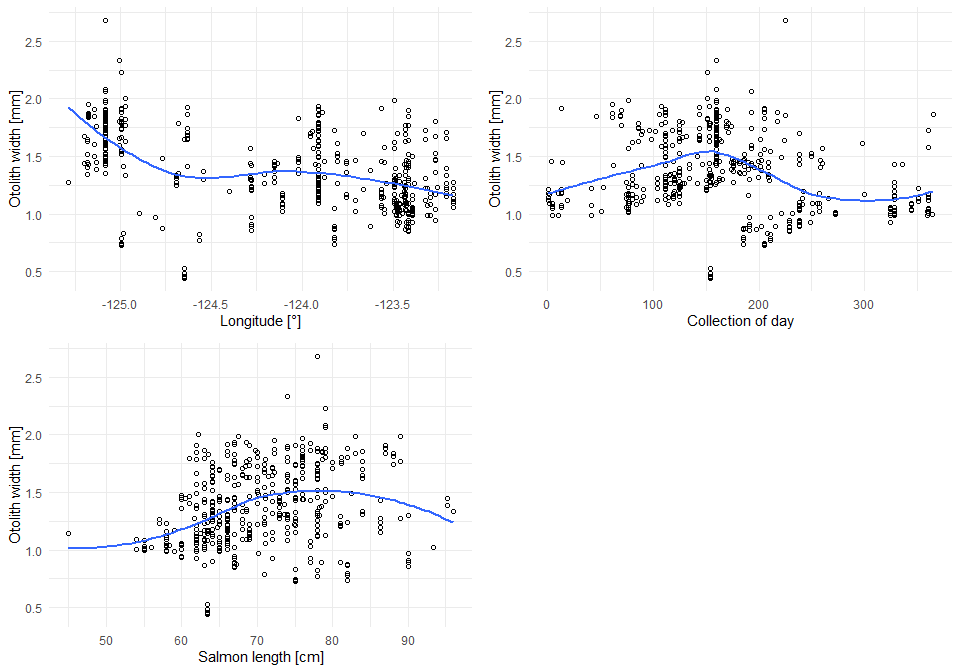
#Add columns for standardized continuous explanatory variables ((x - mean)/ 1 sd)  
herring\_data$sal.length\_standardized <- (herring\_data$sal.length - mean(herring\_data$sal.length)) / sd(herring\_data$sal.length)  
herring\_data$coll.doy\_standardized <- (herring\_data$coll.doy - mean(herring\_data$coll.doy)) / sd(herring\_data$coll.doy)  
herring\_data$lat\_standardized <- (herring\_data$lat - mean(herring\_data$lat)) / sd(herring\_data$lat)  
herring\_data$long\_standardized <- (herring\_data$long - mean(herring\_data$long)) / sd(herring\_data$long)

#Intercept-only model  
model\_0 <- lmer(oto.width ~ (1 | sal.id), data = herring\_data)  
  
#Models with one explanatory variable  
model\_1 <- lmer(oto.width ~ sal.length\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_2 <- lmer(oto.width ~ coll.doy\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_3 <- lmer(oto.width ~ lat\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_4 <- lmer(oto.width ~ long\_standardized + (1 | sal.id), data = herring\_data)  
  
#Models with two explanatory variables  
model\_5 <- lmer(oto.width ~ coll.doy\_standardized + sal.length\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_6 <- lmer(oto.width ~ coll.doy\_standardized + long\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_7 <- lmer(oto.width ~ coll.doy\_standardized + lat\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_8 <- lmer(oto.width ~ sal.length\_standardized + long\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_9 <- lmer(oto.width ~ sal.length\_standardized + lat\_standardized + (1 | sal.id), data = herring\_data)  
  
#Models with three explanatory variables  
model\_10 <- lmer(oto.width ~ coll.doy\_standardized + lat\_standardized + sal.length\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_11 <- lmer(oto.width ~ coll.doy\_standardized + long\_standardized + sal.length\_standardized + (1 | sal.id), data = herring\_data)  
  
#Models with two explanatory variables and one interaction term  
model\_5i <- lmer(oto.width ~ coll.doy\_standardized\*sal.length\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_6i <- lmer(oto.width ~ coll.doy\_standardized\*long\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_7i <- lmer(oto.width ~ coll.doy\_standardized\*lat\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_8i <- lmer(oto.width ~ sal.length\_standardized\*long\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_9i <- lmer(oto.width ~ sal.length\_standardized\*lat\_standardized + (1 | sal.id), data = herring\_data)  
  
#Models with three explanatory variables and one interaction term  
model\_10l <- lmer(oto.width ~ coll.doy\_standardized + lat\_standardized\*sal.length\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_11l <- lmer(oto.width ~ coll.doy\_standardized + long\_standardized\*sal.length\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_10c <- lmer(oto.width ~ coll.doy\_standardized\*sal.length\_standardized + lat\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_11c <- lmer(oto.width ~ coll.doy\_standardized\*sal.length\_standardized + long\_standardized + (1 | sal.id), data = herring\_data)  
  
#Models with three explanatory variables and two interaction terms  
model\_10cl <- lmer(oto.width ~ coll.doy\_standardized\*sal.length\_standardized\*lat\_standardized + (1 | sal.id), data = herring\_data)  
  
model\_11cl <- lmer(oto.width ~ coll.doy\_standardized\*sal.length\_standardized\*long\_standardized + (1 | sal.id), data = herring\_data)

# Section 6: Model Checking

## Plotting otolith.width vs. models’ explanatory variables

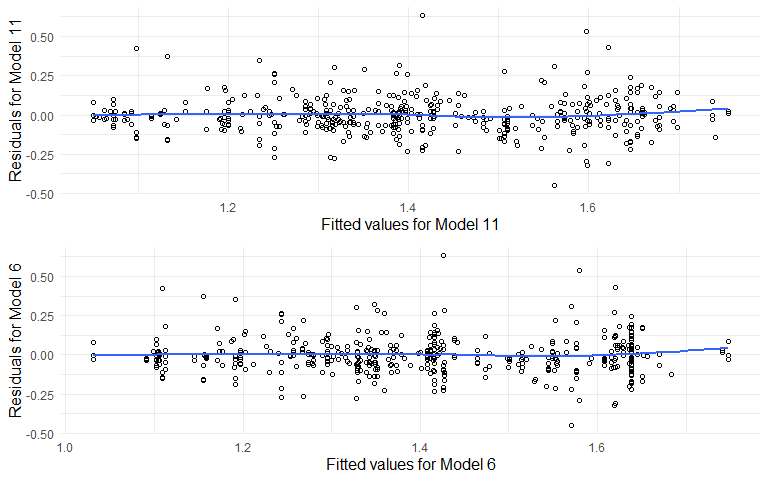
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'  
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'  
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'

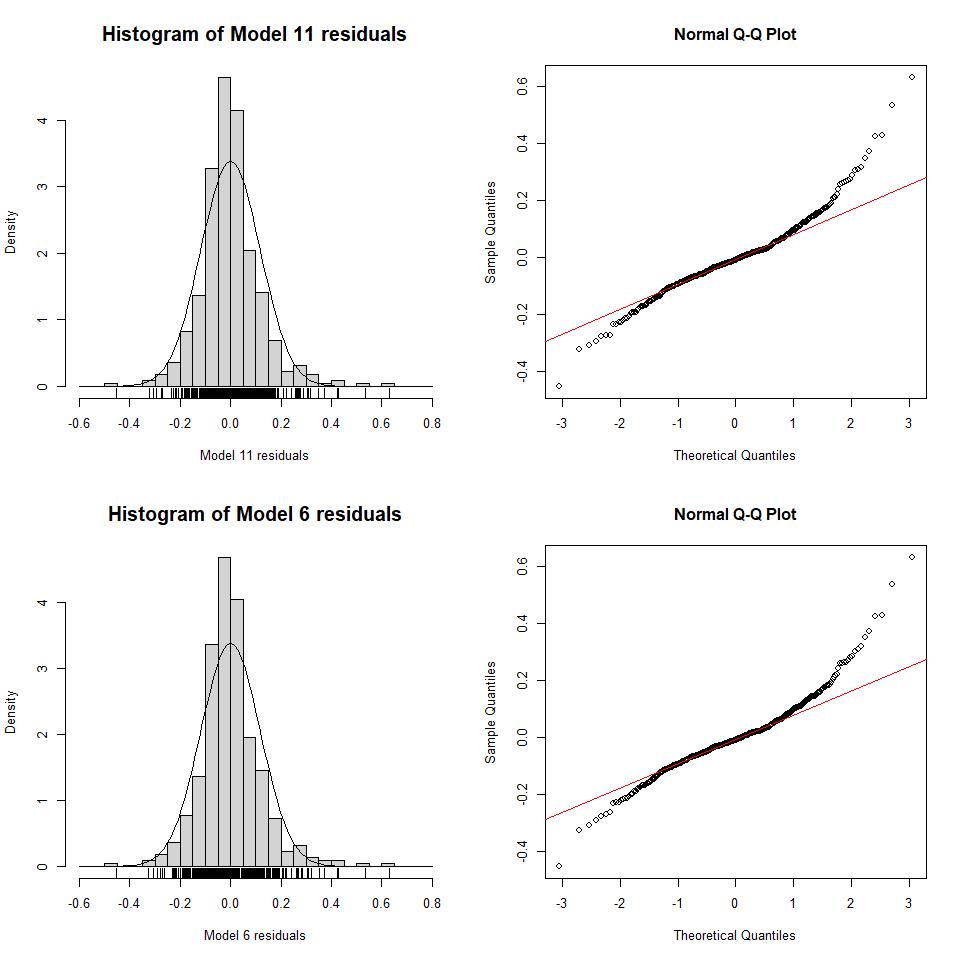


## Plotting residuals for Model 11 and Model 6

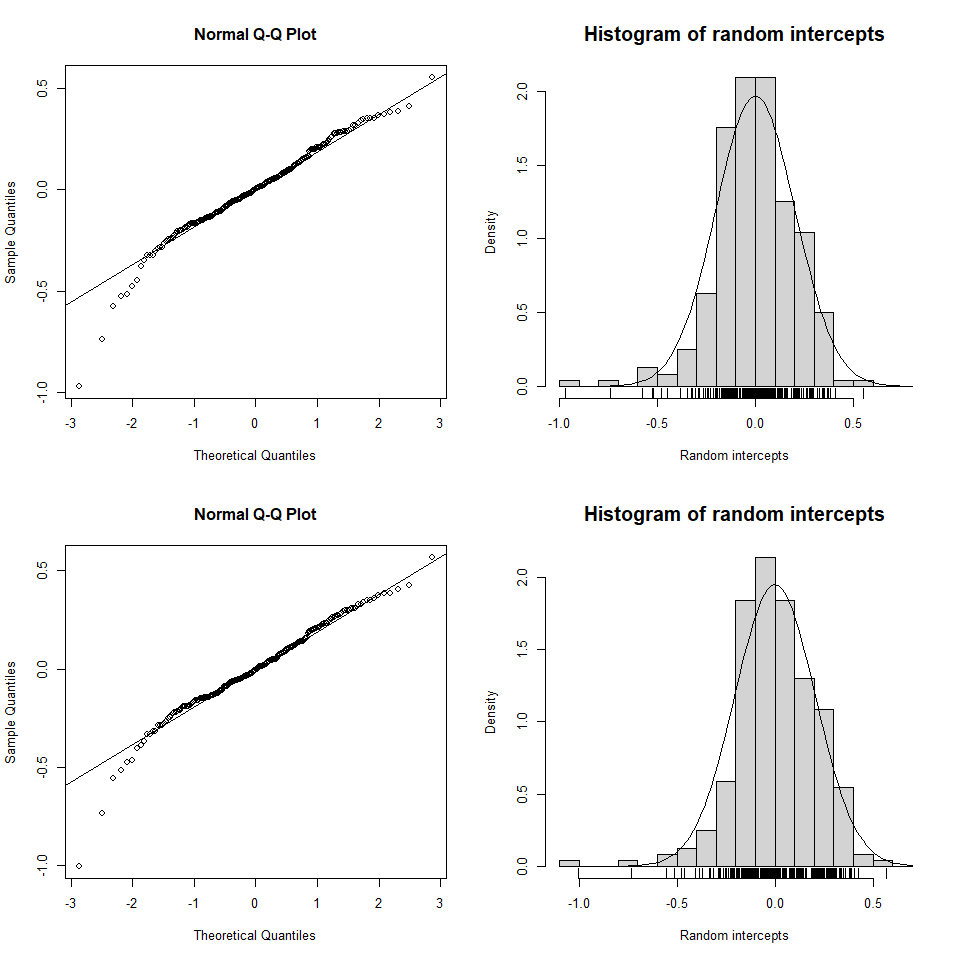
#Add Model 11 residuals and fitted values to data frame  
herring\_data$resid\_11 <- resid(model\_11)   
herring\_data$fitted\_11 <- predict(model\_11, re.form = ~0)  
  
#Add Model 6 residuals and fitted values to data frame  
herring\_data$resid\_6 <- resid(model\_6)   
herring\_data$fitted\_6 <- predict(model\_6, re.form = ~0)

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'  
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



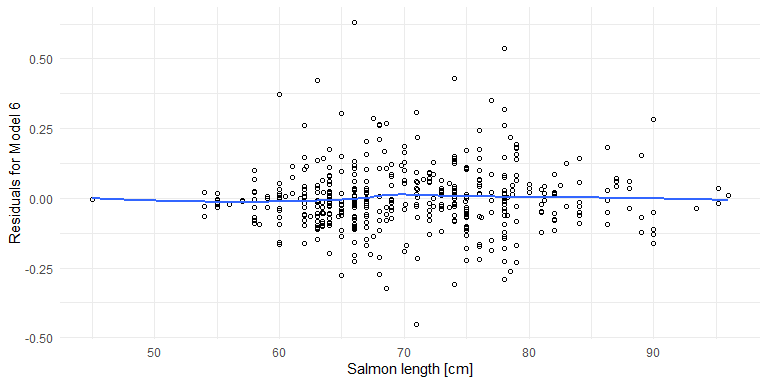


## Plotting random intercepts for Model 11 and Model 6

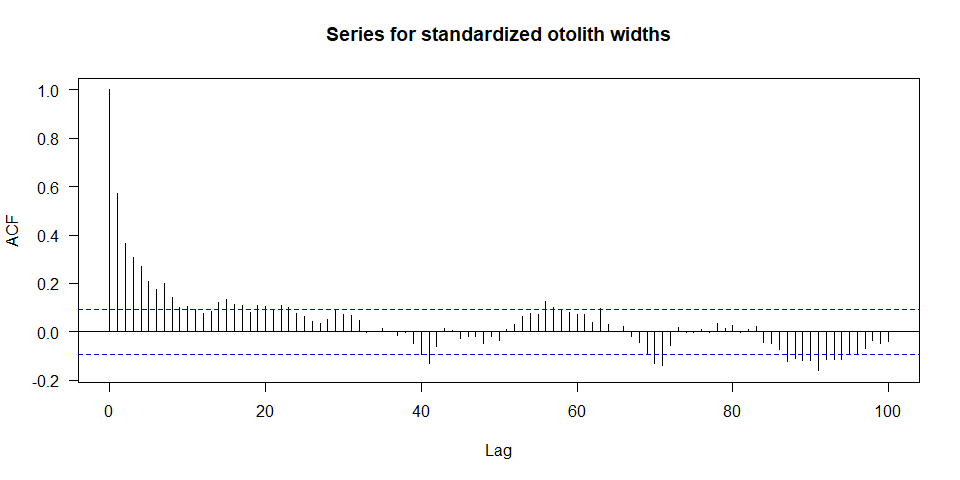


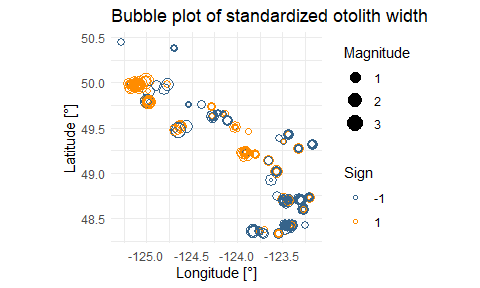
## Plottng Model 6 residuals vs. salmon length

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



#Create column of standardized otolith widths ((x - mean)/ 1 sd)  
herring\_data$oto.width\_standardized <- (herring\_data$oto.width - mean(herring\_data$oto.width)) / sd(herring\_data$oto.width)





Pairwise scatterplots suggest a nonlinear relationship between the response and predictor variables. ACF plot suggests temporal autocorrelation at short time lags ~ 10 days or less. Bubble plots indicate spatial autocorrelation and/or may reflect the spatial size distribution of herring. Model 11 histogram of residuals appear normal compared to the quantile-quantile plot suggesting a non-normal distribution. Residuals fall above the 1:1 line at high quantiles and below the 1:1 line at negative quantiles, suggesting slightly right-skewed residuals. The increase in spread of the Model 11 residuals versus fitted values suggests unequal variance. Residuals for the random intercept appear to fall below the 1:1 line at negative quantiles, suggesting a non-normal distribution of residuals. Model 6 of residuals and random intercept distribution show the same pattern as Model 11, violating assumptions of the linear model. The increase in variance and non-linear relationships indicate that the linear model is not plausible. Outliers may exist above 0.5 and below -40; however, they may just represent the upper and lower range of the data given the pattern of increasing variance. (181)

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

#Rewrote all models from Section 5 with REML = FALSE

#Calculate AIC values  
AIC\_out <- AIC(model\_0, model\_1, model\_2, model\_3, model\_4, model\_5, model\_6, model\_7, model\_8, model\_9, model\_5i, model\_6i, model\_7i, model\_8i, model\_9i, model\_10, model\_11, model\_10l, model\_10c, model\_11l, model\_11c, model\_10cl, model\_11cl)  
  
#Calculate delta AIC values  
AIC\_out$delta\_AIC <- AIC\_out$AIC - min(AIC\_out$AIC)  
  
#Show AIC table with weights  
AICtab(model\_0, model\_1, model\_2, model\_3, model\_4, model\_5, model\_6, model\_7, model\_8, model\_9, model\_5i, model\_6i, model\_7i, model\_8i, model\_9i, model\_10, model\_11, model\_10l, model\_10c, model\_11l, model\_11c, model\_10cl, model\_11cl, base = TRUE, weights = TRUE)

## AIC dAIC df weight  
## model\_11 -21.3 0.0 6 0.328   
## model\_6 -20.3 1.0 5 0.198   
## model\_11c -20.1 1.2 7 0.179   
## model\_11l -19.7 1.6 7 0.146   
## model\_6i -19.5 1.8 6 0.133   
## model\_11cl -15.3 6.1 10 0.016   
## model\_8 -7.3 14.0 5 <0.001  
## model\_8i -5.5 15.8 6 <0.001  
## model\_4 -5.3 16.0 4 <0.001  
## model\_10 -1.6 19.7 6 <0.001  
## model\_10c -0.8 20.6 7 <0.001  
## model\_10l -0.2 21.1 7 <0.001  
## model\_7 0.3 21.7 5 <0.001  
## model\_7i 2.1 23.4 6 <0.001  
## model\_9 3.9 25.3 5 <0.001  
## model\_10cl 5.0 26.3 10 <0.001  
## model\_9i 5.8 27.1 6 <0.001  
## model\_3 5.9 27.2 4 <0.001  
## model\_5 30.6 51.9 5 <0.001  
## model\_5i 31.4 52.7 6 <0.001  
## model\_1 37.2 58.5 4 <0.001  
## model\_2 47.4 68.7 4 <0.001  
## model\_0 54.5 75.8 3 <0.001

#Call coeﬃcient estimates for fixed effects of Model 11  
fixef(model\_11)

## (Intercept) coll.doy\_standardized long\_standardized   
## 1.40823938 -0.07112385 -0.14511708   
## sal.length\_standardized   
## 0.03339340

#Call variation within and between the random effect (salmon ID) of Model 11  
VarCorr(model\_11)

## Groups Name Std.Dev.  
## sal.id (Intercept) 0.23094   
## Residual 0.15423

#Calculate conﬁdence intervals for coeﬃcient estimates of Model 11  
confint(model\_11)

## Computing profile confidence intervals ...

## 2.5 % 97.5 %  
## .sig01 0.205272227 0.25963048  
## .sigma 0.140630587 0.17004989  
## (Intercept) 1.374474504 1.44215054  
## coll.doy\_standardized -0.105569467 -0.03672275  
## long\_standardized -0.181799245 -0.10839302  
## sal.length\_standardized -0.004343185 0.07109330

#Call coeﬃcient estimates for fixed effects of Model 6  
fixef(model\_6)

## (Intercept) coll.doy\_standardized long\_standardized   
## 1.40569338 -0.07358628 -0.15719183

#Call variation within and between the random effect (salmon ID) of Model 6  
VarCorr(model\_6)

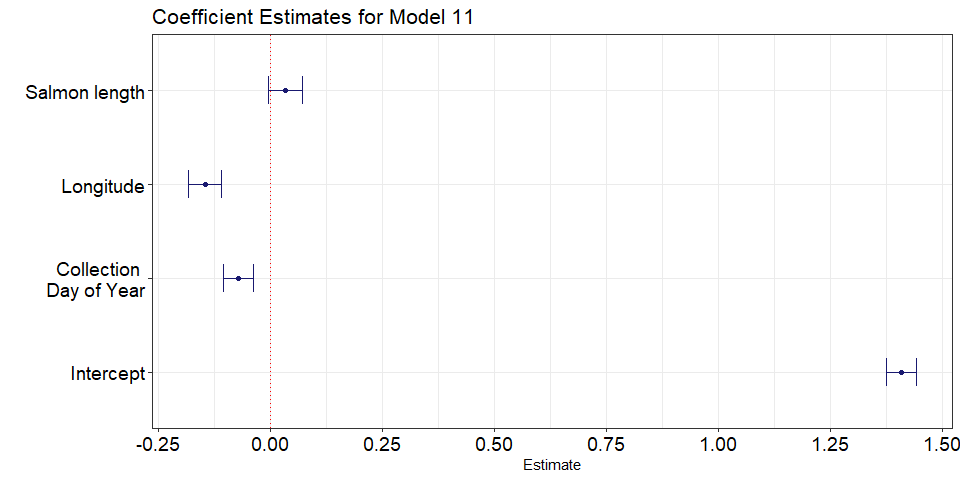
## Groups Name Std.Dev.  
## sal.id (Intercept) 0.23295   
## Residual 0.15418

#Calculate conﬁdence intervals for coeﬃcient estimates of Model 6  
confint(model\_6)

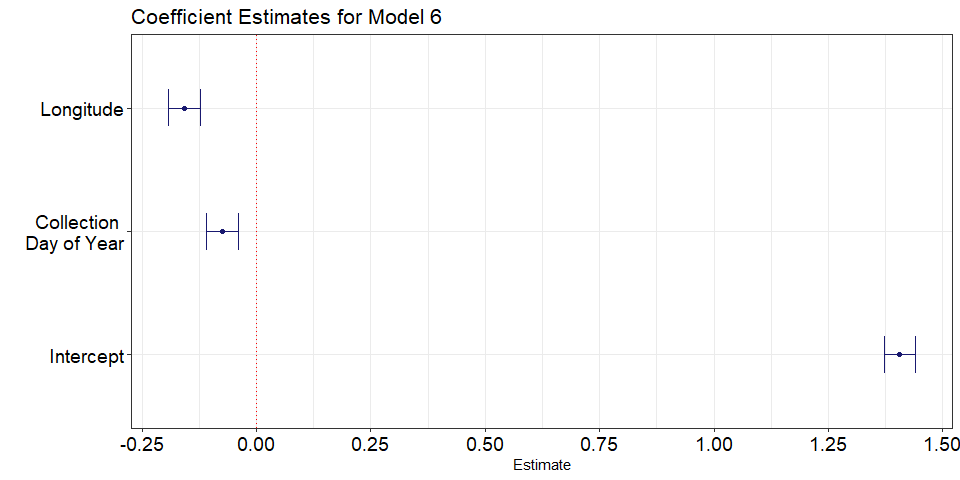
## Computing profile confidence intervals ...

## 2.5 % 97.5 %  
## .sig01 0.2071951 0.26175136  
## .sigma 0.1405957 0.17000163  
## (Intercept) 1.3718305 1.43968577  
## coll.doy\_standardized -0.1081468 -0.03906564  
## long\_standardized -0.1914520 -0.12286438

## Computing profile confidence intervals ...

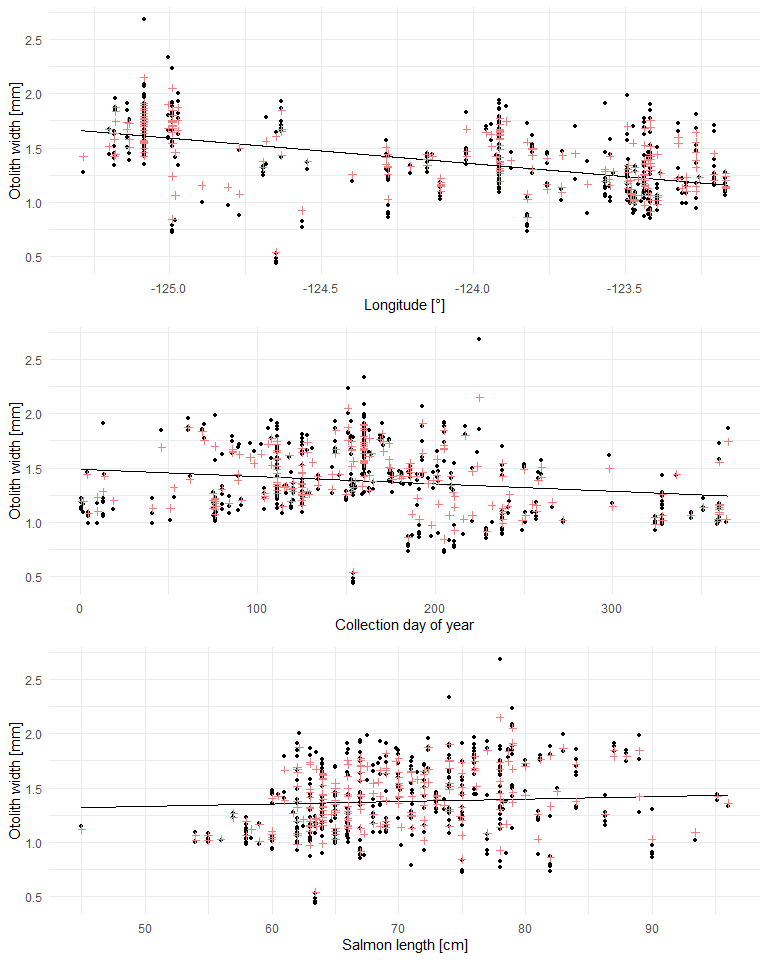


## Computing profile confidence intervals ...

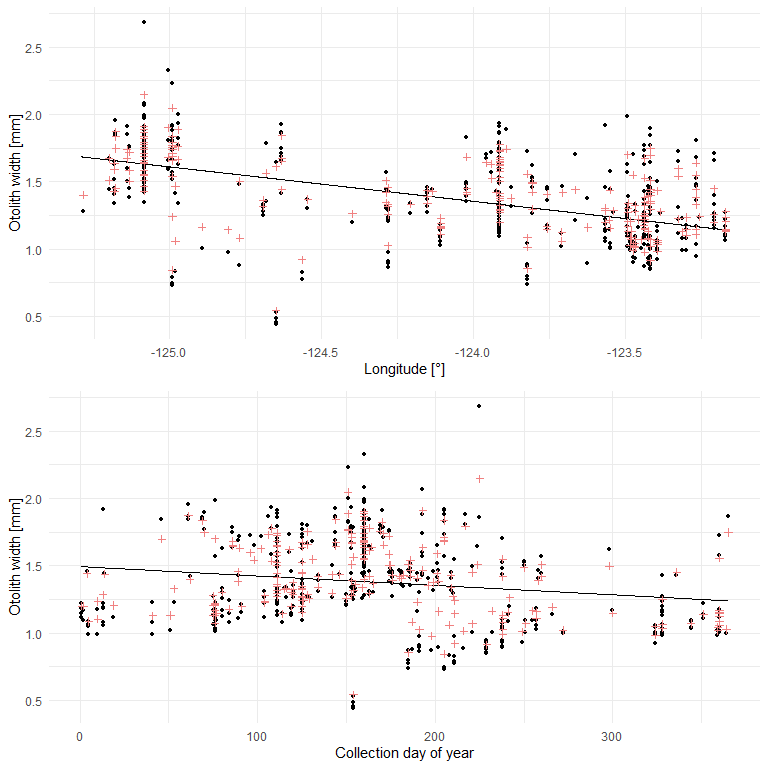


# Section 8: Plotting a Model with the Data

## Visualizing Model 11



## Visualizing Model 6



# Section 9: Discussion

## A)

Model 11 and 6 does not fit the data well and violates linear model assumptions. Both models overpredict otolith width between longitude -124.0 to -125.0. Both models overpredict between 0 to 50 days and underpredict 125 to 175 days. Model 11 overpredicts for salmon between 40 cm to 60 cm.

## B)

Model 11 coefficient plot shows that salmon length has a small positive relationship with otolith width, but the confidence interval overlaps with 0 and is not a useful predictor unlike longitude and day of year that doesn’t overlap with 0 and has a small, negative relationship with otolith width. Similar patterns are observed for Model 6. The results support our hypothesis for a positive relationship between otolith width versus salmon length and longitude but contradicts positive relationship we hypothesized for collection day. Interestingly, longitude better describes otolith width than latitude so we determine that longitude and collection day of year best describe the variation in herring size consumed by Chinook Salmon in the Salish Sea in 2018.

## C)

Longitude is not capturing east-west patterns from shallow to deep water, but the spatial size distribution of herring in westerly regions near Comox Valley and easterly regions near San Juan Islands. Modelling capture location as a categorical variable will allow us to look at regional patterns. Otolith width versus collection day may represent two or more age classes of available herring. Modelling age classes separately may suggest the importance of a different set of factors. Alternatively, we could model herring size as a quadratic relationship. Otolith width versus salmon length shows few salmon below 62cm paired with model overprediction, due to gape limitation or size restrictions >62cm for recreational catch. Generating models with and without fish <62cm can determine whether omitting this subset is justified and changes our result.

## D)

The longitudinal and day of year pattern may reflect high herring abundance along east coast Vancouver Island and peak spawning near Hornby/Denman Island in mid-March (Therriault et al. 2009). Perhaps a greater contribution of large herring consumed in western regions and small larvae appearing in June are driving these trends. Although salmon length is a poor predictor, the overall trend suggests prey size range increases with salmon size, consistent with findings in other marine predators (Scharf et al. 2000). (385)

## References

Baptiste, A. (2017). gridExtra: Miscellaneous Functions for “Grid” Graphics. R package version 2.3. <https://CRAN.R-project.org/package=gridExtra>

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

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. <doi:10.18637/jss.v067.i01>.

Bolker, B. and R Development Core Team (2020). bbmle: Tools for General Maximum Likelihood Estimation. R package version 1.0.23.1. <https://CRAN.R-project.org/package=bbmle>

Faraway, J. (2016). faraway: Functions and Datasets for Books by Julian Faraway. R package version 1.0.7. <https://CRAN.R-project.org/package=faraway>

Scharf, F. S., F. Juanes, and R. A. Rountree. 2000. Predator size - Prey size relationships of marine fish predators: Interspecific variation and effects of ontogeny and body size on trophic-niche breadth. Marine Ecology Progress Series 208:229–248.

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.

Therriault, T. W., D. E. Hay, and J. F. Schweigert. 2009. Biological overview and trends in pelagic forage fish abundance in the Salish Sea (Strait of Georgia, British Columbia). Marine Ornithology 37:3–8.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL <http://www.jstatsoft.org/v40/i01/>.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, <https://doi.org/10.21105/joss.01686>

Zuur, A., et al., 2007. Analysing Ecological Data. Springer: New York.