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 gridextra package  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':  
##   
## combine

#Load lmer package  
library(lme4)

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

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

#Load bbmle package  
library(bbmle)

## Loading required package: stats4

##   
## Attaching package: 'bbmle'

## The following object is masked from 'package:dplyr':  
##   
## slice

#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},  
## }

citation("bbmle")

##   
## To cite package 'bbmle' in publications use:  
##   
## Ben Bolker 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  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Manual{,  
## title = {bbmle: Tools for General Maximum Likelihood Estimation},  
## author = {Ben Bolker and {R Development Core Team}},  
## year = {2020},  
## note = {R package version 1.0.23.1},  
## url = {https://CRAN.R-project.org/package=bbmle},  
## }

citation("gridExtra")

##   
## To cite package 'gridExtra' in publications use:  
##   
## Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for  
## "Grid" Graphics. R package version 2.3.  
## https://CRAN.R-project.org/package=gridExtra  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Manual{,  
## title = {gridExtra: Miscellaneous Functions for "Grid" Graphics},  
## author = {Baptiste Auguie},  
## year = {2017},  
## note = {R package version 2.3},  
## url = {https://CRAN.R-project.org/package=gridExtra},  
## }

##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 (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 found a positive, linear relationship between otolith width (mm) and length (mm) of herring from this dataset (Murchy et al. 2017). 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?

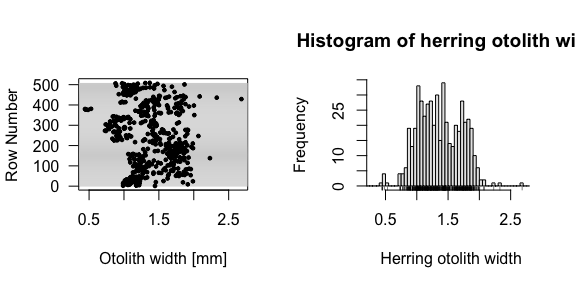
##C) This analysis is descriptive, with the goal of showing how explanatory variables describe patterns in herring size consumed by Chinook in the Salish Sea.

##D) We hypothesize that otolith width will be positively linear as salmon length increases. Otolith width and day of year will show a positive linear relationship with the calendar year. We expect a moderately positive, linear relationship with increasing latitude. We expect a negative linear relationship with longitude if larger herring occur along west Strait of Georgia. However, we could see a non-linear relationship with longitude if maximum herring size occurrs in mid-channel waters. We expect an interaction between salmon length and latitude, as there productivity in northern regions. An interaction between salmon length and longitude may occur if mainland stocks and those on Vancouver Island differ in size. We expect an interaction between salmon length and day of year, as each age class grows through the 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)

par(mfrow = c(1, 2))  
  
#Add column for number of rows  
n <- nrow(herring\_data)  
herring\_data$row <- 1:n  
  
#Make 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))  
  
#Make 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)



#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()

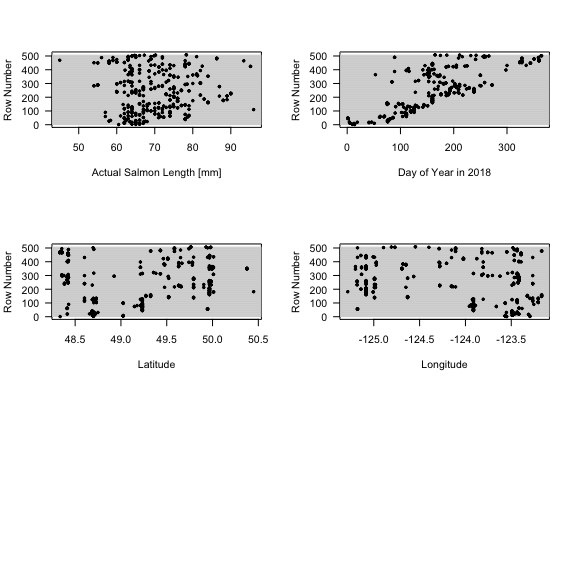
##B) The response variable (otolith width) is continuous and cannot take on zero or negative values. The range of possible response values in this dataset is 0.007 mm to 2.68 mm.

##C) Outliers at < 0.6 mm and > 2.2 mm are reasonable measurements representing a few individuals in the extreme 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.

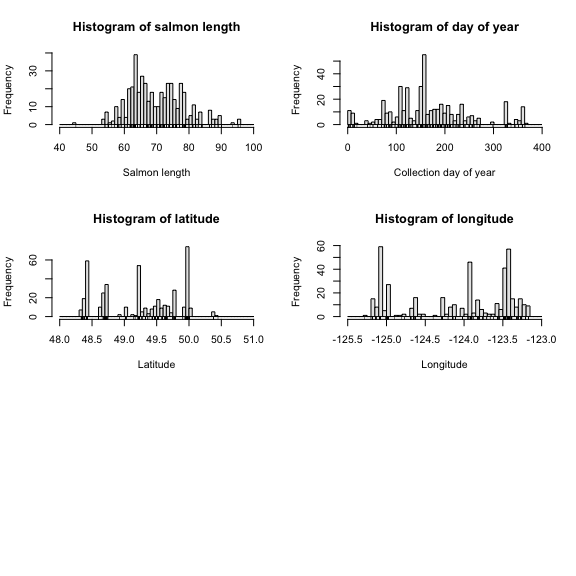
##D) We will consider the normal distribution for modelling.

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

#Make multipanel of dotcharts for continuous explanatory variables  
par(mfrow = c(3, 2))  
  
#Salmon length dotchart  
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 dotchart  
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 dotchart  
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 dotchart  
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))  
  
#Make multipanel of histograms for continuous explanatory variables  
par(mfrow = c(3, 2))

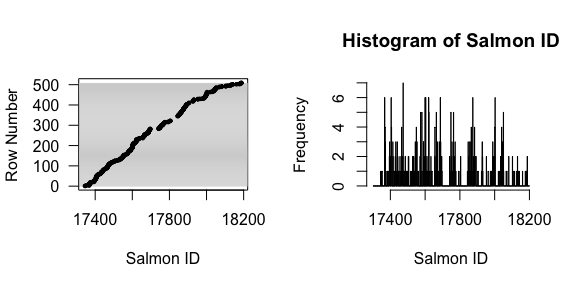


#Salmon length histogram  
hist(herring\_data$sal.length, col = grey(0.9), breaks = seq(40, 100, by = 1), main = "Histogram of salmon length", xlab = "Salmon length")  
rug(herring\_data$sal.length)  
  
#Collection day of year histogram  
hist(herring\_data$coll.doy, col = grey(0.9), breaks = seq(0, 400, by = 7), main = "Histogram of day of year", xlab = "Collection day of year")  
rug(herring\_data$coll.doy)  
  
#Latitude histogram  
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 histogram  
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)



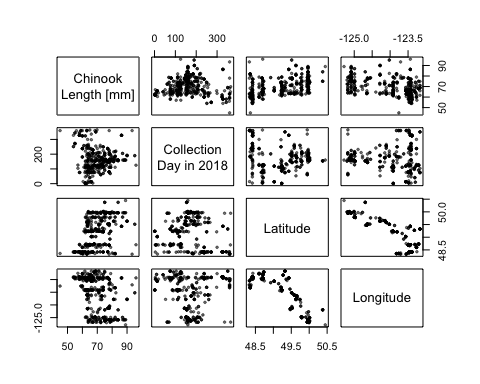
###Categorical explanatory variables

#Make multipanel for plots of categorical explanatory variables  
par(mfrow = c(1, 2))  
  
#Make dotchart of salmon id  
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  
hist(herring\_data$sal.id, col = grey(0.9), breaks = seq(17300, 18200, by = 1), main = "Histogram of Salmon ID", xlab = "Salmon ID")

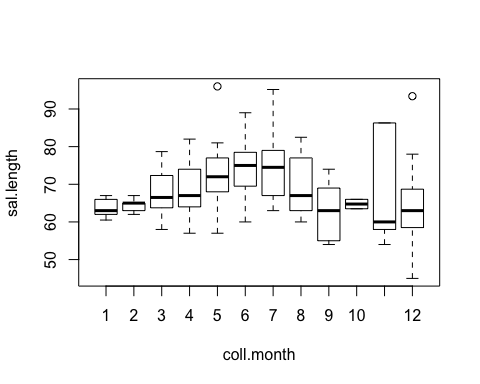
 ## B) One outlier is observed at 45cm and three outliers at 90 - 100cm. However, these are reasonable sizes for recreationally caught chinook salmon. No transformations are necessary as no substantial skewedness is observed.

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

#Create dataframe for pairwise scatter plots that only includes 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)  
  
#Make pairwise scatterplot of explanatory variables  
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"))



#Plot mean salmon size vs. collection month  
boxplot(sal.length ~ coll.month, data = herring\_data)



#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

Our only categorical variable is salmon id so the range in the number of otoliths per salmon was x to x will not confounded any other categorical variables. We plan to model salmon id as a random effect which will account for the effects of any unbalancing associated with salmon id on the continuous variables.

A boxplot of mean salmon length per month shows that mean salmon length in summer months overlapped with mean salmon length observed in earlier months in the year, likely due to unbalanced sampling effort over the year as opposed 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.

\*get otolith number per salmon id range

#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 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 (ref.).

\*ref for > 10 levels

##Standardizing data for models

#Add columns for standardized continuous predictor variable for modelling (x - minus 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)

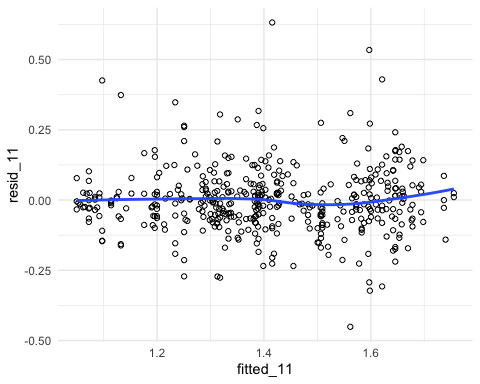
##Creating linear mixed-effect models

#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)

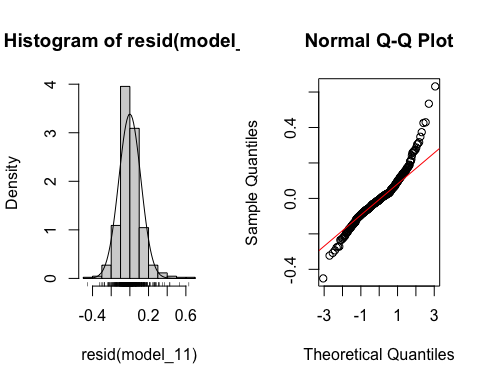
#Section 6: Model Checking ##Checking assumptions of model 11

#Plot residuals versus fitted values (shown with a smoothing curve)  
herring\_data$resid\_11 <- resid(model\_11)   
herring\_data$fitted\_11 <- predict(model\_11, re.form = ~0)  
  
ggplot(herring\_data) +  
 theme\_minimal() +   
 geom\_point(aes(fitted\_11, resid\_11), shape = 1) +  
 geom\_smooth(aes(fitted\_11, resid\_11), se = FALSE)

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



#Make multipanel of residual plots  
par(mfrow = c(1, 2))  
  
#Plot histogram of residuals as probability density with rug marks and reference curve  
hist(resid(model\_11), prob = TRUE, col = "lightgrey")  
rug(resid(model\_11))  
sd\_hat <- sd(resid(model\_11))  
curve(dnorm(x, sd = sd\_hat), add = TRUE)  
  
#Make quantile-quantile plot with reference line  
qqnorm(resid(model\_11))  
qqline(resid(model\_11), col = "red")

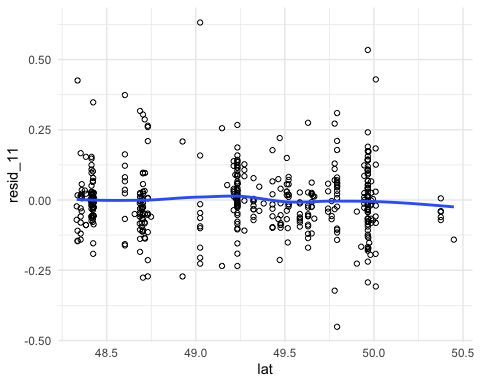


#Run a Shapiro-Wilk test for normality of the residuals  
shapiro.test(resid(model\_11))

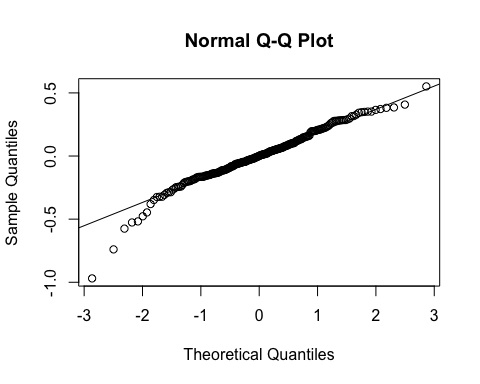
##   
## Shapiro-Wilk normality test  
##   
## data: resid(model\_11)  
## W = 0.94396, p-value = 7.724e-12

#Plot residuals vs. lat (shown with smoothing curve)  
resid\_vs\_lat\_11 <- ggplot(herring\_data) +  
 theme\_minimal() +   
 geom\_point(aes(lat, resid\_11), shape = 1) +  
 geom\_smooth(aes(lat, resid\_11), se = FALSE)  
  
resid\_vs\_lat\_11

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

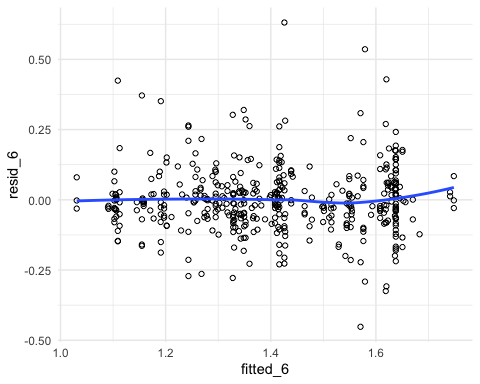


#Make multipanel of residual plots  
par(mfrow = c(1, 1))  
  
#Make quantile-quantile plot for checking whether the random intercept is normally distributed  
qqnorm(ranef(model\_11)$sal.id[, "(Intercept)"])  
qqline(ranef(model\_11)$sal.id[, "(Intercept)"])

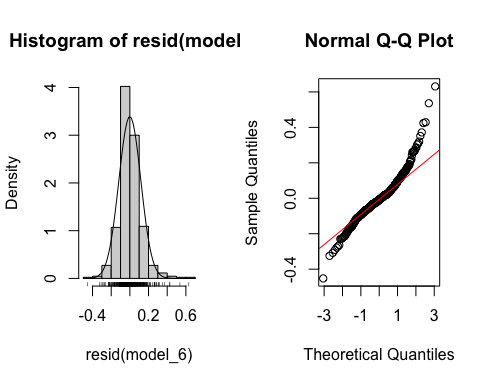
 ##Checking assumptions of model 6

#Plot residuals versus fitted values (shown with a smoothing curve)  
herring\_data$resid\_6 <- resid(model\_6)   
herring\_data$fitted\_6 <- predict(model\_6, re.form = ~0)  
  
ggplot(herring\_data) +  
 theme\_minimal() +   
 geom\_point(aes(fitted\_6, resid\_6), shape = 1) +  
 geom\_smooth(aes(fitted\_6, resid\_6), se = FALSE)

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



#Make multipanel of residual plots  
par(mfrow = c(1, 2))  
  
#Plot histogram of residuals as probability density with rug marks and reference curve  
hist(resid(model\_6), prob = TRUE, col = "lightgrey")  
rug(resid(model\_6))  
sd\_hat <- sd(resid(model\_6))  
curve(dnorm(x, sd = sd\_hat), add = TRUE)  
  
#Make quantile-quantile plot with reference line  
qqnorm(resid(model\_6))  
qqline(resid(model\_6), col = "red")

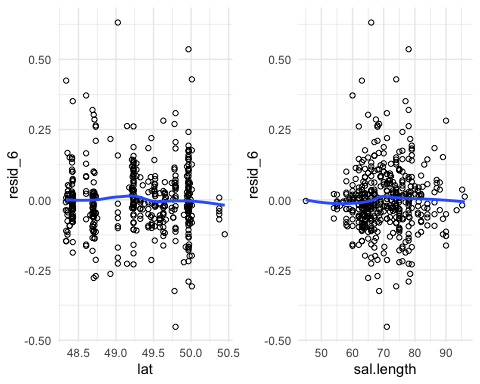


#Run a Shapiro-Wilk test for normality of the residuals  
shapiro.test(resid(model\_6))

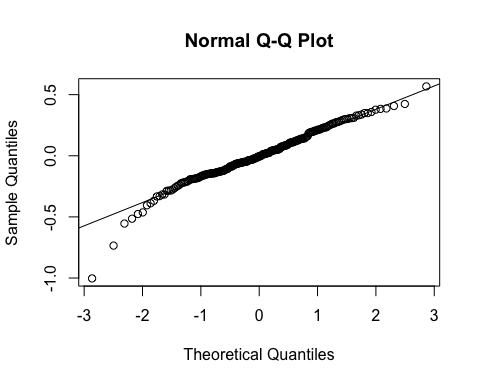
##   
## Shapiro-Wilk normality test  
##   
## data: resid(model\_6)  
## W = 0.94325, p-value = 6.278e-12

#Plot residuals vs. longitude (shown with smoothing curve)  
resid\_vs\_lat\_6 <- ggplot(herring\_data) +  
 theme\_minimal() +   
 geom\_point(aes(lat, resid\_6), shape = 1) +  
 geom\_smooth(aes(lat, resid\_6), se = FALSE)  
  
#Plot residuals vs. sal.length (shown with smoothing curve)  
resid\_vs\_sal.length\_6 <- ggplot(herring\_data) +  
 theme\_minimal() +   
 geom\_point(aes(sal.length, resid\_6), shape = 1) +  
 geom\_smooth(aes(sal.length, resid\_6), se = FALSE)  
  
grid.arrange(resid\_vs\_lat\_6, resid\_vs\_sal.length\_6, nrow = 1)

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



#Make multipanel of residual plots  
par(mfrow = c(1, 1))  
  
#Make quantile-quantile plot for checking whether the random intercept is normally distributed  
qqnorm(ranef(model\_6)$sal.id[, "(Intercept)"])  
qqline(ranef(model\_6)$sal.id[, "(Intercept)"])

 1. Written summary (200 words)

-assessment of plausibility of model assumptions (LINE) -check for correlation in the residuals due to grouping variables (e.g., repeated measurements on individuals or within study sites), spatial autocorrelation and/or temporal autocorrelation -comment on potential outliers, as relevant

#Section 7: Model Summary, Conﬁdence Intervals and Model Comparison ##Rewriting all models with REML= FALSE

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

##Running the AIC

#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

##Summarizing model 11

#Call coeﬃcient estimates for fixed effects  
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 salmon.id  
VarCorr(model\_11)

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

#Calculate conﬁdence intervals for coeﬃcient estimate  
confint(model\_11)

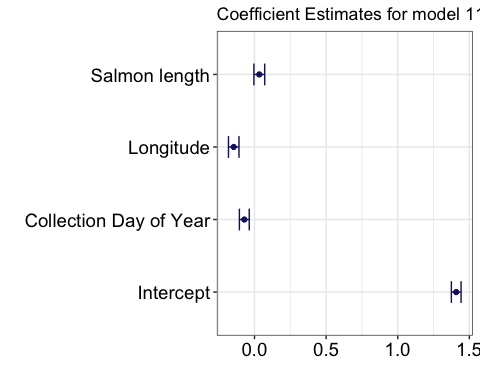
## Computing profile confidence intervals ...

## 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

#Rearrange data for a plot of coefficients with standard errors  
#Make data frame for plot of coefficients  
confint\_fixed\_11 <- as.data.frame(confint(model\_11, parm = c("(Intercept)", "coll.doy\_standardized", "long\_standardized", "sal.length\_standardized")))

## Computing profile confidence intervals ...

##Compute profile confidence intervals  
confint\_fixed\_11 <- as.data.frame(confint\_fixed\_11)  
confint\_fixed\_11$term <- rownames(confint\_fixed\_11)  
confint\_fixed\_11$estimate <- fixef(model\_11)  
  
##ggplot  
ggplot(confint\_fixed\_11) +   
 theme\_bw() +  
 geom\_point(aes(y = term, x = estimate), colour = "midnightblue") +  
 geom\_errorbarh(aes(y = term, xmin = `2.5 %`, xmax = `97.5 %`), colour = "midnightblue", height = 0.3) +  
 scale\_y\_discrete(labels = c("Intercept", "Collection Day of Year", "Longitude", "Salmon length")) +  
 xlab("") +  
 ylab("") +   
 ggtitle("Coefficient Estimates for model 11") +  
 theme(axis.text = element\_text(size = 14, colour = "black"))

 ##Summarizing model 6

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

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

#Call variation within and between salmon.id  
VarCorr(model\_6)

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

#Calculate conﬁdence intervals for coeﬃcient estimate  
confint(model\_6)

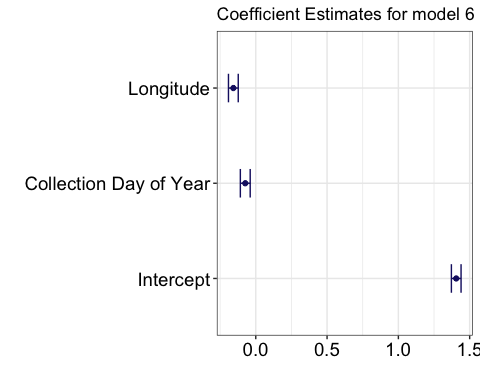
## Computing profile confidence intervals ...

## 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

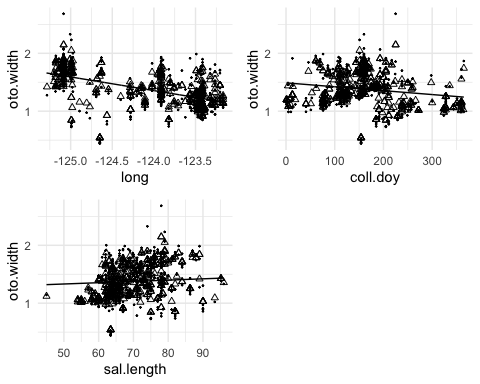
#Rearrange data for a plot of coefficients with standard errors  
#Make data frame for plot of coefficients  
confint\_fixed\_6 <- as.data.frame(confint(model\_6, parm = c("(Intercept)", "coll.doy\_standardized", "long\_standardized")))

## Computing profile confidence intervals ...

##Compute profile confidence intervals  
confint\_fixed\_6 <- as.data.frame(confint\_fixed\_6)  
confint\_fixed\_6$term <- rownames(confint\_fixed\_6)  
confint\_fixed\_6$estimate <- fixef(model\_6)  
  
##ggplot  
ggplot(confint\_fixed\_6) +   
 theme\_bw() +  
 geom\_point(aes(y = term, x = estimate), colour = "midnightblue") +  
 geom\_errorbarh(aes(y = term, xmin = `2.5 %`, xmax = `97.5 %`), colour = "midnightblue", height = 0.3) +  
 scale\_y\_discrete(labels = c("Intercept", "Collection Day of Year", "Longitude")) +  
 xlab("") +  
 ylab("") +   
 ggtitle("Coefficient Estimates for model 6") +  
 theme(axis.text = element\_text(size = 14, colour = "black"))

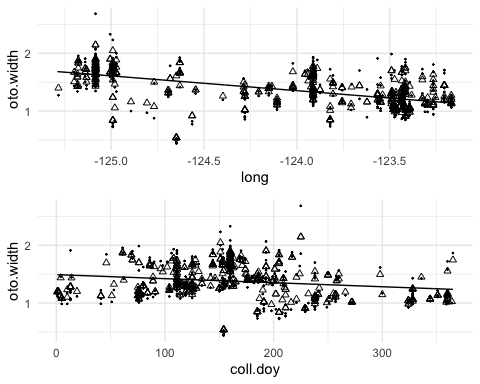
 #Section 8: Plotting a Model with the Data ##Plot model 11

#Create data frame for plot of model 11 predictions with longitude on the x axis  
oto.width\_vs\_long\_data\_11 <- data.frame(long = seq(-125.29, -123.16, 0.001), coll.doy\_standardized = median(herring\_data$coll.doy\_standardized), sal.length\_standardized = median(herring\_data$sal.length\_standardized))  
  
#Add standardized long data to data frame  
oto.width\_vs\_long\_data\_11$long\_standardized <- ((oto.width\_vs\_long\_data\_11$long - mean(oto.width\_vs\_long\_data\_11$long))/sd(oto.width\_vs\_long\_data\_11$long))  
  
#Generate model predictions at population level   
oto.width\_vs\_long\_data\_11$fit\_11\_pop <- predict(model\_11, newdata = oto.width\_vs\_long\_data\_11, re.form = ~0)  
  
#Generate model predictions at salmon id level  
herring\_data$fit\_11\_sal.id <- predict(model\_11)  
  
#Plot the model vs. longitude  
model\_11\_long <- ggplot(herring\_data) +  
 theme\_minimal() +   
 geom\_point(aes(long, oto.width), shape = 1, size = 0.2) +  
 geom\_line(aes(long, fit\_11\_pop), data = oto.width\_vs\_long\_data\_11) +   
 geom\_point(aes(long, fit\_11\_sal.id, group = sal.id), shape = 2)  
  
#Create data frame for plot of model 11 predictions with coll.doy on the x axis  
oto.width\_vs\_coll.doy\_data\_11 <- data.frame(coll.doy = seq(1, 365, 1), long\_standardized = median(herring\_data$long\_standardized), sal.length\_standardized = median(herring\_data$sal.length\_standardized))  
  
#Add standardized coll.doy data to data frame  
oto.width\_vs\_coll.doy\_data\_11$coll.doy\_standardized <- ((oto.width\_vs\_coll.doy\_data\_11$coll.doy - mean(oto.width\_vs\_coll.doy\_data\_11$coll.doy))/sd(oto.width\_vs\_coll.doy\_data\_11$coll.doy))  
  
#Generate model predictions at population level   
oto.width\_vs\_coll.doy\_data\_11$fit\_11\_pop <- predict(model\_11, newdata = oto.width\_vs\_coll.doy\_data\_11, re.form = ~0)  
  
#Generate model predictions at salmon id level  
herring\_data$fit\_11\_sal.id <- predict(model\_11)  
  
#Plot the model vs.coll.doy  
model\_11\_coll.doy <- ggplot(herring\_data) +  
 theme\_minimal() +   
 geom\_point(aes(coll.doy, oto.width), shape = 1, size = 0.2) +  
 geom\_line(aes(coll.doy, fit\_11\_pop), data = oto.width\_vs\_coll.doy\_data\_11) +   
 geom\_point(aes(coll.doy, fit\_11\_sal.id, group = sal.id), shape = 2)  
  
#Create data frame for plot of model 11 predictions with sal.length on the x axis  
oto.width\_vs\_sal.length\_data\_11 <- data.frame(sal.length = seq(45, 96, 0.5), long\_standardized = median(herring\_data$long\_standardized), coll.doy\_standardized = median(herring\_data$coll.doy\_standardized))  
  
#Add standardized sal.length data to data frame  
oto.width\_vs\_sal.length\_data\_11$sal.length\_standardized <- ((oto.width\_vs\_sal.length\_data\_11$sal.length - mean(oto.width\_vs\_sal.length\_data\_11$sal.length))/sd(oto.width\_vs\_sal.length\_data\_11$sal.length))  
  
#Generate model predictions at population level   
oto.width\_vs\_sal.length\_data\_11$fit\_11\_pop <- predict(model\_11, newdata = oto.width\_vs\_sal.length\_data\_11, re.form = ~0)  
  
#Plot the model vs. sal.length  
model\_11\_sal.length <- ggplot(herring\_data) +  
 theme\_minimal() +   
 geom\_point(aes(sal.length, oto.width), shape = 1, size = 0.2) +  
 geom\_line(aes(sal.length, fit\_11\_pop), data = oto.width\_vs\_sal.length\_data\_11) +   
 geom\_point(aes(sal.length, fit\_11\_sal.id, group = sal.id), shape = 2)  
  
grid.arrange(model\_11\_long, model\_11\_coll.doy, model\_11\_sal.length, nrow = 2)



##Plot model 6

#Create data frame for plot of model 6 predictions with longitude on the x axis  
oto.width\_vs\_long\_data\_6 <- data.frame(long = seq(-125.29, -123.16, 0.001), coll.doy\_standardized = median(herring\_data$coll.doy\_standardized), sal.length\_standardized = median(herring\_data$sal.length\_standardized))  
  
#Add standardized long data to data frame  
oto.width\_vs\_long\_data\_6$long\_standardized <- ((oto.width\_vs\_long\_data\_6$long - mean(oto.width\_vs\_long\_data\_6$long))/sd(oto.width\_vs\_long\_data\_6$long))  
  
#Generate model predictions at population level   
oto.width\_vs\_long\_data\_6$fit\_6\_pop <- predict(model\_6, newdata = oto.width\_vs\_long\_data\_6, re.form = ~0)  
  
#Generate model predictions at salmon id level  
herring\_data$fit\_6\_sal.id <- predict(model\_6)  
  
#Plot the model vs. longitude  
model\_6\_long <- ggplot(herring\_data) +  
 theme\_minimal() +   
 geom\_point(aes(long, oto.width), shape = 1, size = 0.2) +  
 geom\_line(aes(long, fit\_6\_pop), data = oto.width\_vs\_long\_data\_6) +   
 geom\_point(aes(long, fit\_6\_sal.id, group = sal.id), shape = 2)  
  
#Create data frame for plot of model 6 predictions with coll.doy on the x axis  
oto.width\_vs\_coll.doy\_data\_6 <- data.frame(coll.doy = seq(1, 365, 1), long\_standardized = median(herring\_data$long\_standardized), sal.length\_standardized = median(herring\_data$sal.length\_standardized))  
  
#Add standardized coll.doy data to data frame  
oto.width\_vs\_coll.doy\_data\_6$coll.doy\_standardized <- ((oto.width\_vs\_coll.doy\_data\_6$coll.doy - mean(oto.width\_vs\_coll.doy\_data\_6$coll.doy))/sd(oto.width\_vs\_coll.doy\_data\_6$coll.doy))  
  
#Generate model predictions at population level   
oto.width\_vs\_coll.doy\_data\_6$fit\_6\_pop <- predict(model\_6, newdata = oto.width\_vs\_coll.doy\_data\_6, re.form = ~0)  
  
#Generate model predictions at salmon id level  
herring\_data$fit\_6\_sal.id <- predict(model\_6)  
  
#Plot the model vs.coll.doy  
model\_6\_coll.doy <- ggplot(herring\_data) +  
 theme\_minimal() +   
 geom\_point(aes(coll.doy, oto.width), shape = 1, size = 0.2) +  
 geom\_line(aes(coll.doy, fit\_6\_pop), data = oto.width\_vs\_coll.doy\_data\_6) +   
 geom\_point(aes(coll.doy, fit\_6\_sal.id, group = sal.id), shape = 2)  
  
#Create data frame for plot of model 6 predictions with sal.length on the x axis  
oto.width\_vs\_sal.length\_data\_6 <- data.frame(sal.length = seq(45, 96, 0.5), long\_standardized = median(herring\_data$long\_standardized), coll.doy\_standardized = median(herring\_data$coll.doy\_standardized))  
  
grid.arrange(model\_6\_long, model\_6\_coll.doy, nrow = 2)



#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.