STATS5620_Q2_Analysis

STAT 5620 Project Outline

Linking Diet and Reproductive Success in Northwest Atlantic Grey Seals

Max Henry-Adams (B00952295) and Peter Braithwaite (B00504978)

Data Analysis STAT 4620/5620 Winter 24-25"

Link to GitHub Repository: https://github.com/MaxHA17/STAT5620.Project

Abstract:

Keywords: Capital breeding strategy, Generalized Linear Models, Pinnipeds, Quantitative Fatty Acid Signature Analysis, Reproductive Ecology

Introduction

For marine predators, prey distribution may vary unpredictably in time and space, undergoing seasonal, inter annual, and longer-term changes in environmental conditions. Understanding how top predators use different foraging strategies to navigate and adapt to patchy and unpredictable prey availability has important implications for individual fitness and population dynamics, yet studies exploring the consequences of such decisions on fitness are relatively rare (Nathan et al., 2008).

Vertebrates fuel the costs of reproduction along a continuum of an income-capital breeding strategy, where animals that use an income breeding strategy rely on concurrent energy accumulation during the breeding period as opposed to a capital breeding strategy, where energy acquired and stored prior to the breeding period is used solely to finance major reproductive costs (Stephens et al., 2009; Stephens et al., 2014). A capital breeding strategy is thought to be advantageous for large animals capable of carrying large energy stores and offers energetic benefits in areas of patchy or unreliable food availability and allows parents to direct more resources (e.g. time and energy) towards offspring, as opposed to foraging, during the breeding and lactation period (Stephens et al., 2009; Stephens et al., 2014). Given the importance of pre-breeding and lactation period foraging success and energy acquisition in capital breeders, these species offer interesting model systems to study the relationships between diet and reproductive success.

Phocid seals are wide-ranging, large-bodied, and long-lived marine predators, with individuals of many species foraging across a wide range of habitats over large temporal scales (Bowen and Jonsen, 2022). Grey seals (*Halichoerus grypus*) are a long-lived, sexually dimorphic, phocid species and are capital breeders. In the Sable Island population, pregnant females haul out in late December or early January, give birth to a single pup and nurse that pup for 16-18 days, relying solely on energy accumulated during the foraging period prior to parturition to support the costs of lactation (Iverson et al., 1993; Stephens et al., 2009). At or near the abrupt weaning of that pup, females are mated and return to sea to replenish spent body reserves and, after a several month period of delayed implantation, to support gestation and most importantly preparation for the subsequent December or early January lactation period. As in all mammals, the costs of lactation, and its contribution to reproductive success, far exceed the costs of gestation. Like most other phocids, grey seals are generalist predators (Bowen and Jonsen, 2022), but females tend to feed on a narrower range of energy rich species (predominantly sand lance, redfish, and other pelagic fishes) following the lactation period and expand their prey diversity in the several months leading up to the December-January breeding and lactation period (Beck et al., 2007). Female grey seals reach sexual maturity at age 4-5 years and can continue to reproduce as late as into their early 40s (Bowen et al., 2006). During the brief 16–18-day lactation period, females utilize approximately 25-38% of parturition body mass and daily maternal mass loss is a significant predictor of pup growth rates during lactation, with heavier mothers at parturition weaning heavier pups (Iverson et al., 1993; Mellish et al., 1999). While pup survival is positively related to pup weaning mass up to the mean population weaning mass of 51.5 kg, body length at weaning is the best predictor of grey seal pup survival, as predicted by a bigger-is-better hypothesis for size selective mortality (Bowen et al., 2015). The phenomena of heavier females weaning larger offspring is evident even from primiparity where it is believed that heavier females have larger resource stores that they can mobilize during lactation to produce larger offspring at weaning (Iverson et al., 1993; Mellish et al., 1999; Bowen et al., 2006). Following the lactation period, parental care abruptly ends and the pup is left to fast and survive for several weeks on the energy acquired during lactation before undertaking its first foraging trip (Noren et al., 2008). Thus, answering questions relating to the diet of females prior to parturition will improve our understanding of the importance of habitat use and energy acquisition as they relate to life history characteristics, such as reproductive success, in these capital breeders.

Northwest Atlantic grey seals breeding on Sable Island, Nova Scotia, tend to be central place foragers and have high fidelity to Sable Island for breeding and as a haul-out site, making them excellent candidates for long-term studies (Bowen et al., 2015). Using diet estimations methods, such as quantitative fatty acid (FA) signature analysis (QFASA; Iverson et al., 2004, Beck et al., 2007), it becomes possible to explore the relationship between diet variation and reproductive success at the individual level. QFASA is used to estimate predator diet by comparing the FA signatures of predator adipose tissue to those of candidate prey species (Iverson, 2009; Karnovsky et al., 2012). Predator FA signatures alone have historically been used to provide a qualitative assessment of spatial and temporal patterns in diet diversity. However, since many FAs that are biosynthesized within unique prey species accumulate in predator adipose tissue over time, evaluating the amount of prey-specific FAs relative to the predator’s overall FA stores can provide a quantitative estimate of diet, known as QFASA (Iverson et al., 2004). For accurate diet estimation, species-specific calibration coefficients must be experimentally derived to account for the altered incorporation of prey FAs into the predator adipose tissue due to lipid metabolic processes within the predator (Iverson et al., 2004; Karnovsky et al., 2012). When calibration coefficients are derived and the FA signature of many or all potential prey species in an ecosystem is known, as in the case of the Northwest Atlantic grey seal range (Iverson et al., 2004; Beck et al., 2007), QFASA can provide a statistical approach to accurately depicting diet and prey diversity over ecologically relevant time scales (e.g., weeks to several months).

This study will test the hypothesis that differences in diet will affect an individual’s ability to store energy prior to parturition, and thus influence their reproductive success. Specifically, we will explore two research questions:

1. Does variation in diet influence maternal mass change over the foraging period leading up to parturition in Northwest Atlantic grey seals?

2. Does female diet predict pup weaning mass in Northwest Atlantic grey seals?

Data description

Response variables (organized by research question):

1. Maternal mass change during the pre-breeding season foraging period: continuous variable ranging from -45 – 88 kg. Calculated by subtracting the maternal recovery mass from her deployment mass.

2. Pup weaning mass: continuous variable ranging from 34.5 – 63 kg. Denotes the mass of the pup after the mother has terminated the lactation period and departed the pup and/or Sable Island.

Predictor or supporting variables

3. Year: discrete variable with values ranging from 1996-2015. Denotes the year in which each observation was collected.

4. Maternal ID: discrete variable with a unique identifier for each individual female.

5. Pup sex: categorical; male or female.

6. Maternal age: discrete variable with values ranging from 7-40 years old.

7. Maternal deployment mass: continuous variable with values ranging from 116 – 250 kg. Describes the female mass when she was located pregnant on Sable Island 3-6 months prior to parturition and prior to undergoing her pre-parturition foraging period.

8. Maternal mass at parturition: continuous variable ranging from ~ 137 – 250 kg. This is the mass of the female after returning to Sable Island and giving birth to her pup.

9. Maternal dominant prey species: discrete variable with 6 unique values. Represents the species making up the greatest proportion of the female diet.

10. Maternal dietary energy density: continuous variable ranging from 4.88 – 6.51 kJ/g. Energy density (kJ/g) of the female diet is calculated from the average energy content per gram of prey body tissue.

11. Maternal diet diversity: continuous variable 0.18 – 0.58, no specific units. Diversity is calculated using a standardized Shannon-Weaver diversity index to quantify individual diet diversity.

Method

The initial step was data cleaning to ensure all readily apparent data errors were rectified prior to data analysis and model fitting. Most of the data errors discovered resulted from the absence of critical data and thus these observations were removed from the dataset. The first question dataset resulted an initial dataset of 56 observations (n=56) and the second question resulted in 76 observations (n=76). It should also be noted that additional observations were filtered from the dataset throughout the analysis process as = outliers with high leverage were also removed. Thus, the final dataset used for question one had 72 observation and the final dataset for question two had 50 observations.

Once cleaned and uploaded to RStudio, initial data plots were made using the 'flexplot' and 'plot_explore,' the R function we have included in our Git Hub. These visualizations facilitated the exploration of variable distributions, specifically the response variables. From this it was determined that the response variables for both questions, Maternal Mass Change and Pup Weaning Mass, were Gaussian distributed. This normality allowed the exploration of the data through General Linear Models as well as more complex Generalized Linear Models and Mixed Models. Although our original analytical approach proposal Generalized Linear Mixed Models, it felt prudent to explore the efficiency, interpret ability and computational simplify of Linear Models if the assumption were met and the data suggested linearity.

As a result, both questions were analysed first as Linear Models (LM), then as Generalized Linear Models (GLM) with Gaussian distributions, and finally as Generalized Linear Mixed Models (GLMM) with Random Effects and Fixed Effects on some and all the categorical data. With each model type we began with all the predictor variables prior to using the backward and forward "Step" function to reduce the number of variables through elimination based on the p-values for Linear Models and AIC for GLM and GLMM.

Once we had established the best fit model for both questions using the AIC these models were tested using Cross Validation, question one with 7 folds and question two with 11 folds.

Analysis

Research Question 1

Here we are exploring whether variation in diet influence maternal mass change over the foraging period leading up to parturition in Northwest Atlantic grey seals?

The response variable of interest is maternal mass change. We will first look at the distribution of this variable:

```
#Load the data
library(readr)
Data_Q1<- read.csv("/Users/peterbraithwaite/Desktop/IDPhD Classes/Stat 5620_Updated/Final

#Ensure the data types are correctly assigned
Data_Q1$MomID <- as.factor(Data_Q1$MomID)
Data_Q1$Year <- factor(Data_Q1$Year, levels = sort(unique(Data_Q1$Year)), ordered = TRUE)
```

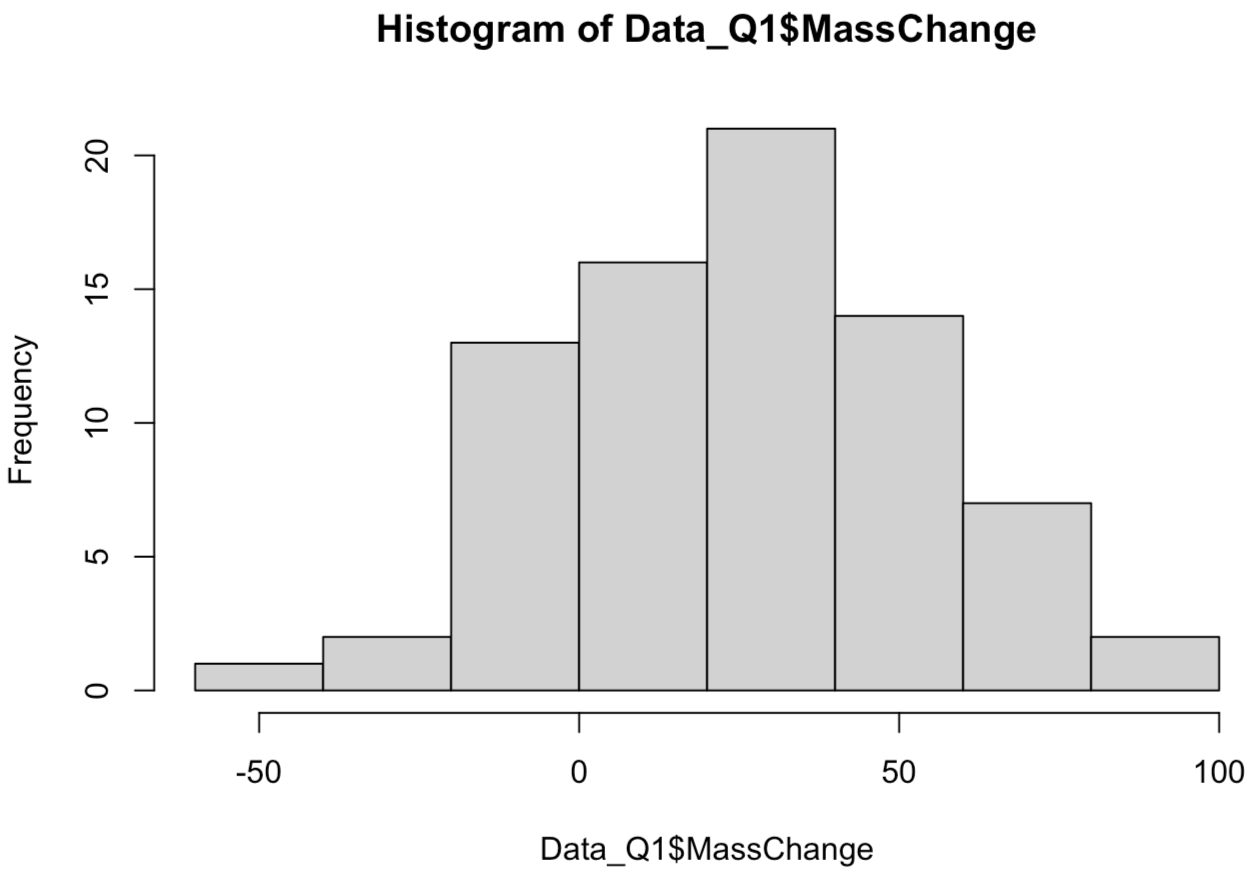
```r
Data_Q1$`Dominant.prey.species`<- as.character(Data_Q1$`Dominant.prey.species`)
Data_Q1$`Diet.diversity` <- as.numeric(Data_Q1$`Diet.diversity`)
Data_Q1$`Dietary.energy.density` <- as.numeric(Data_Q1$`Dietary.energy.density`)

summary(Data_Q1)
```

```
    MomID          Year    Dietary.energy.density Diet.diversity  
 4269   : 3   2013   :13   Min.   :4.881          Min.   :0.1788  
 24     : 2   2011   :11   1st Qu.:5.623          1st Qu.:0.3063  
 146    : 2   2010   : 9   Median :5.753          Median :0.3572  
 829    : 2   2012   : 9   Mean   :5.736          Mean   :0.3625  
 3271   : 2   2009   : 6   3rd Qu.:5.893          3rd Qu.:0.4181  
 3616   : 2   2015   : 5   Max.   :6.517          Max.   :0.5782  
 (Other):63   (Other):23                                          
 Dominant.prey.species  Mass.change    
 Length:76             Min.   :-40.50  
 Class :character      1st Qu.:  6.25  
 Mode  :character      Median : 22.75  
                       Mean   : 26.02  
                       3rd Qu.: 45.00  
                       Max.   : 88.00  
```

```r
#rename columns
library(dplyr)
```

```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
Data_Q1 <- Data_Q1 %>%
  rename(
    DietEngDen = `Dietary.energy.density`,
    DietDiv = `Diet.diversity`,
    DomSpp = `Dominant.prey.species`,
    MassChange = `Mass.change`)

hist(Data_Q1$MassChange)
```



There are a total of n=76 observations and mass change appears to follow a normal distribution a Gaussian family distribution is a good starting point for an initial model.

One of the covariates, dominant prey species, is a categorical variable. We will first check the number of observations in each category:

```
table(Data_Q1$DomSpp)
```

```
       AtlanticCod            Capelin NorthernSandlance           Pollock 
                 6                  3                29                 5 
           Redfish          WhiteHake 
                32                  1 
```

With only one observation of white hake and 3 observations of capelin, there is likely not enough data to estimate the effect of these species on the response variable and we are unable to draw meaningful conclusions. These points also cause extreme residual values. Capelin and white hake observations will be removed from the data set:

```
Data_Q1 <- Data_Q1 %>%
  filter(DomSpp != "WhiteHake")
```

```
Data_Q1 <- Data_Q1 %>%
  filter(DomSpp != "Capelin")
```

The initial model seeks to model maternal mass change as a function of dominant prey species, diet diversity, dietary energy density, and deployment year. We will now explore the apparent relationships between each continuous and categorical covariate using the plot_explore function built and contained in the STAT5620.Project R package:

Initially, it appears that diet diversity is negatively related to mass change, with energy density possibly having a slight negative relationship. It also appears that animals with cod, redfish or sand lance as the dominant prey species gain more mass during foraging than females who primarily prey on Pollock. There does not appear to be any clear trends in mass change over time.

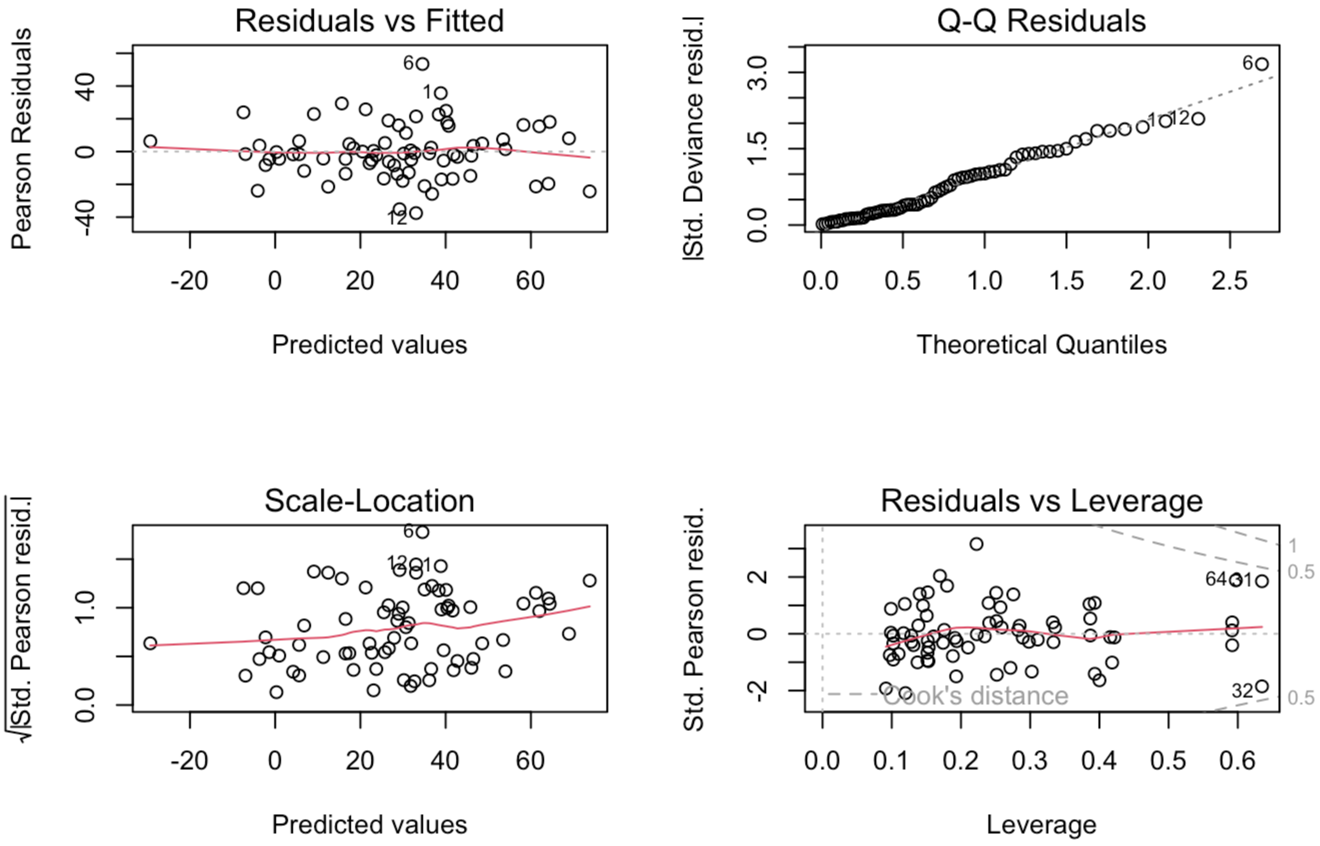
We will begin with an initial model with all covariates of interest. Based on the distribution of mass change, we will begin with a simple model with a gaussian distribution and an identity link function:

```
library(lme4)
```

```
Loading required package: Matrix
```

```
mom_mod <- glm(data=Data_Q1, family = gaussian(link="identity"), formula = MassChange ~ D

par(mfrow = c(2, 2))
plot(mom_mod)
```

```
Warning: not plotting observations with leverage one:
  45
```



The residual plots from this initial model show good homoscedasticity and normality. There is one observation (observation 6) that appears to be an extreme value. This observation comes from a female that experiences an 88 kg increase in mass over the pre-breeding period foraging period, which is well above the mean mass change value of 28.5. However, none of the covariate values associated with this observation are abnormal and this point does not have high leverage. As a results, the decision is to retain this data point since it is not highly influential and knowledge of the system suggests it is a valid data point.

We can now use step selection to preform repeated hypothesis tests for variable selection to determine which covariates best explain maternal mass change while minimizing model complexity and colinearity between covariates.

```
step(mom_mod, direction = "both")
```

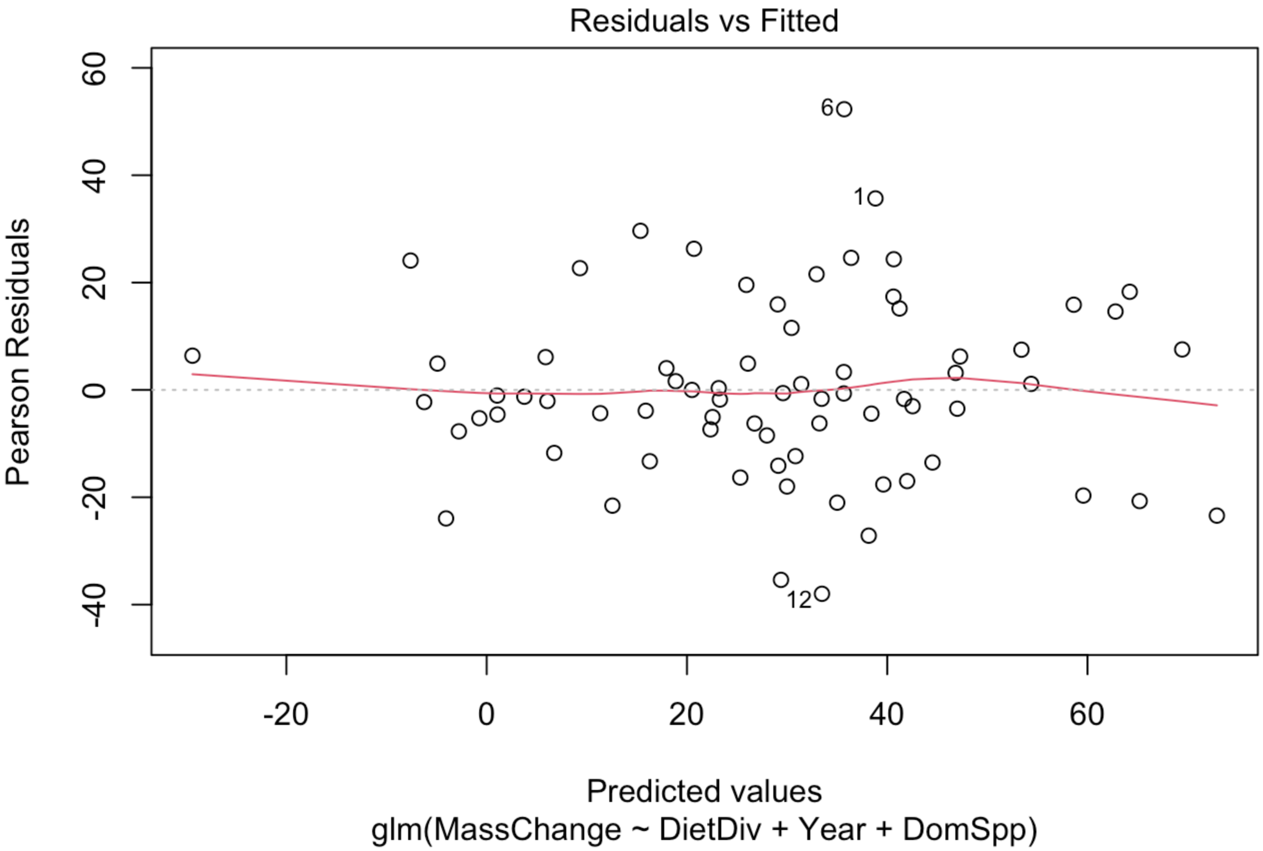
```
Start:  AIC=647.55
MassChange ~ DietDiv + DietEngDen + Year + DomSpp

             Df Deviance    AIC
- DietEngDen  1    19508 645.67
<none>             19478 647.55
- DietDiv     1    20858 650.48
- DomSpp      3    23417 654.82
- Year       13    38977 671.50
```

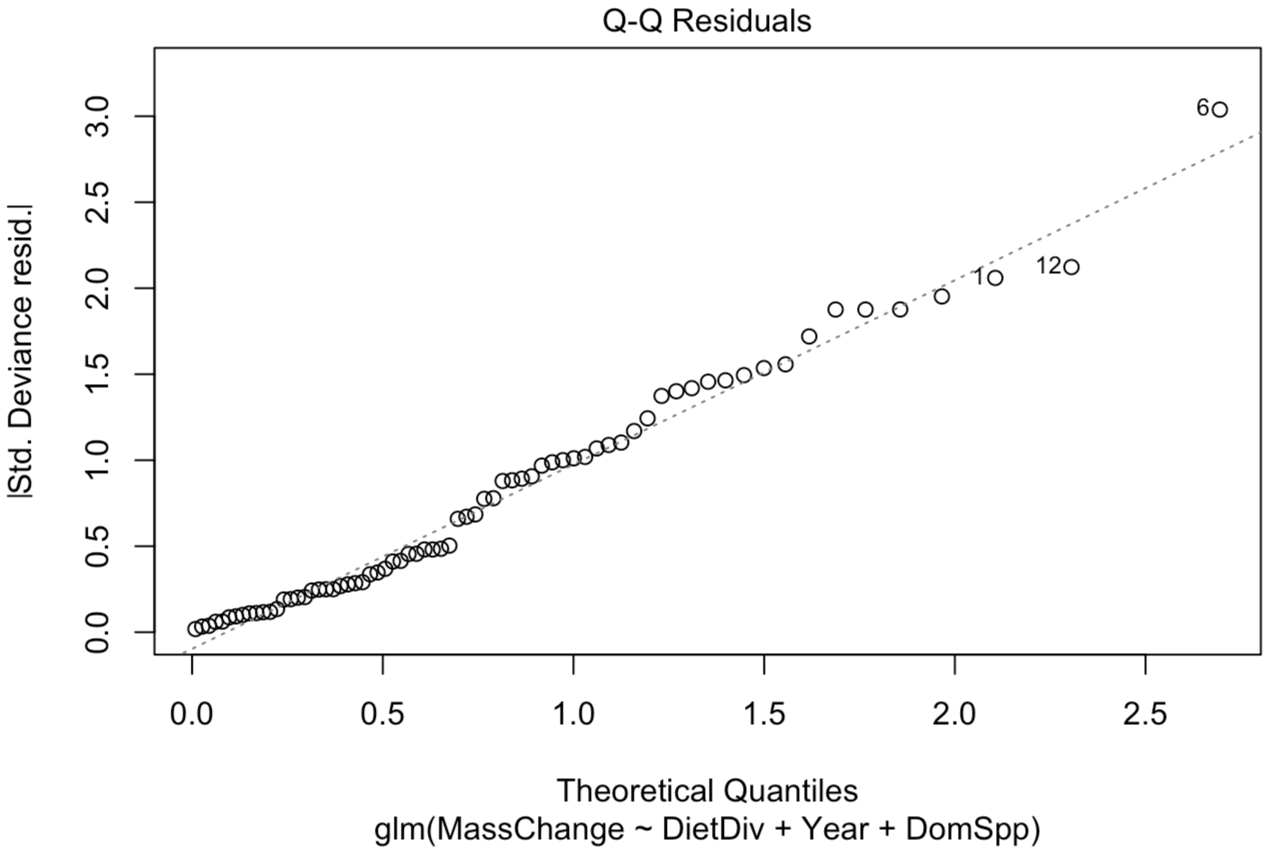
```
Step:  AIC=645.67
MassChange ~ DietDiv + Year + DomSpp

             Df Deviance    AIC
<none>             19508 645.67
+ DietEngDen  1    19478 647.55
- DietDiv     1    20891 648.60
- DomSpp      3    23602 653.38
- Year       13    39769 670.95

Call:  glm(formula = MassChange ~ DietDiv + Year + DomSpp, family = gaussian(link =
"identity"),
    data = Data_Q1)

Coefficients:
            (Intercept)                  DietDiv                   Year.L
                 70.599                  -63.415                   28.367
                 Year.Q                   Year.C                   Year^4
                 12.387                   18.307                   15.184
                 Year^5                   Year^6                   Year^7
                -15.549                    3.305                   23.361
                 Year^8                   Year^9                  Year^10
                  2.392                  -23.277                  -11.186
                Year^11                  Year^12                  Year^13
                 13.261                   21.328                  -39.502
DomSppNorthernSandlance             DomSppPollock            DomSppRedfish
                -18.212                  -47.479                  -20.303

Degrees of Freedom: 71 Total (i.e. Null);  54 Residual
Null Deviance:      48570
Residual Deviance: 19510    AIC: 645.7
```

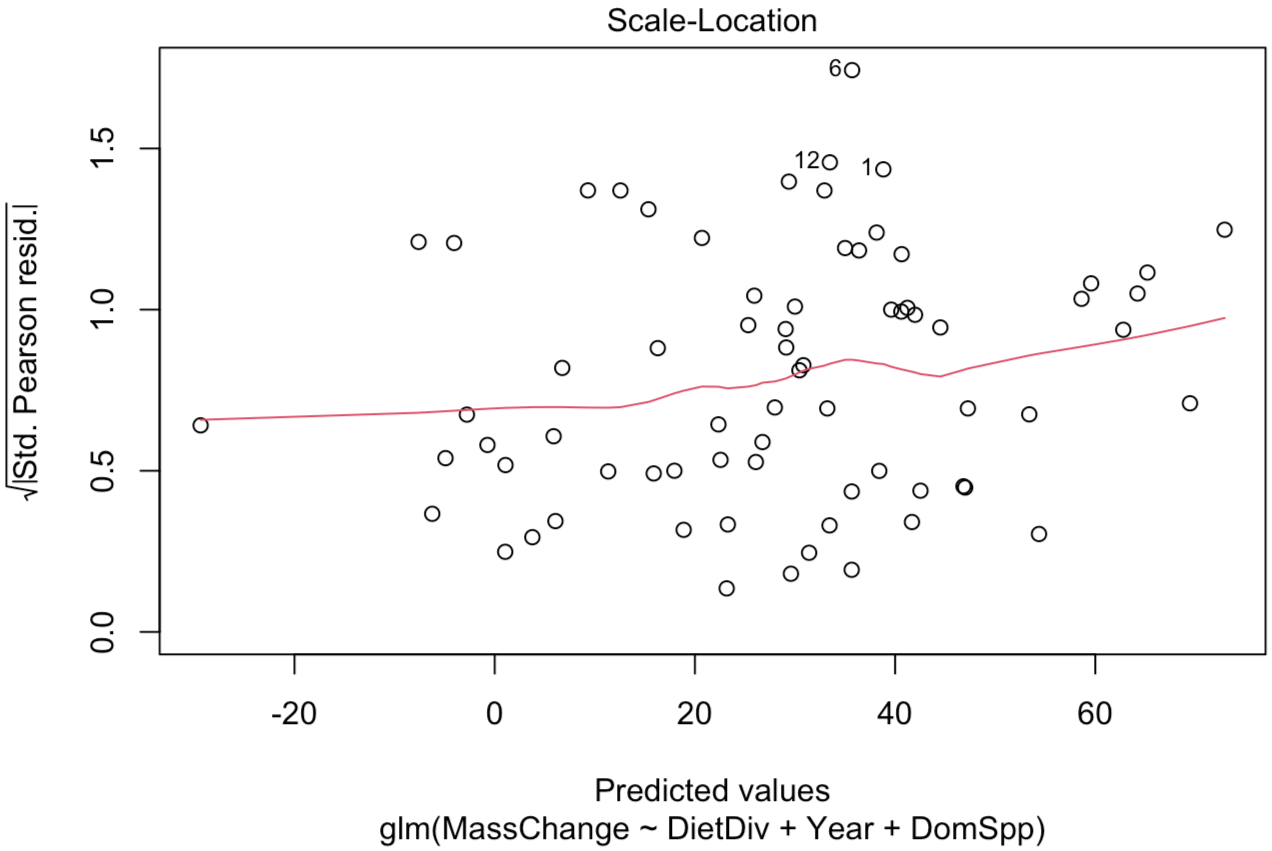
The lowest AIC model drops the dietary energy density covariate. We will update the initial model to drop the energy density covariate:

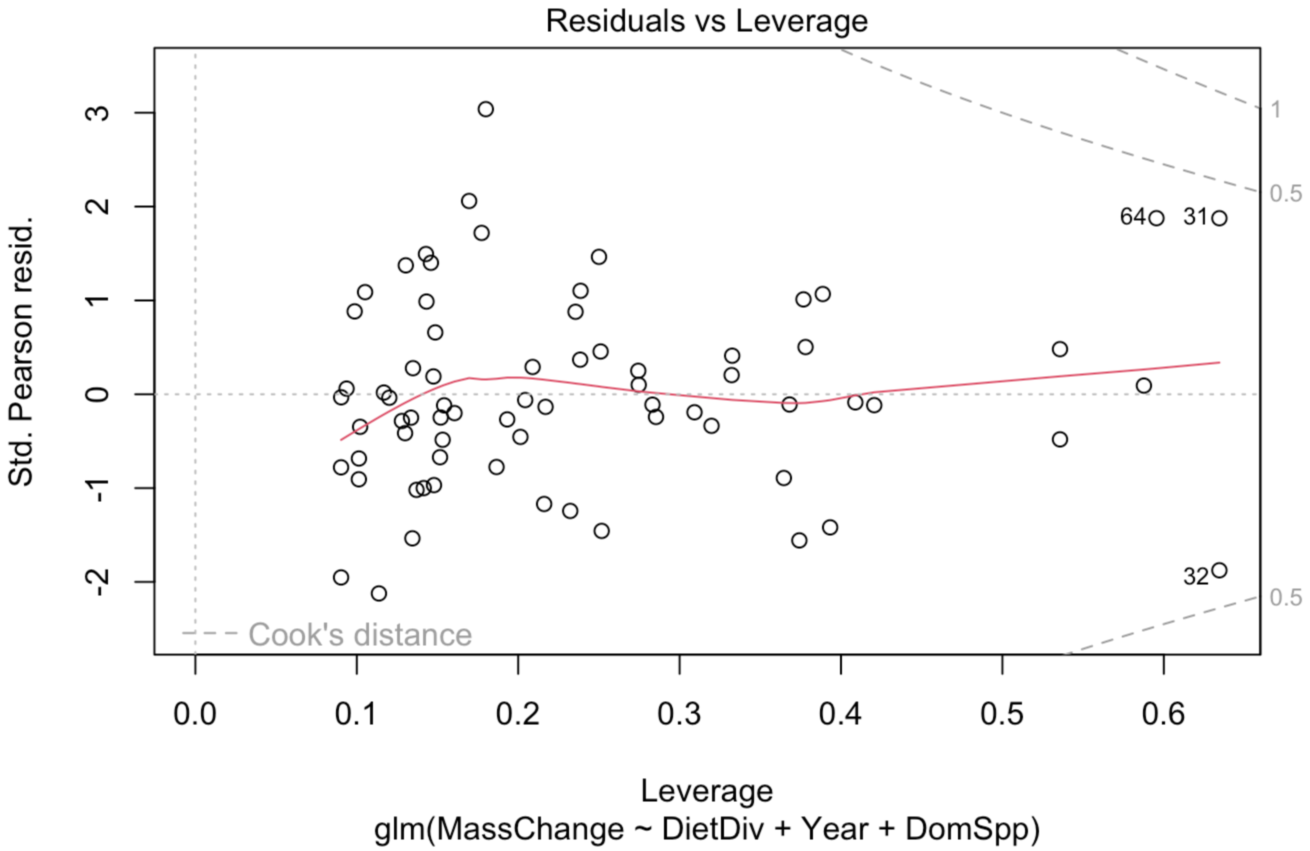
```
mom_mod2 <- glm(formula = MassChange ~ DietDiv + Year + DomSpp, family = gaussian(link =
                data = Data_Q1)
plot(mom_mod2)
```

```
Warning: not plotting observations with leverage one:
  45
```


Residuals vs Fitted
Pearson Residuals
60
40
20
0
-20
-40
-20
0
20
40
60
6
1
12
Predicted values
glm(MassChange ~ DietDiv + Year + DomSpp)


Q-Q Residuals
|Std. Deviance resid.|
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Theoretical Quantiles
glm(MassChange ~ DietDiv + Year + DomSpp)
6
1
12


Scale-Location
6
12
1
1.5
1.0
0.5
0.0
√|Std. Pearson resid.|
-20
0
20
40
60
Predicted values
glm(MassChange ~ DietDiv + Year + DomSpp)



All model residuals still look good. We can proceed to interpreting model output:

```
summary(mom_mod2)
```

```

Call:
glm(formula = MassChange ~ DietDiv + Year + DomSpp, family = gaussian(link = "identity"), 
    data = Data_Q1)

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)               70.599     16.241   4.347 6.15e-05 ***
DietDiv                  -63.415     32.417  -1.956 0.055621 .  
Year.L                    28.367     15.657   1.812 0.075584 .  
Year.Q                    12.387     12.620   0.982 0.330717    
Year.C                    18.307     13.274   1.379 0.173517    
Year^4                    15.184     14.787   1.027 0.309065    
Year^5                   -15.549     12.334  -1.261 0.212857    
Year^6                     3.305     10.924   0.303 0.763373    
Year^7                    23.361     10.525   2.219 0.030674 *  
Year^8                     2.392     10.067   0.238 0.813123    
Year^9                   -23.277     10.057  -2.314 0.024475 *  
Year^10                  -11.186     10.451  -1.070 0.289261    
Year^11                   13.261      9.842   1.347 0.183483    
```

```
Year^12                   21.328      8.874   2.403 0.019708 *  
Year^13                  -39.502      9.864  -4.005 0.000191 ***
DomSppNorthernSandlance  -18.212     14.318  -1.272 0.208824    
DomSppPollock            -47.479     16.434  -2.889 0.005550 ** 
DomSppRedfish            -20.303     13.938  -1.457 0.150992    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 361.2663)

    Null deviance: 48567  on 71  degrees of freedom
Residual deviance: 19508  on 54  degrees of freedom
AIC: 645.67

Number of Fisher Scoring iterations: 2
```

While several years have mass change values that significantly differ from the expected mass change value, the effect of time is not of direct interest to our research question and is only present to account for changes in prey dynamics over time. As a result, the interpretation of this output is not relvant. However, we see that females who feed primarily on pollock experience lower mass change than the reference level, in this case being females who primarily feed on cod. Being an identity link function, the model is essentially a simple linear model and there is no transformations applied to the response variable. As such, coefficient estimates can be interpreted directly. In this case, individuals who feed on pollock experience 47.5% lower mass change than individuals who primarily feed on cod. This effect is presented in figure 1.

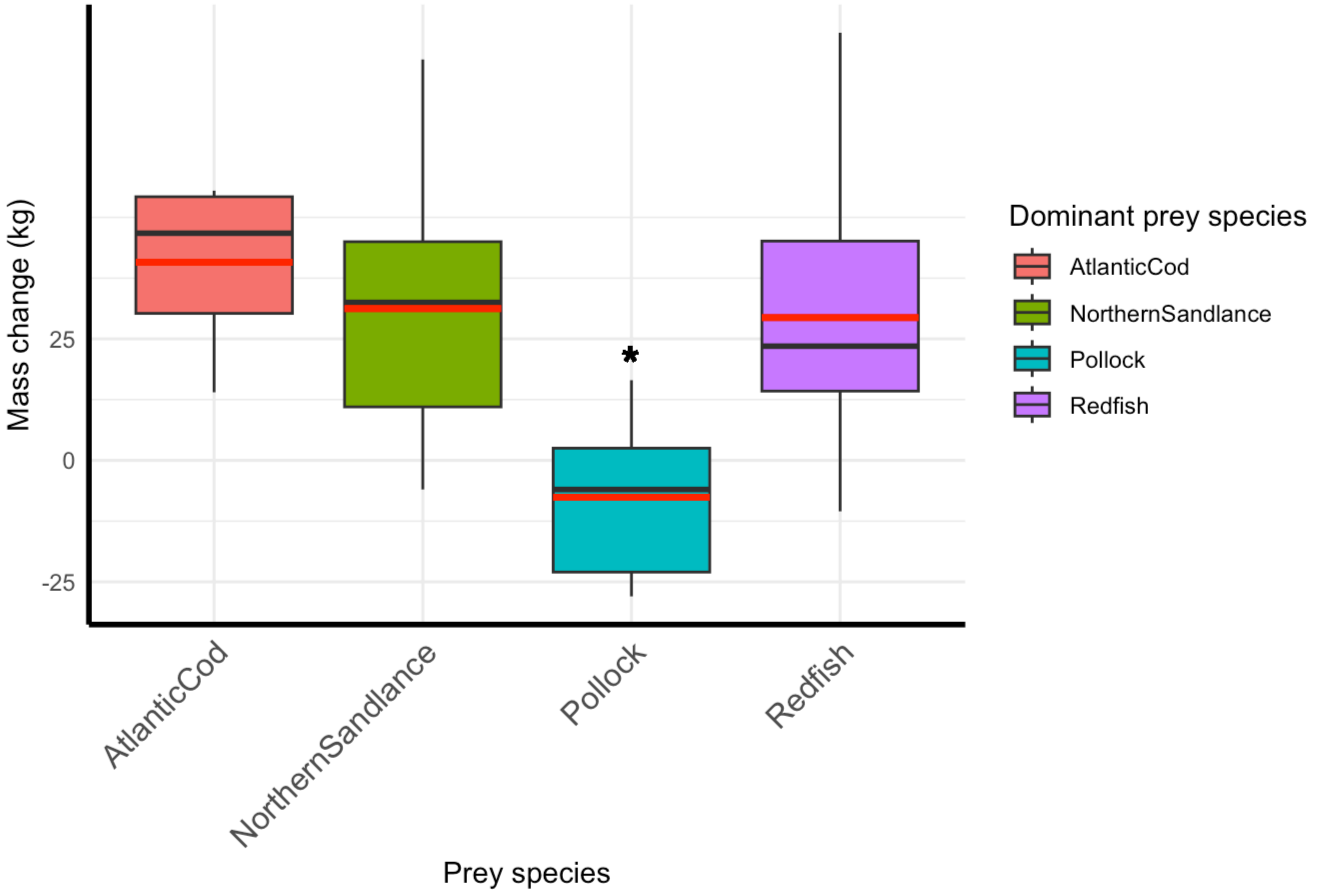
```r
require(ggplot2)
```

```
Loading required package: ggplot2
```

```r
ggplot(Data_Q1, aes(x = DomSpp, y = MassChange, fill = DomSpp)) +
  geom_boxplot() +
  stat_summary(
    fun = "mean", geom = "errorbar",
    aes(ymax = ..y.., ymin = ..y..),
    width = 0.75, color = "red", size = 1.2
  ) + geom_text(x = "Pollock", y = 20, label = "*",
                aes(x = x, y = y, label = label),
                color = "black", size = 6, fontface = "bold"
  ) +
  labs(
    title = "Figure 1. Median (black line) and mean (red line) mass change by dominant pr
    x = "Prey species",
    y = "Mass change (kg)", fill = "Dominant prey species"
  ) +
  theme_minimal() + theme(
    axis.line = element_line(color = "black", size = 1),  # Add axis lines
    axis.text.x = element_text(size = 12, angle = 45, hjust = 1)
  ) +
  scale_y_continuous(
    breaks = seq(0, max(Data_Q1$MassChange), by = 25) - 50
  )
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.

Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
ℹ Please use the `linewidth` argument instead.

Warning: The dot-dot notation (`..y..`) was deprecated in ggplot2 3.4.0.
ℹ Please use `after_stat(y)` instead.
```

Figure 1. Median (black line) and mean (red line) mass change by dominant prey



It may also be worth noting that diet diversity is almost significant at p = 0.0556. While this p-value is not significant, the negative relationship between diet diversity and mass change may support the observed effects of dominant prey species, where individuals who primarily prey on cod, sand lance, or redfish and have a lower diet diversity due to this focus on a few profitable species are more successful in gaining mass than individuals who feed on a more diverse array of prey species which include less profitable species.

We can perform cross-validation to to test this models predictions. In order to have each fold represent about 10% of the data, we will perform 7 fold cross validation:

```r
set.seed(123)

#Number of folds
k <- 7

# Split the data into folds
DatCV <- Data_Q1 %>% mutate(fold = sample(rep(1:k, length.out = n())))

# Create a vector to store RMSE for each fold
rmse_values <- numeric(k)

#Perform cross validation
for (i in 1:k) {
  train_data <- DatCV %>% filter(fold != i)
  test_data  <- DatCV %>% filter(fold == i)

  # Fit the model on training data
modCV <- glm(MassChange ~ DietDiv + Year + DomSpp, family = gaussian(link = "identity"),

  # Predict on the test data
  predictions <- predict(modCV, newdata = test_data)

  # Compute RMSE for this fold
  actuals <- test_data$MassChange
  rmse_values[i] <- sqrt(mean((predictions - actuals)^2))
}

# Summarize
rmse_values
```

```
[1] 22.44760 11.88442 18.05650 11.05129 18.98355 15.02906 14.37965
```

```r
mean_rmse <- mean(rmse_values)
cat("Average RMSE across", k, "folds:", round(mean_rmse, 2), "\n")
```

```
Average RMSE across 7 folds: 15.98
```

The average root mean squared error is 15.98, indicating the model generally predicts values within 16 kg of the true value. Check the standard deviation to compare the average RMSE to:

```r
sd(Data_Q1$MassChange)
```

```
[1] 26.15433
```

Compared to the standard deviation of mass change, the RMSE value is about 40% lower than the standard deviation, meaning the model predictions are much better than using the population mean to predict mass change. However, compared to the overall mean mass change within our dataset (28.5

kg), there is relatively large prediction error. We can also visualize prediction accuracy with a fitted vs observed value plot (figure 2):

```
fit_mom <- fitted(mom_mod2)

# Get the observed response variable (actual values)
obs_mom <- Data_Q1$MassChange

# Plot the fitted vs. observed values
ggplot(data = Data_Q1, aes(x = fit_mom, y = obs_mom)) +
  geom_point(color = "black") +
  geom_abline(slope = 1, intercept = 0, color = "red", linetype = "dashed") +  # Add 1:1
  labs(
    title = "Figure 2. Fitted vs. Observed Values with 1:1 line",
    x = "Fitted Values",
    y = "Observed Values"
  ) +
  theme_minimal()
```

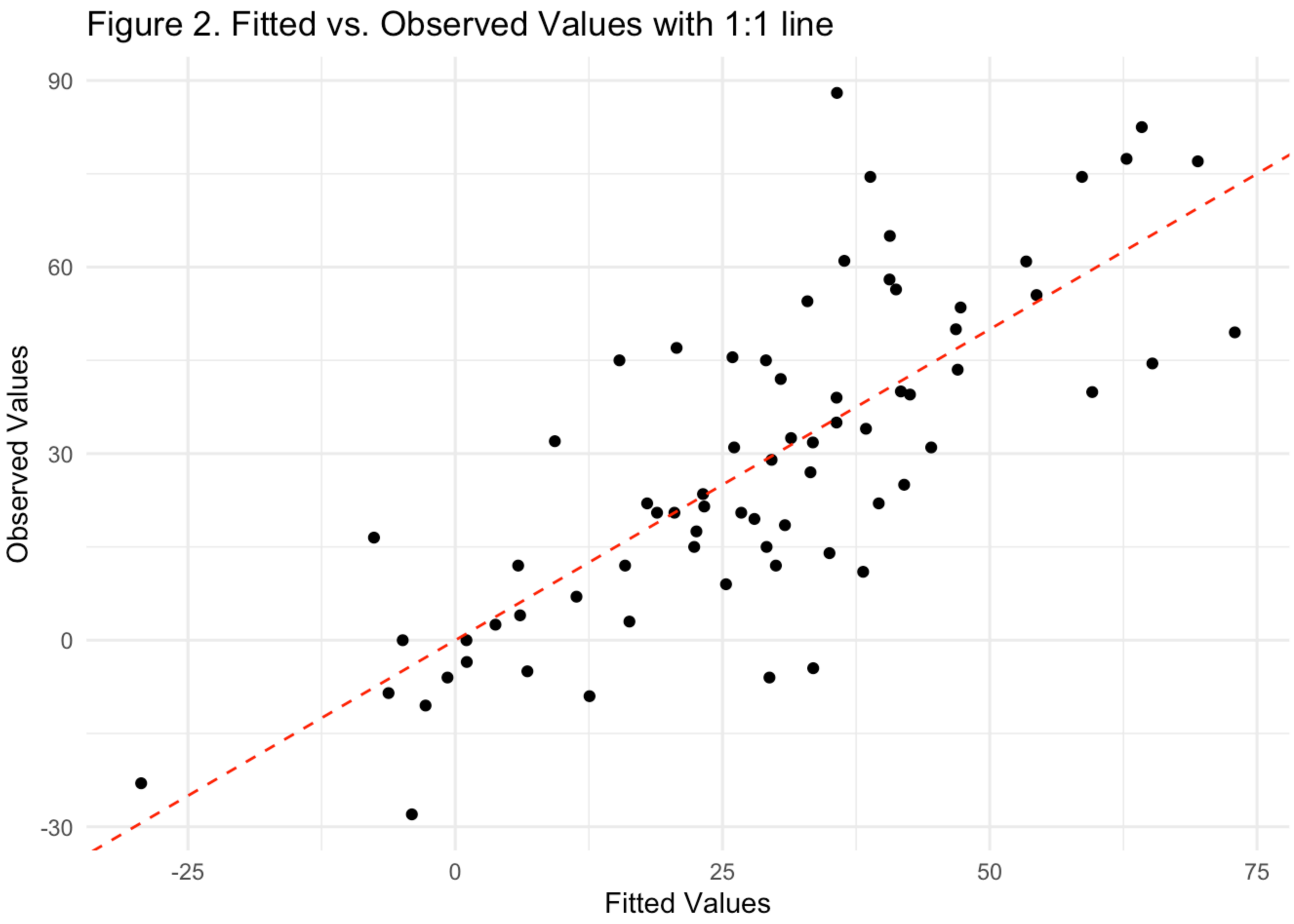


Figure 2. Fitted vs. Observed Values with 1:1 line

While figure 2 shows that the model predictions have consistent accuracy across the range of fitted values, as suggested by the average RMSE value obtained from cross-validation, there is reasonably large prediction error. Given that the model residuals suggest good model fit, the relatively large prediction error suggest the set of covariates used in this model are not sufficient in predicting maternal mass change. From this we can conclude that females who feed on redfish, sand lance, or cod experience greater mass change than individuals who feed on Pollock, but other factors outside of maternal diet, such as individual physiology or characteristics associated with individual foraging behavior, are responsible for explaining a large proportion of variation in maternal mass change.

Research Question 2

Does variation in diet of maternal female predict pup weaning mass in Northwest Atlantic grey seals?

Load Data

require(tinytex) quarto install tinytex

Cleaned data file was uploaded. seal_data = read.csv(“Data_Q2.csv”, header = T)

```
seal_data <- read.csv("/Users/peterbraithwaite/Desktop/IDPhD Classes/Stat 5620_Updated/Fi
summary(seal_data)
```

```
     MomID            Year      Dietary.energy.density Diet.diversity
 Min.   :   19   Min.   :1996   Min.   :5.207          Min.   :0.1788
 1st Qu.:  467   1st Qu.:2002   1st Qu.:5.676          1st Qu.:0.3108
 Median : 4688   Median :2011   Median :5.784          Median :0.3564
 Mean   : 4999   Mean   :2008   Mean   :5.772          Mean   :0.3622
 3rd Qu.: 9412   3rd Qu.:2013   3rd Qu.:5.900          3rd Qu.:0.4181
 Max.   :10690   Max.   :2015   Max.   :6.517          Max.   :0.5290
 Dominant.prey.species    Mom.Age         Pup.sex      Pup.Wean.Mass
 Length:56              Min.   : 9.00   Min.   :1.000   Min.   :34.50
 Class :character       1st Qu.:23.00   1st Qu.:1.000   1st Qu.:45.50
 Mode  :character       Median :25.00   Median :1.000   Median :50.75
                        Mean   :23.46   Mean   :1.464   Mean   :50.31
                        3rd Qu.:26.25   3rd Qu.:2.000   3rd Qu.:55.12
                        Max.   :31.00   Max.   :2.000   Max.   :63.00
```

Data Exploration

The package “flexplot” was used to explore the predictor and response variables.

```
require(flexplot)
```

```
Loading required package: flexplot
```

```
Attaching package: 'flexplot'

The following object is masked from 'package:ggplot2':

    flip_data
```

```
a = flexplot (MomID~1, data= seal_data)
b = flexplot (Year~1, data= seal_data)
c = flexplot (Dietary.energy.density~1, data= seal_data)
d = flexplot (Diet.diversity~1, data= seal_data)
e = flexplot (Dominant.prey.species~1, data= seal_data)
f = flexplot (Mom.Age~1, data= seal_data)
g = flexplot (Pup.sex~1, data= seal_data)
h = flexplot (Pup.Wean.Mass~1, data= seal_data)

require (cowplot)
```

```
Loading required package: cowplot
```

```
plot_grid(a,b,c,d,e,f,g,h)
```

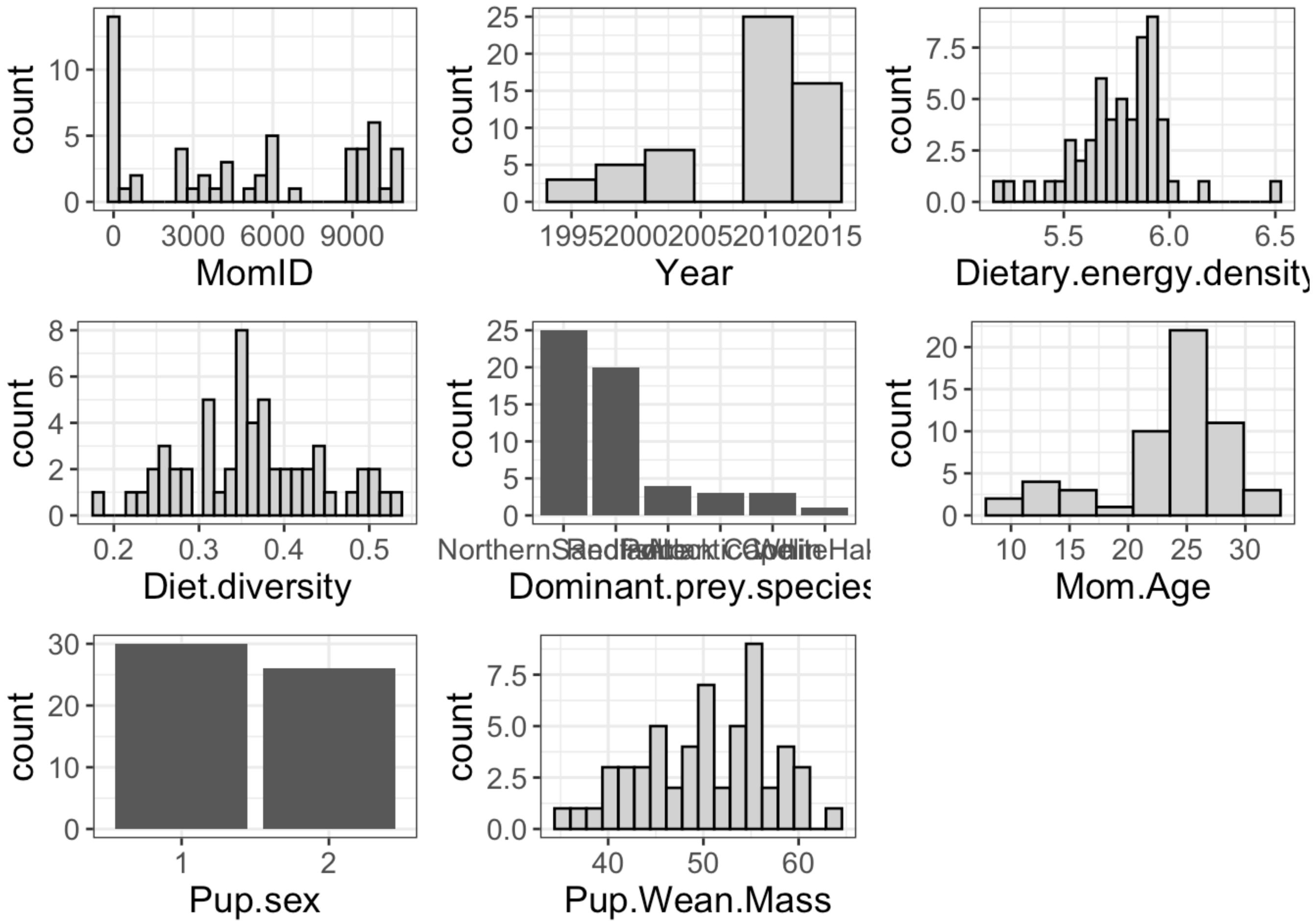


Figure 1 - Exploring Variable Distributions

From the plots above it seems that the response variable (Pup Wean Mass) has a Gaussian distribution.

To further explore the relationship between variable the package “ggplot2” was used plot individual variables in relation to the repsonse variables.

Pup Wean Mass and Dietary Engery Density

```
require(ggplot2)
# Pup Wean Mass + Dietary.energy.density
ggplot(seal_data) + geom_point(aes(Dietary.energy.density, Pup.Wean.Mass, color = Dominan
```

```
`geom_smooth()` using formula = 'y ~ x'
```

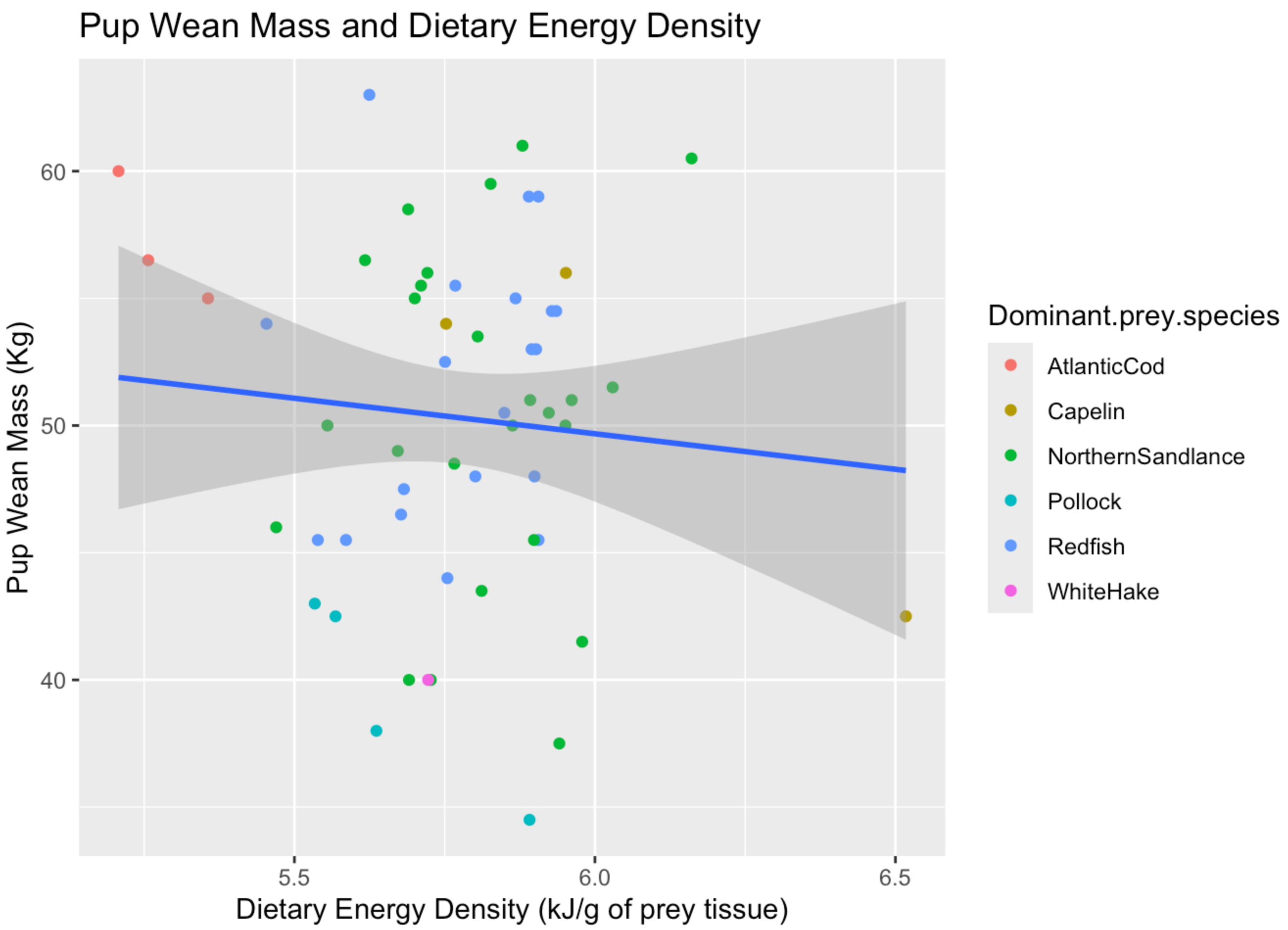


Figure 2 - Pup Wean Mass and Dietary Energy Density

From this plot we can see a negative or inverse relationship between Pup Wean Mass and Dietary Energy Density. This indicates that maternal mothers that feed on prey species resulting in greater dietary energy density tend to have pups with a lower wean mass. This is somewhat at odds with what might have been expected. We might have expected the greater the dietary energy density of the materanl mother would result in heavier and ultimately more fit pups.

We can also see that the Dominant Species Capelin has only three observation and one of these appears to be an outlier. It should also be noted that White Hake has only one observation and Atlantic Cod has only three. This lack of observations may be problematic while fitting a model and we may need to remove these Dominant Prey Species.

Pup Wean Mass and Dietary Engery Density

```
# Pup Wean Mass + Diet Diveristy
ggplot(seal_data) + geom_point(aes(Diet.diversity, Pup.Wean.Mass, color = Dominant.prey.s
  labs(title = "Pup Wean Mass and Diet Diversity", x = "Diet Diveristy (no specific units
  geom_smooth(aes(Diet.diversity, Pup.Wean.Mass), method="lm", se=T)
```

```
`geom_smooth()` using formula = 'y ~ x'
```

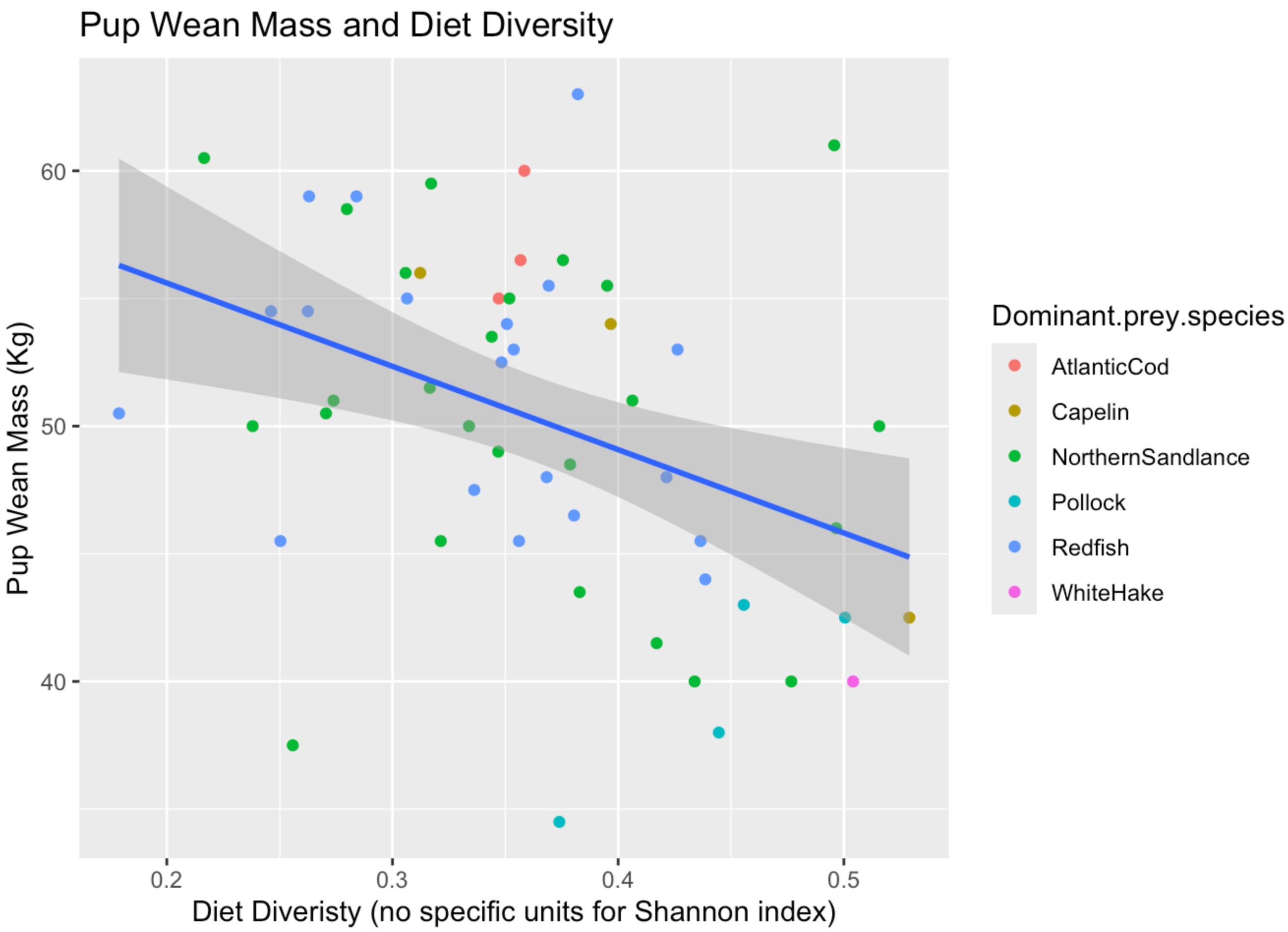


Figure 3 - Pup Wean Mass and Diet Diversity

From this plot we can see a negative or inverse relationship between Pup Wean Mass and Diet Diversity. This indicates that as the maternal mothers feed on more diverse prey species their pup wean mass decreases. This may indicates that mothers that reduce the diversity of their diet have healthier and more fit pups.

Pup Wean Mass + Year

Exploring if specific years showed abnormally high or low Pup Wean Mass and if these were related to specific dominant prey species.

```
ggplot(seal_data) + geom_point(aes(Year, Pup.Wean.Mass, color = Dominant.prey.species)) +
  labs(title = "Pup Wean Mass and Year", x = "Year", y = "Pup Wean Mass (Kg)") +
  geom_smooth (aes(Year, Pup.Wean.Mass), method="lm", se=T)
```

```
`geom_smooth()` using formula = 'y ~ x'
```

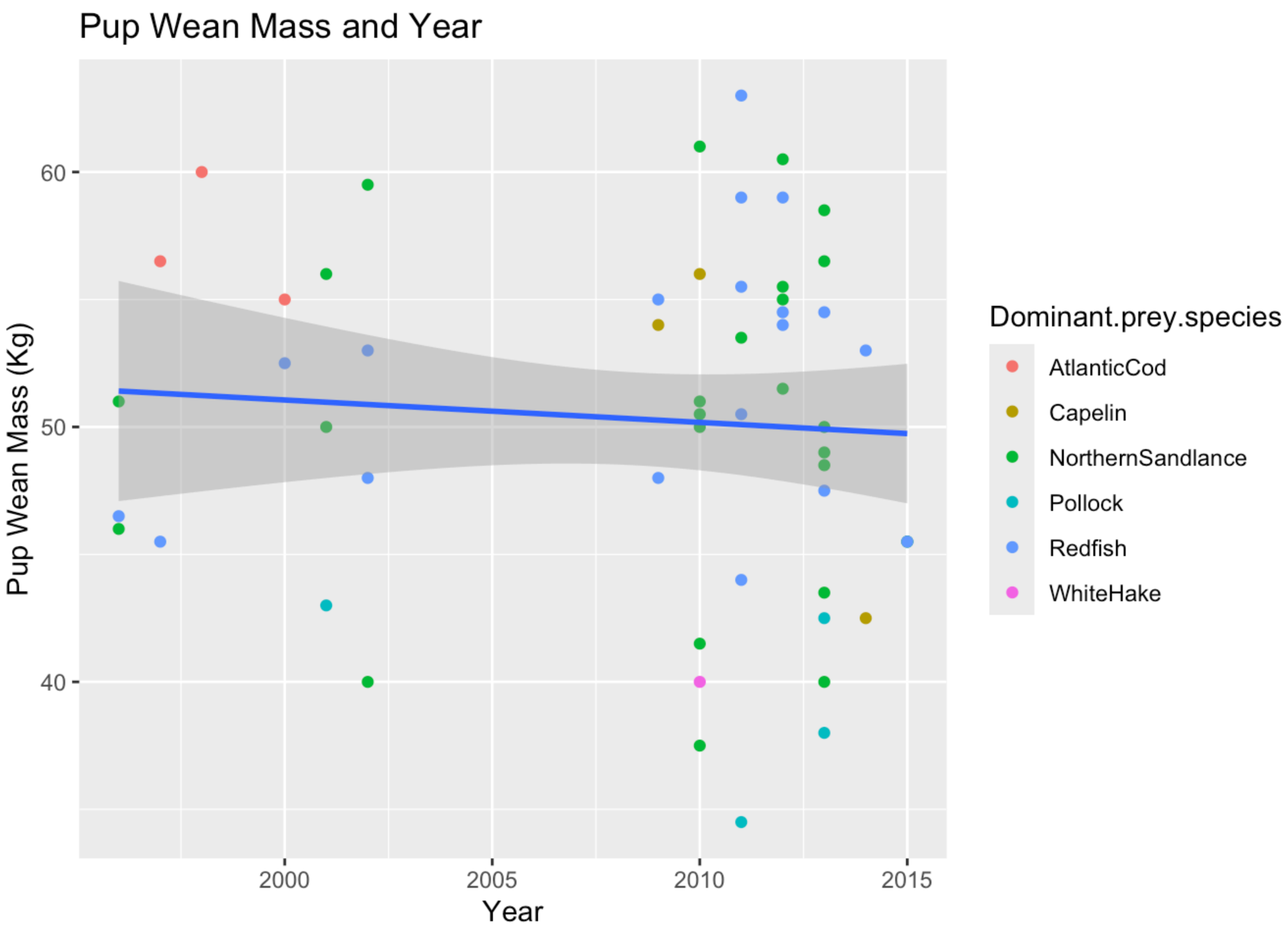


Figure 4 - Pup Wean Mass and Year

From this plot it seems Pup Wean Mass seems to be fairly evenly distributed over the observed years and it does not seem that a specific Dominant Prey Species had a disproportionate effect on the Pup Wean Mass.

Pup Wean Mass + MomID

Exploring if Pup Wean Mass shows significant differences based on the maternal mother ID.

```
ggplot(seal_data) + geom_point(aes(MomID, Pup.Wean.Mass, color = Year )) +
  labs(title = "Pup Wean Mass and Mom ID", x = "Mom ID", y = "Pup Wean Mass (Kg)") +
geom_smooth(aes(MomID, Pup.Wean.Mass), method="lm", se=T)
```

```
`geom_smooth()` using formula = 'y ~ x'
```

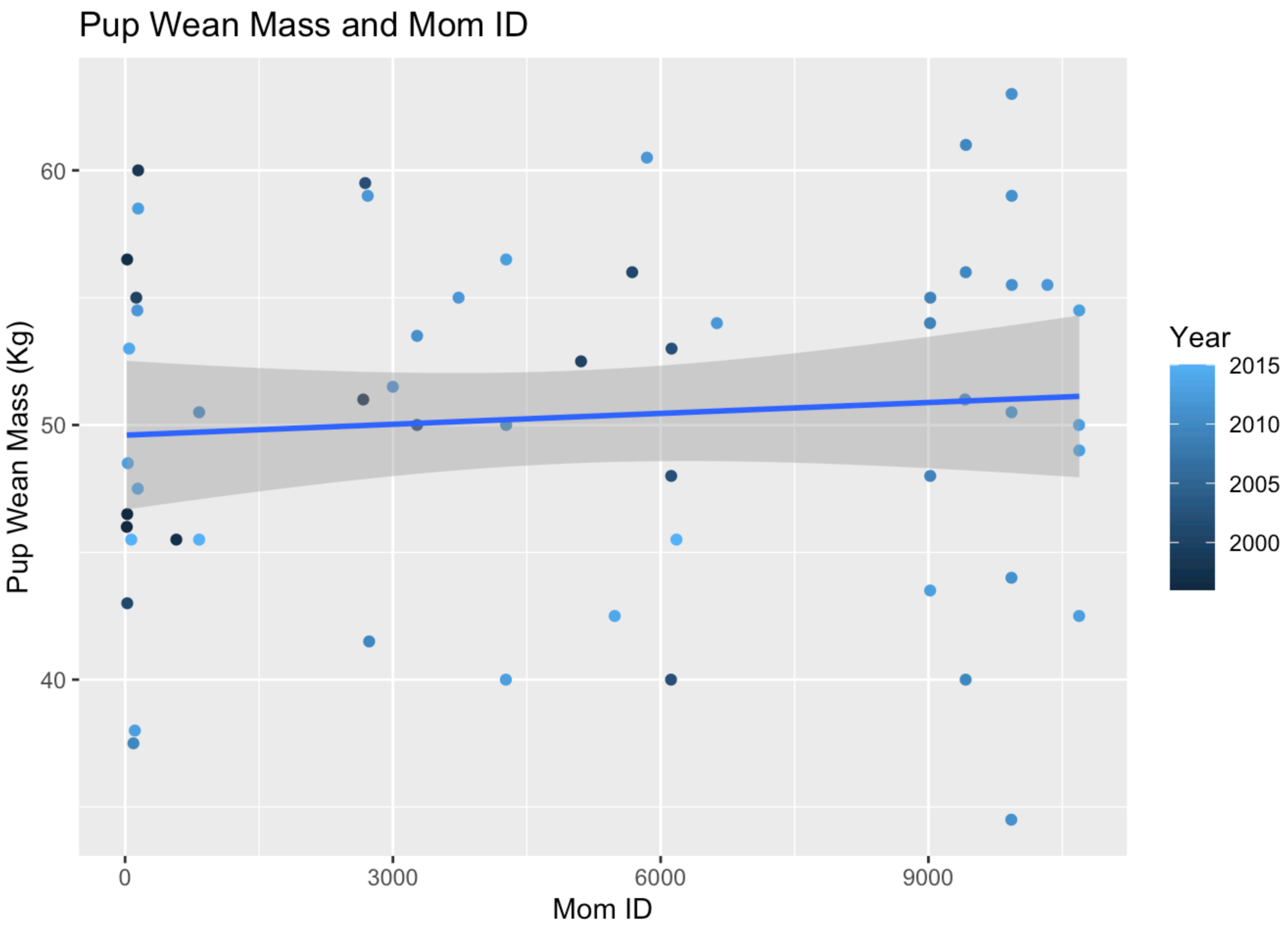


Figure 5 - Pup Wean Mass and Materianl Mother ID

With the vertical rows of the this plot representing unique maternal mothers, it does not seem any individual is creating disproportionately large or small pup weights. Rather, it seems most maternal mother are creating a fairly evenly distribution of pup weights dependent on the year.

Pup Wean Mass + Dominant Prey Species

```
ggplot(seal_data) + geom_point(aes(Dominant.prey.species, Pup.Wean.Mass, color = Year ))
```

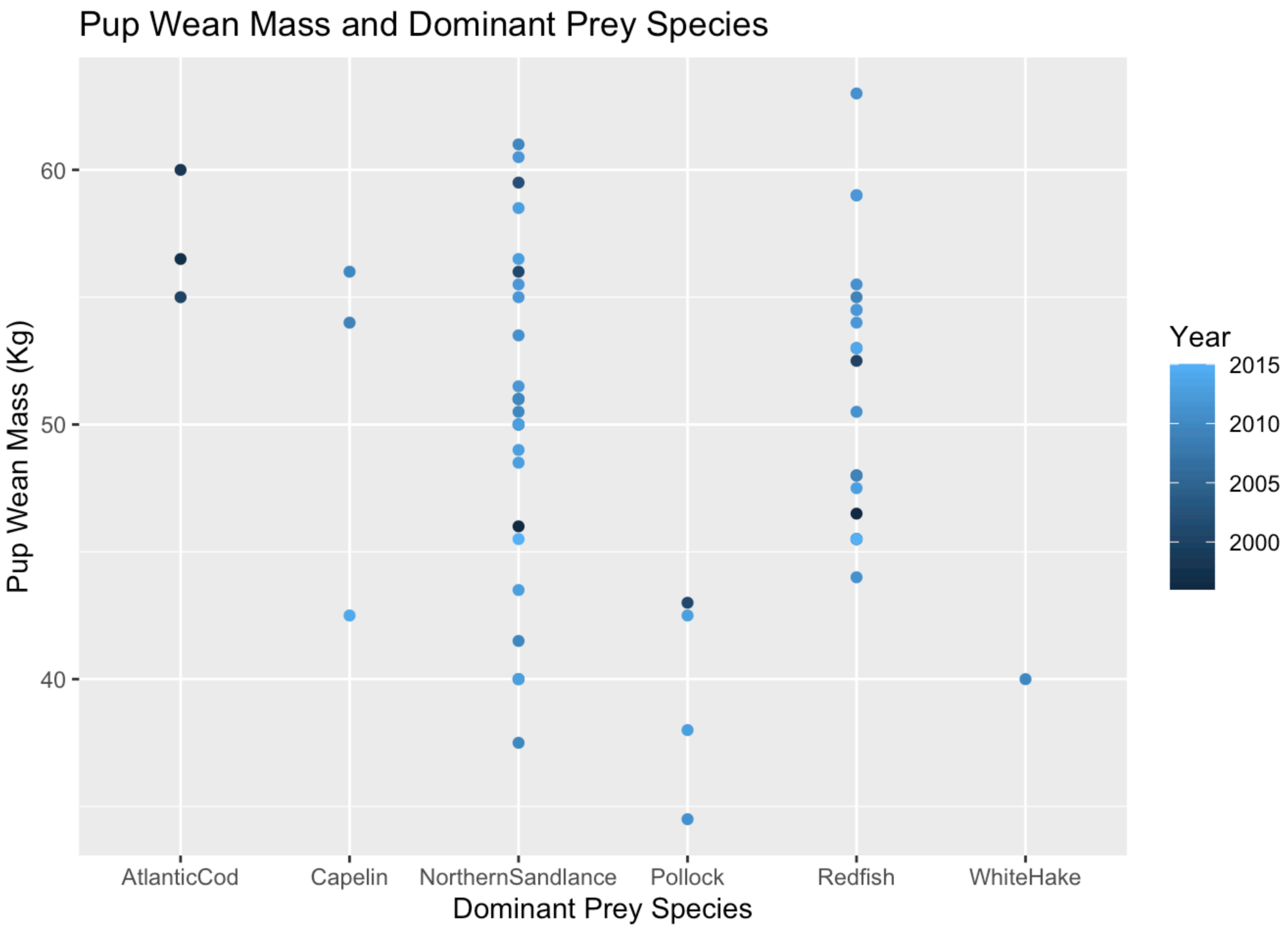


Figure 6 - Pup Wean Mass and Dominant Prey Species

A considerable amount of insight can be gained from this plot including: The Dominant Prey Species with the largest number of observations (Northern Sandlance and Redfish) show a variety of Pup Wean Masses and these occur over a wide variety of years; the Dominant Prey Species with the smallest number of observations tend to show either predominantly large (Atlantic Cod and Capelin) or very small (Pollock and White Hake) Pup Wean sizes and these tend to occur in a variety of years.

Pup Wean Mass + Pup Sex

```
ggplot(seal_data) + geom_point(aes(Pup.sex, Pup.Wean.Mass, color = Year )) +
  labs(title = "Pup Wean Mass and Pup Sex", x = "Pup Sex", y = "Pup Wean Mass (Kg)") +
  geom_smooth(aes(Pup.sex, Pup.Wean.Mass), method="lm", se=T)
```

```
`geom_smooth()` using formula = 'y ~ x'
```

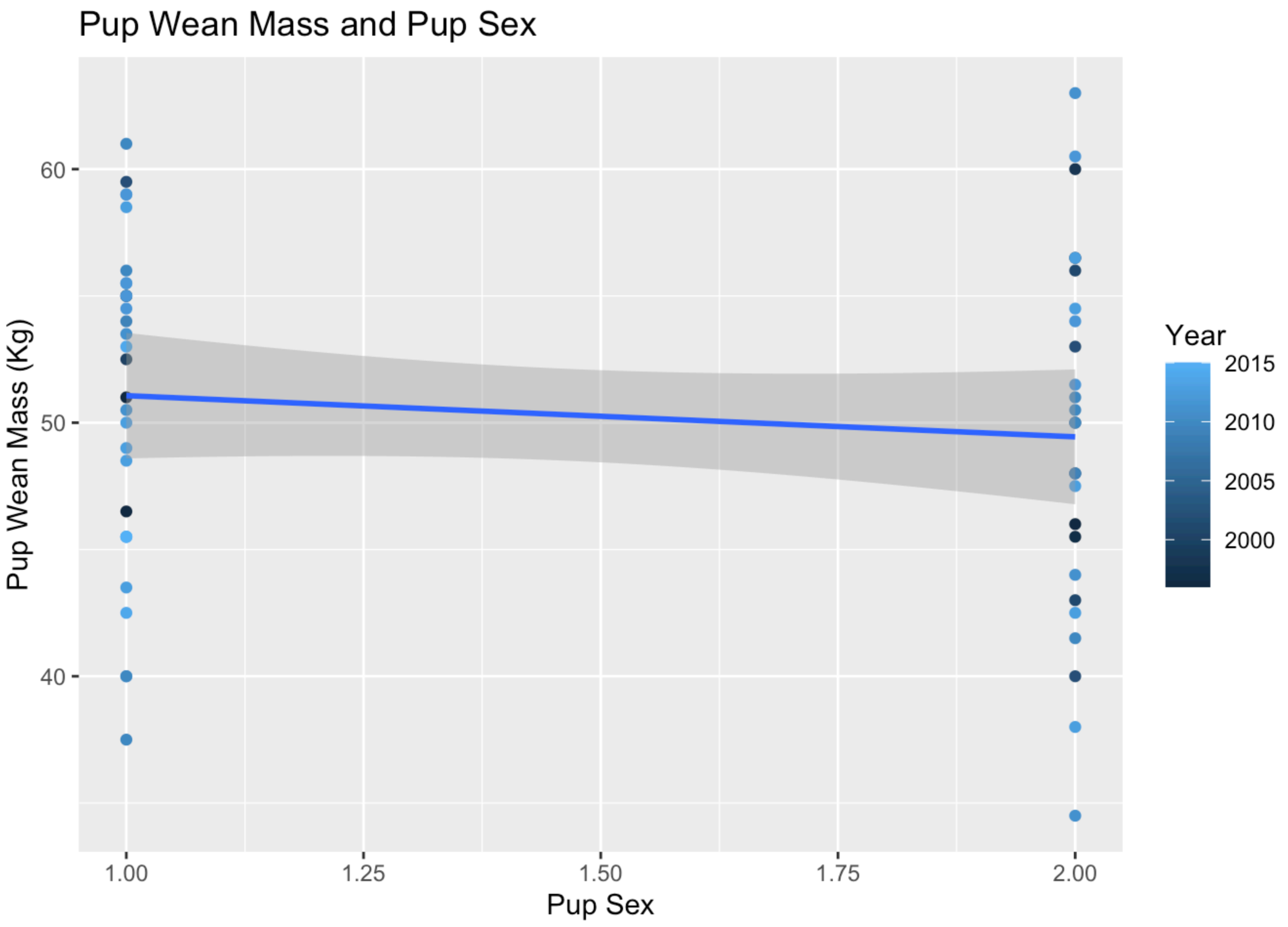


Figure 7 - Pup Wean Mass and Pup Sex

This plot shows that the Pup Sex seems to have little effect on the average Pup Mean Mass. Both Pup Sex seem to create a fairly even distribution of Pup Mean Mass over a variety of years. It is expected that this variable will have a small effect on the fit model.

Pup Wean Mass + Maternal Age

```
ggplot(seal_data) + geom_point(aes(Mom.Age, Pup.Wean.Mass, colour = Dominant.prey.species
  labs(title = "Pup Wean Mass and Maternal Age", x = "Maternal Age", y = "Pup Wean Mass (
```

```
`geom_smooth()` using formula = 'y ~ x'
```

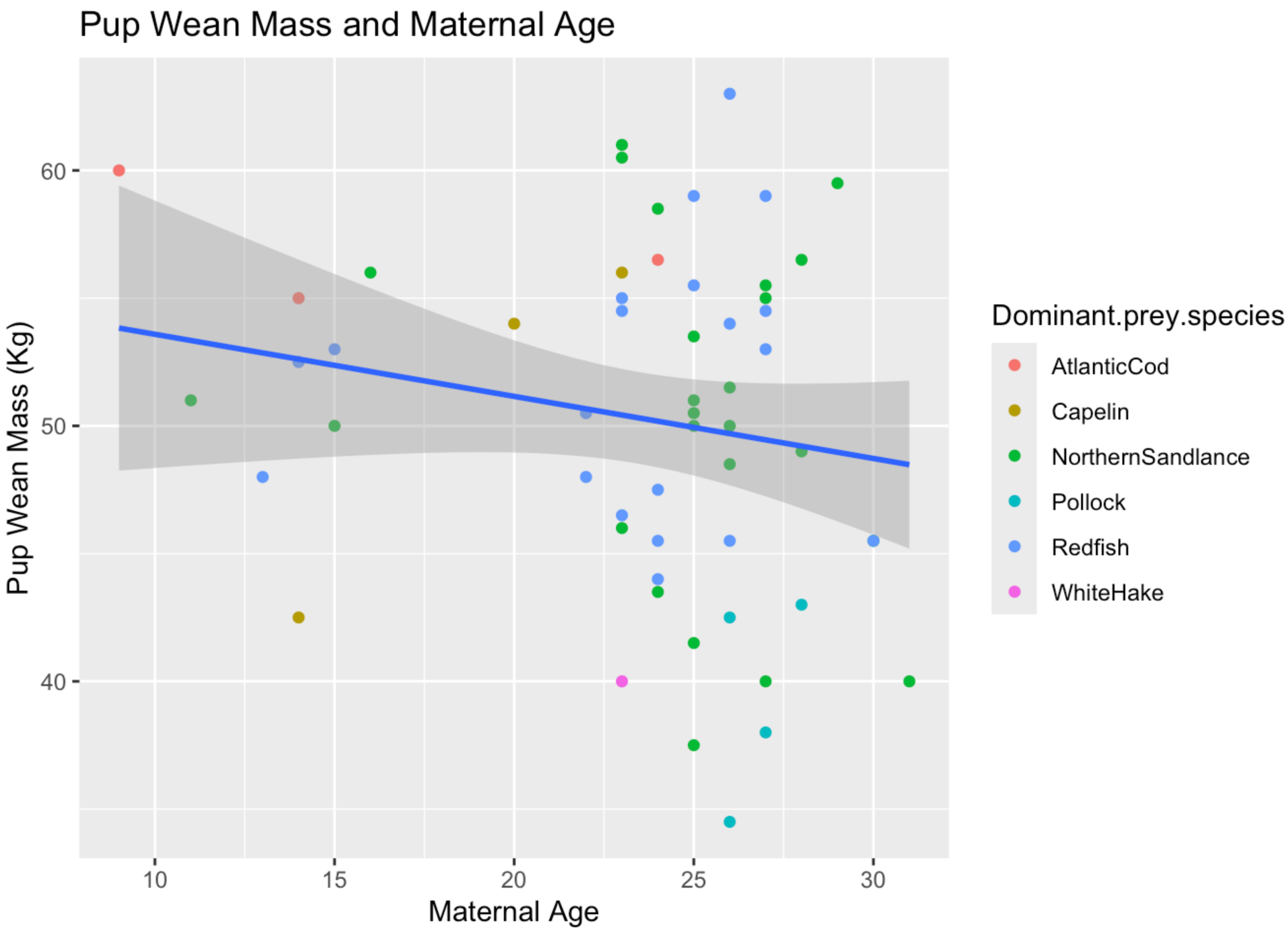


Figure 7 - Pup Wean Mass and Maternal Age

A considerable amount of insight can be gained from this plot. It seems younger maternal seals have larger pups but it is also noteworthy that the majority of older mothers prey on the most observed Dominant Prey Species including Redfish and Northern Sandlance. Alternatively it seems less frequent observed Dominant Prey Species were preyed on by young maternal mothers. It seems as though there might be a social aspects to why younger mothers prey on less frequently observed Dominant Prey Species.

Fit Linear Model

We now explore fitting general linear models to further explore the variables, the residuals and their relationships.

First we convert the catagorical data to factors and determine the number of levels within each catagorical group.

```
# convert catagorical data to factors

seal_data$MomID = as.factor(seal_data$MomID)
seal_data$Year = as.factor(seal_data$Year)
```

```
seal_data$Dominant.prey.species = as.factor(seal_data$Dominant.prey.species)
seal_data$Pup.sex = as.factor(seal_data$Pup.sex)
seal_data$Pup.Mom.Age = as.factor(seal_data$Mom.Age)

nlevels(seal_data$MomID)
```

```
[1] 51
```

```
# 51
nlevels(seal_data$Year)
```

```
[1] 13
```

```
# 13
nlevels(seal_data$Dominant.prey.species)
```

```
[1] 6
```

```
# 5
nlevels(seal_data$Pup.sex)
```

```
[1] 2
```

```
# 2
nlevels(seal_data$Pup.Mom.Age)
```

```
[1] 17
```

```
# 17
```

Fit Linear Model

All Predictor Variables

```
Q1 = lm(Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.density
        + Year + Mom.Age + Pup.sex + MomID, data = seal_data)
summary(Q1)
```

```

Call:
lm(formula = Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity +
    Dietary.energy.density + Year + Mom.Age + Pup.sex + MomID,
    data = seal_data)

Residuals:
```

```
ALL 56 residuals are 0: no residual degrees of freedom!

Coefficients: (16 not defined because of singularities)
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            -127.5808        NaN     NaN      NaN
Dominant.prey.speciesCapelin             14.0800        NaN     NaN      NaN
Dominant.prey.speciesNorthernSandlance    2.5039        NaN     NaN      NaN
Dominant.prey.speciesPollock            -19.7677        NaN     NaN      NaN
Dominant.prey.speciesRedfish            -16.6515        NaN     NaN      NaN
Dominant.prey.speciesWhiteHake            4.6941        NaN     NaN      NaN
Diet.diversity                            8.3424        NaN     NaN      NaN
Dietary.energy.density                   35.8485        NaN     NaN      NaN
Year1997                                 19.2395        NaN     NaN      NaN
Year1998                                 -1.3181        NaN     NaN      NaN
Year2000                                 14.1916        NaN     NaN      NaN
Year2001                                 -0.8124        NaN     NaN      NaN
Year2002                                 -3.0383        NaN     NaN      NaN
Year2009                                 29.8803        NaN     NaN      NaN
Year2010                                 -2.9017        NaN     NaN      NaN
Year2011                                 37.2376        NaN     NaN      NaN
Year2012                                 23.6996        NaN     NaN      NaN
Year2013                                 20.8844        NaN     NaN      NaN
Year2014                                -55.5366        NaN     NaN      NaN
Year2015                                 16.4409        NaN     NaN      NaN
Mom.Age                                  -1.8940        NaN     NaN      NaN
Pup.sex2                                 14.4153        NaN     NaN      NaN
MomID23                                   4.4626        NaN     NaN      NaN
MomID24                                  27.5950        NaN     NaN      NaN
MomID30                                  -7.9254        NaN     NaN      NaN
MomID45                                  88.7703        NaN     NaN      NaN
MomID69                                  26.2285        NaN     NaN      NaN
MomID93                                  -2.2734        NaN     NaN      NaN
MomID109                                 -4.5789        NaN     NaN      NaN
MomID125                                      NA         NA      NA       NA
MomID137                                  3.6242        NaN     NaN      NaN
MomID142                                 -4.6121        NaN     NaN      NaN
MomID146                                  1.8643        NaN     NaN      NaN
MomID574                                      NA         NA      NA       NA
MomID829                                 -3.1820        NaN     NaN      NaN
MomID2668                               -19.0778        NaN     NaN      NaN
MomID2690                                31.0252        NaN     NaN      NaN
MomID2718                                16.5836        NaN     NaN      NaN
MomID2734                               -15.3957        NaN     NaN      NaN
MomID2999                               -31.0853        NaN     NaN      NaN
MomID3271                               -22.2778        NaN     NaN      NaN
MomID3736                                 0.2347        NaN     NaN      NaN
MomID4266                               -12.6464        NaN     NaN      NaN
MomID4269                                -5.2029        NaN     NaN      NaN
MomID5108                                     NA         NA      NA       NA
MomID5485                                     NA         NA      NA       NA
MomID5681                                -9.8719        NaN     NaN      NaN
```

```
MomID5846                                      -31.6400        NaN     NaN       NaN
MomID6116                                        3.4943        NaN     NaN       NaN
MomID6118                                       -9.0816        NaN     NaN       NaN
MomID6122                                            NA         NA      NA        NA
MomID6177                                        8.7215        NaN     NaN       NaN
MomID6630                                       10.9350        NaN     NaN       NaN
MomID9018                                      -34.0206        NaN     NaN       NaN
MomID9019                                      -18.3770        NaN     NaN       NaN
MomID9020                                      -21.8694        NaN     NaN       NaN
MomID9021                                            NA         NA      NA        NA
MomID9410                                       -2.7034        NaN     NaN       NaN
MomID9417                                            NA         NA      NA        NA
MomID9418                                            NA         NA      NA        NA
MomID9420                                       17.6372        NaN     NaN       NaN
MomID9928                                      -34.8701        NaN     NaN       NaN
MomID9930                                      -27.9118        NaN     NaN       NaN
MomID9931                                      -12.0164        NaN     NaN       NaN
MomID9932                                            NA         NA      NA        NA
MomID9933                                            NA         NA      NA        NA
MomID9934                                            NA         NA      NA        NA
MomID10333                                           NA         NA      NA        NA
MomID10687                                           NA         NA      NA        NA
MomID10688                                           NA         NA      NA        NA
MomID10689                                           NA         NA      NA        NA
MomID10690                                           NA         NA      NA        NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:    NaN 
F-statistic:   NaN on 55 and 0 DF,  p-value: NA
```

When this model is run we received an error suggesting we have more columns than rows. This suggest we have more variables than observations. As a result we must remove some of the categorical variables. The maternal mother ID seems to be adding a lot of categories (n -51) and did not seem very important in Figure 6 above. As a result Mom ID was removed and the model was fit again.

All Predictor Variables Except Mom ID

```
Q2 = lm(Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.density +
summary
```

```
standardGeneric for "summary" defined from package "base"

function (object, ...) 
standardGeneric("summary")
<environment: 0x1249eff90>
Methods may be defined for arguments: object
Use  showMethods(summary)  for currently available ones.
```

```
# Adjusted R-squared:  0.209
# p-value: 0.08393
```

Although this model was not significant, it did provide a model that works as a starting point for the Step Function.

Step Function (Forward)

We now used the Step Function to determine the best fit model using AIC.

```
fwd.model = step (Q2, direction='forward')
```

```
Start:  AIC=216.76
Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.density +
    Year + Mom.Age + Pup.sex
```

The forward step retained all variables with AIC = 216.76

Step Function (Backward)

```
backward.model = step(Q2, direction='backward')
```

```
Start:  AIC=216.76
Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.density +
    Year + Mom.Age + Pup.sex

                         Df Sum of Sq    RSS    AIC
- Year                   12    347.70 1572.2 206.75
- Mom.Age                 1      1.61 1226.1 214.83
- Pup.sex                 1     17.14 1241.6 215.53
<none>                                1224.5 216.76
- Dietary.energy.density  1     47.89 1272.4 216.91
- Diet.diversity          1     69.12 1293.6 217.83
- Dominant.prey.species   5    329.04 1553.5 220.08

Step:  AIC=206.75
Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.density +
    Mom.Age + Pup.sex

                         Df Sum of Sq    RSS    AIC
- Mom.Age                 1      1.51 1573.7 204.81
- Pup.sex                 1      3.13 1575.3 204.87
- Dietary.energy.density  1     50.68 1622.9 206.53
<none>                                1572.2 206.75
- Dominant.prey.species   5    364.88 1937.1 208.44
- Diet.diversity          1    162.65 1734.8 210.27

Step:  AIC=204.81
Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.density +
    Pup.sex
```

```
                         Df Sum of Sq    RSS    AIC
- Pup.sex                 1      2.57 1576.3 202.90
- Dietary.energy.density  1     49.35 1623.0 204.54
<none>                                1573.7 204.81
- Dominant.prey.species   5    415.52 1989.2 207.93
- Diet.diversity          1    164.37 1738.1 208.37

Step:  AIC=202.9
Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.density

                         Df Sum of Sq    RSS    AIC
- Dietary.energy.density  1     49.16 1625.4 202.62
<none>                                1576.3 202.90
- Diet.diversity          1    166.48 1742.8 206.52
- Dominant.prey.species   5    456.31 2032.6 207.14

Step:  AIC=202.62
Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity

                        Df Sum of Sq    RSS    AIC
<none>                               1625.4 202.62
- Diet.diversity         1    124.06 1749.5 204.74
- Dominant.prey.species  5    493.24 2118.7 207.46
```

The backward step retained only the Dominant Prey Species and Diet Diversity variables with AIC 202.62.

Explore The Resulting Linear Model From Backward Step Function

```
Q1_Reduced = lm(Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity,
                data = seal_data)
summary(Q1_Reduced)
```

```

Call:
lm(formula = Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity,
    data = seal_data)

Residuals:
     Min       1Q   Median       3Q      Max
-15.0508  -4.1950   0.0968   3.2694  13.3653

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                              64.422      5.013  12.850  < 2e-16 ***
Dominant.prey.speciesCapelin             -5.133      4.743  -1.082 0.284509
Dominant.prey.speciesNorthernSandlance   -6.630      3.519  -1.884 0.065539 .
Dominant.prey.speciesPollock            -15.831      4.500  -3.518 0.000949 ***
Dominant.prey.speciesRedfish             -5.797      3.570  -1.624 0.110838
Dominant.prey.speciesWhiteHake          -14.093      6.838  -2.061 0.044627 *
```

```
Diet.diversity                               -20.491     10.596  -1.934 0.058916 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.76 on 49 degrees of freedom
Multiple R-squared:  0.3509,	Adjusted R-squared:  0.2715 
F-statistic: 4.416 on 6 and 49 DF,  p-value: 0.00121
```

```
# Adjusted R-squared: 0.2715
# p-value: 0.00121
```

This found that the redueced Linear Model is significant with a p-value = 0.00121 and an Adjusted R-Squared = 0.2715 meaning the model accounts for about 27% of the variation of the Pup Wean Mass. It should also be noted that the most significant variables are Pollock with a p-value of 0.000949 and White Hake with a p-value = 0.044627.

We will now look at the residuals for the this model.

Explore Residuals for Reduced Linear Model

```
library(ggfortify)
autoplot(Q1_Reduced)
```

```
Warning: Removed 5 rows containing missing values or values outside the scale range
(`geom_line()`).

Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_point()`).

Warning: Removed 1 row containing missing values or values outside the scale range
(`geom_line()`).
```

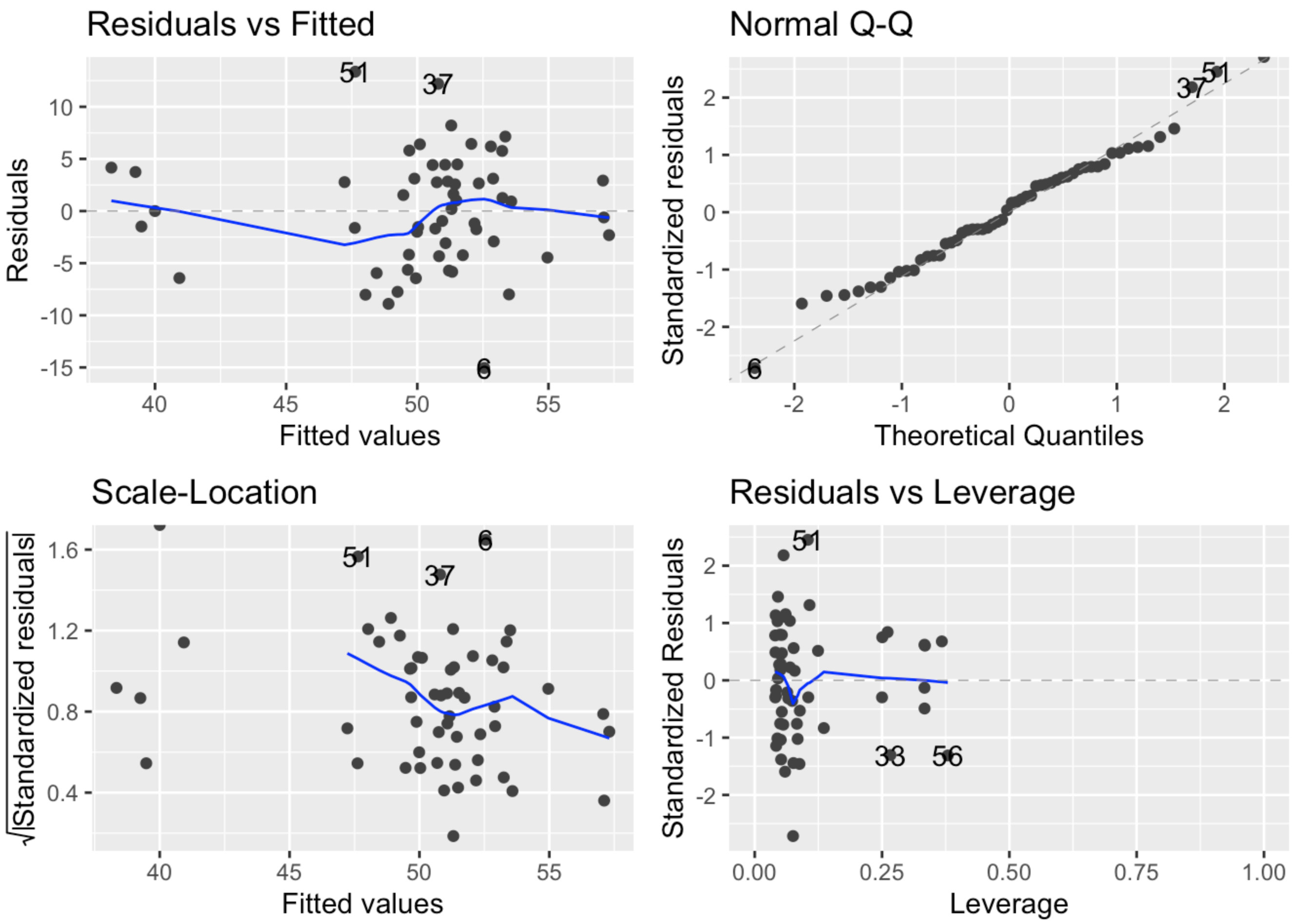


Figure 8 - Residual for Reduced Variables Linear Model

```
#residuals
res = resid(Q1_Reduced)
plot(fitted(Q1_Reduced), res)
abline(0,0)
```

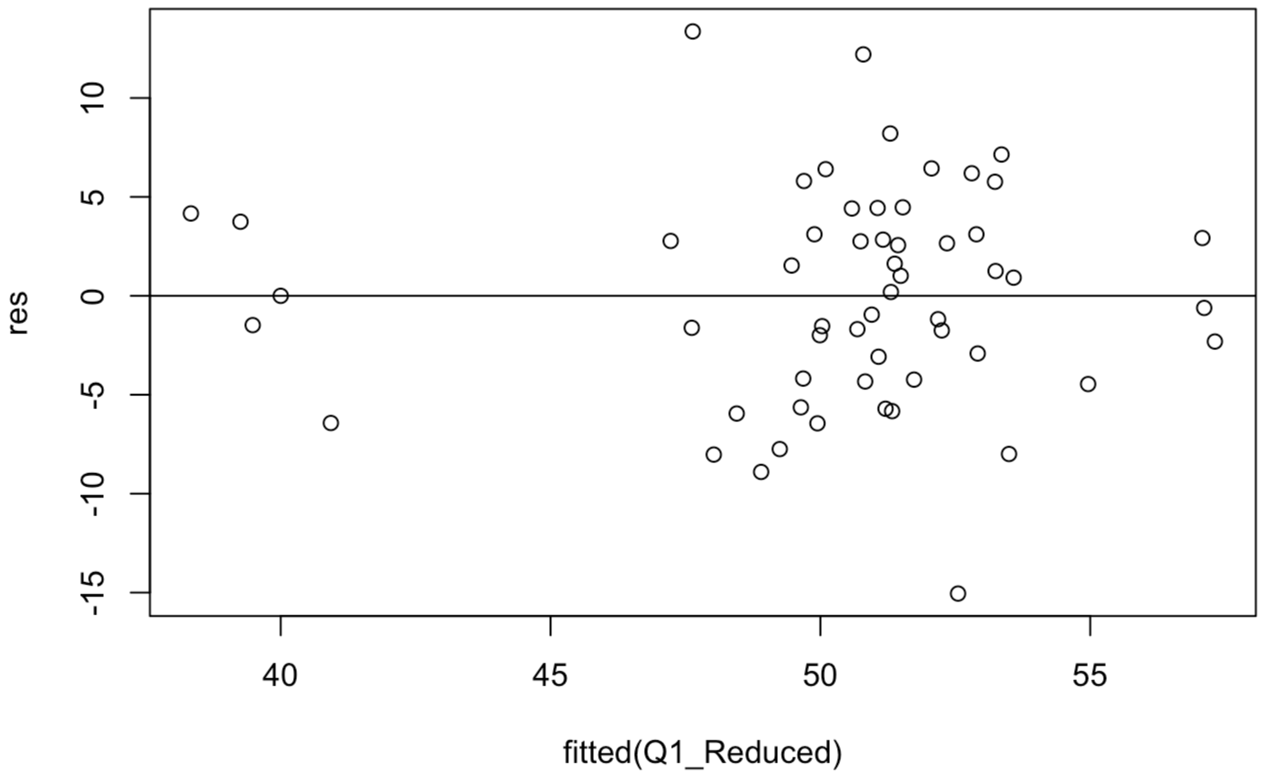


Figure 9 - Residual for Reduced Variables Linear Model

```
ggplot(Q1_Reduced, aes(x = .fitted, y = .resid, colour = Dominant.prey.species )) + geom_
```

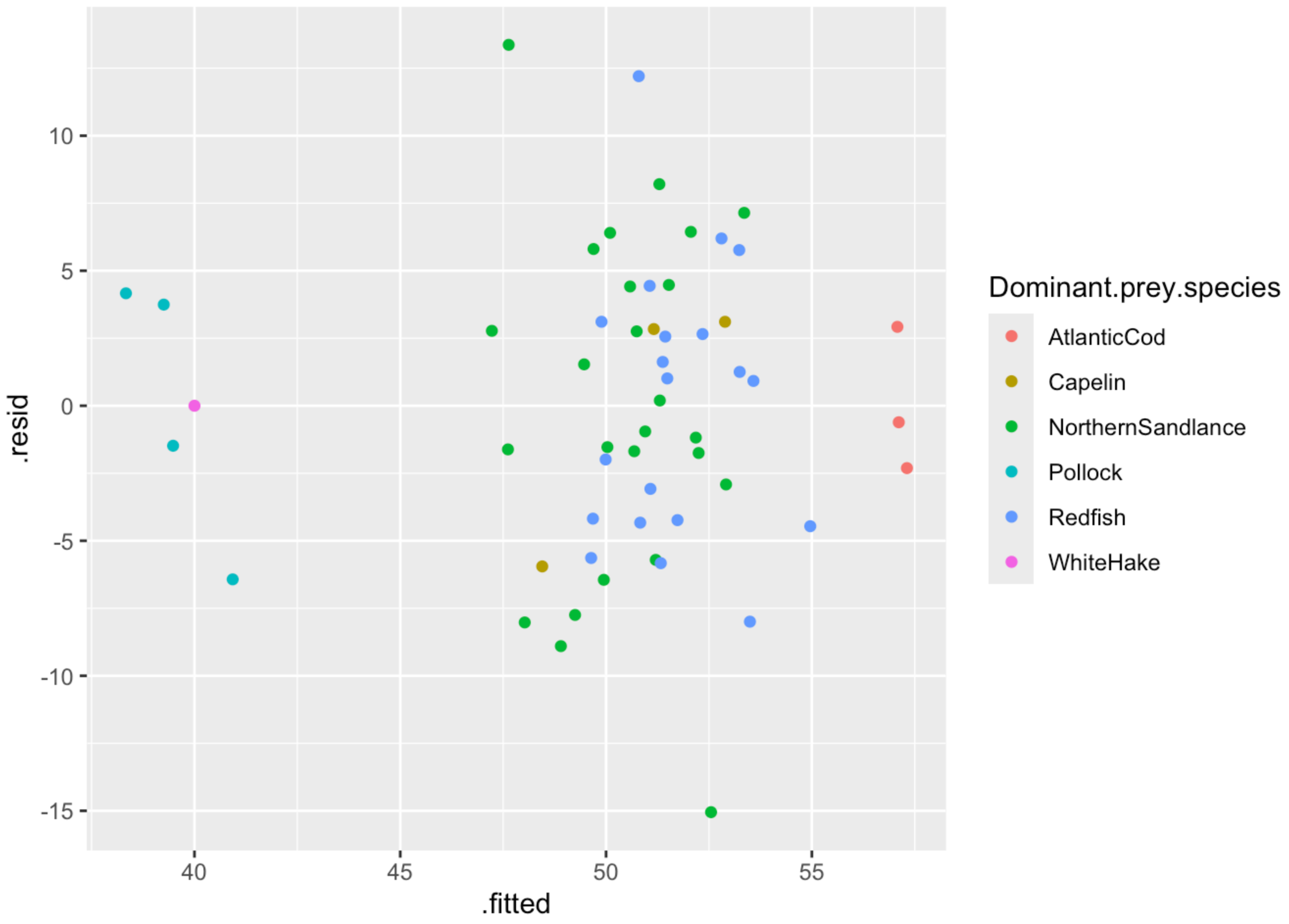


Figure 10 - Residual for Reduced Variables Linear Model

From these residual visualization it seems clear that a number observations seem to be outliers and they should be removed or we should try robust regression strategies. We can see from Figure 10 that it is the observation with White Hake and Pollock as the Dominant Prey Species are clearly outliers.

We will now try to fit a robust regression first.

```
require(robustbase)
```

```
Loading required package: robustbase
```

```
Q1_Reduced_Robust = lmrob(Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity,
                data = seal_data)
summary(Q1_Reduced_Robust)
```

```

Call:
lmrob(formula = Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity, data = seal_data)
 \--> method = "MM"
Residuals:
      Min        1Q    Median        3Q       Max
```

```
-15.74108  -3.76149  -0.08415   3.69887  13.97206

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)                              66.326      4.925  13.467  < 2e-16 ***
Dominant.prey.speciesCapelin             -4.698      2.646  -1.776 0.081995 .  
Dominant.prey.speciesNorthernSandlance   -6.460      1.825  -3.540 0.000889 ***
Dominant.prey.speciesPollock            -15.208      3.194  -4.761 1.74e-05 ***
Dominant.prey.speciesRedfish             -6.031      1.719  -3.508 0.000977 ***
Dominant.prey.speciesWhiteHake          -13.271      2.405  -5.518 1.28e-06 ***
Diet.diversity                          -25.898     13.446  -1.926 0.059903 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 5.928 
Multiple R-squared:  0.3699,    Adjusted R-squared:  0.2928 
Convergence in 12 IRWLS iterations

Robustness weights: 
 4 weights are ~= 1. The remaining 52 ones are summarized as
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.4607  0.9074  0.9590  0.9238  0.9847  0.9974 
Algorithmic parameters: 
       tuning.chi                bb        tuning.psi        refine.tol 
        1.548e+00         5.000e-01         4.685e+00         1.000e-07 
          rel.tol         scale.tol         solve.tol          zero.tol 
        1.000e-07         1.000e-10         1.000e-07         1.000e-10 
      eps.outlier             eps.x warn.limit.reject warn.limit.meanrw 
        1.786e-03         1.819e-12         5.000e-01         5.000e-01 
     nResample         max.it       best.r.s       k.fast.s          k.max 
           500             50              2              1            200 
   maxit.scale      trace.lev            mts     compute.rd fast.s.large.n 
           200              0           1000              0           2000 
                  psi           subsampling                   cov 
           "bisquare"         "nonsingular"         ".vcov.avar1" 
compute.outlier.stats 
                 "SM" 
seed : int(0) 
```

The robust regression returned a better Adjusted R-Squared = 2928 (appoximately 29% of variation) than the regular regression with that showed approximately 27%. Interestingly all the Dominant Prey Species are now significant coefficients except Capelin. We will now look at the residuals for the robust regression model.

```
#residuals
res = resid(Q1_Reduced_Robust)
plot(fitted(Q1_Reduced_Robust), res)
abline(0,0)
```

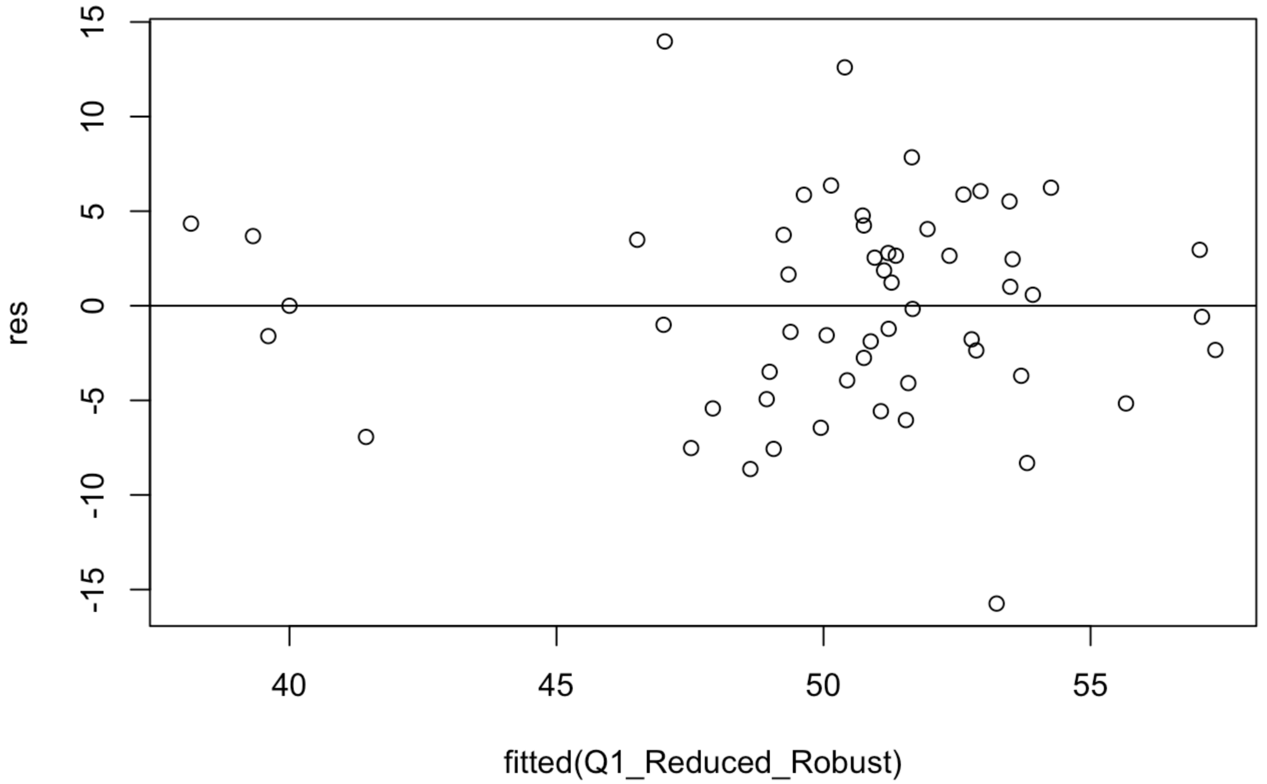


Figure 11 - Residual for Reduced Variables with Robust Linear Model

The residuals look a bit better but not a huge improvement over the linear model above. We will try removing the outlier observation. A new dataset was uploaded below with Pollock and White Hake Dominant Prey Species Removed.

```
seal_data_2 <- read.csv("/Users/peterbraithwaite/Desktop/IDPhD Classes/Stat 5620_Updated/
summary(seal_data_2)
```

```
     MomID             Year      Dietary.energy.density Diet.diversity  
 Min.   :   19.0   Min.   :1996   Min.   :5.207          Min.   :0.1788  
 1st Qu.:  701.5   1st Qu.:2002   1st Qu.:5.686          1st Qu.:0.3061  
 Median : 4269.0   Median :2011   Median :5.805          Median :0.3518  
 Mean   : 4897.8   Mean   :2008   Mean   :5.782          Mean   :0.3531  
 3rd Qu.: 9020.5   3rd Qu.:2013   3rd Qu.:5.904          3rd Qu.:0.3959  
 Max.   :10690.0   Max.   :2015   Max.   :6.517          Max.   :0.5290  
 Dominant.prey.species    Mom.Age          Pup.sex      Pup.Wean.Mass  
 Length:51             Min.   : 9.00   Min.   :1.000   Min.   :37.50  
 Class :character      1st Qu.:23.00   1st Qu.:1.000   1st Qu.:47.00  
 Mode  :character      Median :25.00   Median :1.000   Median :51.50  
                       Mean   :23.22   Mean   :1.431   Mean   :51.36  
                       3rd Qu.:26.00   3rd Qu.:2.000   3rd Qu.:55.50  
                       Max.   :31.00   Max.   :2.000   Max.   :63.00  
```

Re-fit Linear Model with Reduced Data-set

Removed Pollock and White Hake from Data-set

```
Q3_Full= lm(Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.densi

summary(Q3_Full)
```

```

Call:
lm(formula = Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + 
    Dietary.energy.density + Year + Mom.Age + Pup.sex, data = seal_data_2)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.9432  -3.2854   0.9013   3.1288  14.2253 

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)  
(Intercept)                            -258.8154   396.4474  -0.653   0.5174  
Dominant.prey.speciesCapelin             -0.8433     6.9966  -0.121   0.9046  
Dominant.prey.speciesNorthernSandlance   -3.2757     5.2773  -0.621   0.5381  
Dominant.prey.speciesRedfish             -2.9541     5.0552  -0.584   0.5621  
Diet.diversity                          -26.8081    11.8105  -2.270   0.0284 *
Dietary.energy.density                   -7.2831     6.0669  -1.200   0.2367  
Year                                      0.1835     0.2045   0.897   0.3747  
Mom.Age                                  -0.1580     0.2413  -0.655   0.5160  
Pup.sex                                  -0.1694     1.7915  -0.095   0.9251  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.95 on 42 degrees of freedom
Multiple R-squared:  0.1856,	Adjusted R-squared:  0.03048 
F-statistic: 1.197 on 8 and 42 DF,  p-value: 0.3243
```

```
# p-value: 0.3243
# Adjusted R-squared:  0.03048
```

Step Function (Forward)

We now used the Step Function to determine the best fit model using AIC.

```
fwd.model = step (Q3_Full, direction='forward')
```

```
Start:  AIC=190
Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.density + 
    Year + Mom.Age + Pup.sex
```

The forward step retained all variables with AIC = 190. It should be noted that this AIC is better than the best AIC 202.62 for the simple linear regression above.

Step Function (Backward)

```
backward.model = step(Q3_Full, direction='backward')
```

```
Start:  AIC=190
Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.density + 
    Year + Mom.Age + Pup.sex

                         Df Sum of Sq    RSS    AIC
- Dominant.prey.species   3    27.752 1514.5 184.94
- Pup.sex                 1     0.316 1487.0 188.01
- Mom.Age                 1    15.188 1501.9 188.52
- Year                    1    28.503 1515.2 188.97
- Dietary.energy.density  1    51.011 1537.7 189.72
<none>                                1486.7 190.00
- Diet.diversity          1   182.378 1669.1 193.90

Step:  AIC=184.94
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + Year + 
    Mom.Age + Pup.sex

                         Df Sum of Sq    RSS    AIC
- Pup.sex                 1     1.573 1516.0 182.99
- Year                    1    36.520 1551.0 184.16
- Mom.Age                 1    52.015 1566.5 184.66
<none>                                1514.5 184.94
- Dietary.energy.density  1   134.905 1649.4 187.29
- Diet.diversity          1   195.457 1709.9 189.13

Step:  AIC=182.99
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + Year + 
    Mom.Age

                         Df Sum of Sq    RSS    AIC
- Year                    1    41.970 1558.0 182.39
- Mom.Age                 1    52.962 1569.0 182.75
<none>                                1516.0 182.99
- Dietary.energy.density  1   134.114 1650.2 185.32
- Diet.diversity          1   195.334 1711.4 187.17

Step:  AIC=182.39
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + Mom.Age

                         Df Sum of Sq    RSS    AIC
- Mom.Age                 1    17.150 1575.2 180.94
<none>                                1558.0 182.39
```

```
- Dietary.energy.density  1    93.118 1651.1 183.35
- Diet.diversity          1   187.904 1745.9 186.19

Step:  AIC=180.94
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density

                         Df Sum of Sq    RSS    AIC
<none>                                1575.2 180.94
- Dietary.energy.density  1    93.233 1668.4 181.88
- Diet.diversity          1   197.610 1772.8 184.97
```

The backward step retained the Diet Diversity and Dietary.energy.density variables with AIC 180.94. It should be noted that this AIC is better than the best AIC 202.62 for the simple linear regression above.

Explore Backward Step Linear Model

```
Q3_Reduced = lm(Pup.Wean.Mass ~ Dietary.energy.density + Diet.diversity,
                data = seal_data_2)
summary(Q3_Reduced)
```

```

Call:
lm(formula = Pup.Wean.Mass ~ Dietary.energy.density + Diet.diversity,
    data = seal_data_2)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.3764  -4.3276   0.7659   3.9409  13.9590 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)              97.382     22.844   4.263 9.39e-05 ***
Dietary.energy.density   -6.374      3.781  -1.686   0.0984 .  
Diet.diversity          -25.954     10.576  -2.454   0.0178 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.728 on 48 degrees of freedom
Multiple R-squared:  0.1372,    Adjusted R-squared:  0.1012 
F-statistic: 3.815 on 2 and 48 DF,  p-value: 0.029
```

```
# p-value: 0.029
# Adjusted R-squared:  0.1012
```

Explore Residuals For Reduced DataSet Romoving White Hake and Pollock

```r
library(ggfortify)
autoplot(Q3_Reduced)
```

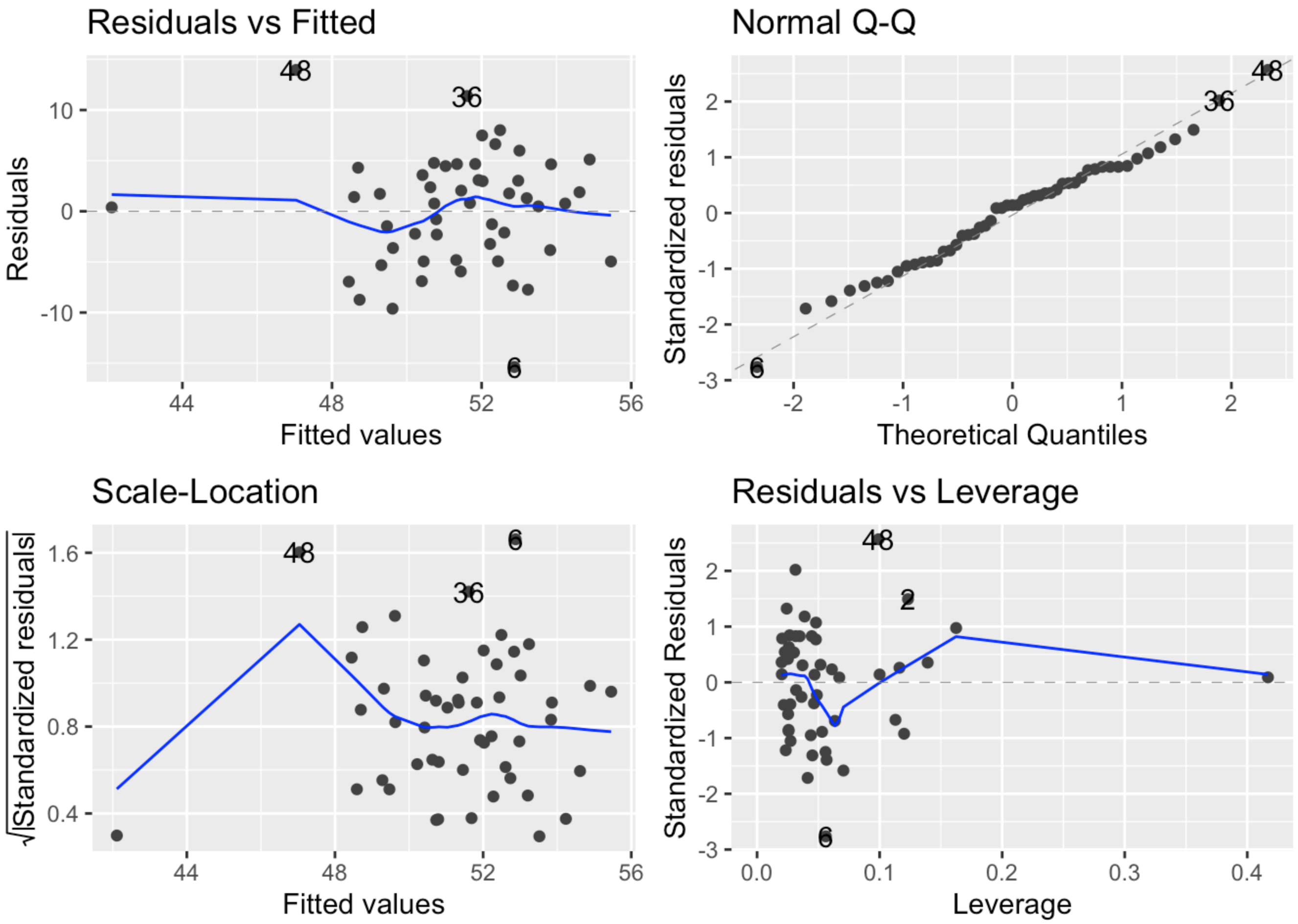


Figure 12 - Residual for Reduced Dataset (White Hake and Pollock removed).

```r
ggplot(Q3_Reduced, aes(x = .fitted, y = .resid, color = seal_data_2$Dominant.prey.species
```

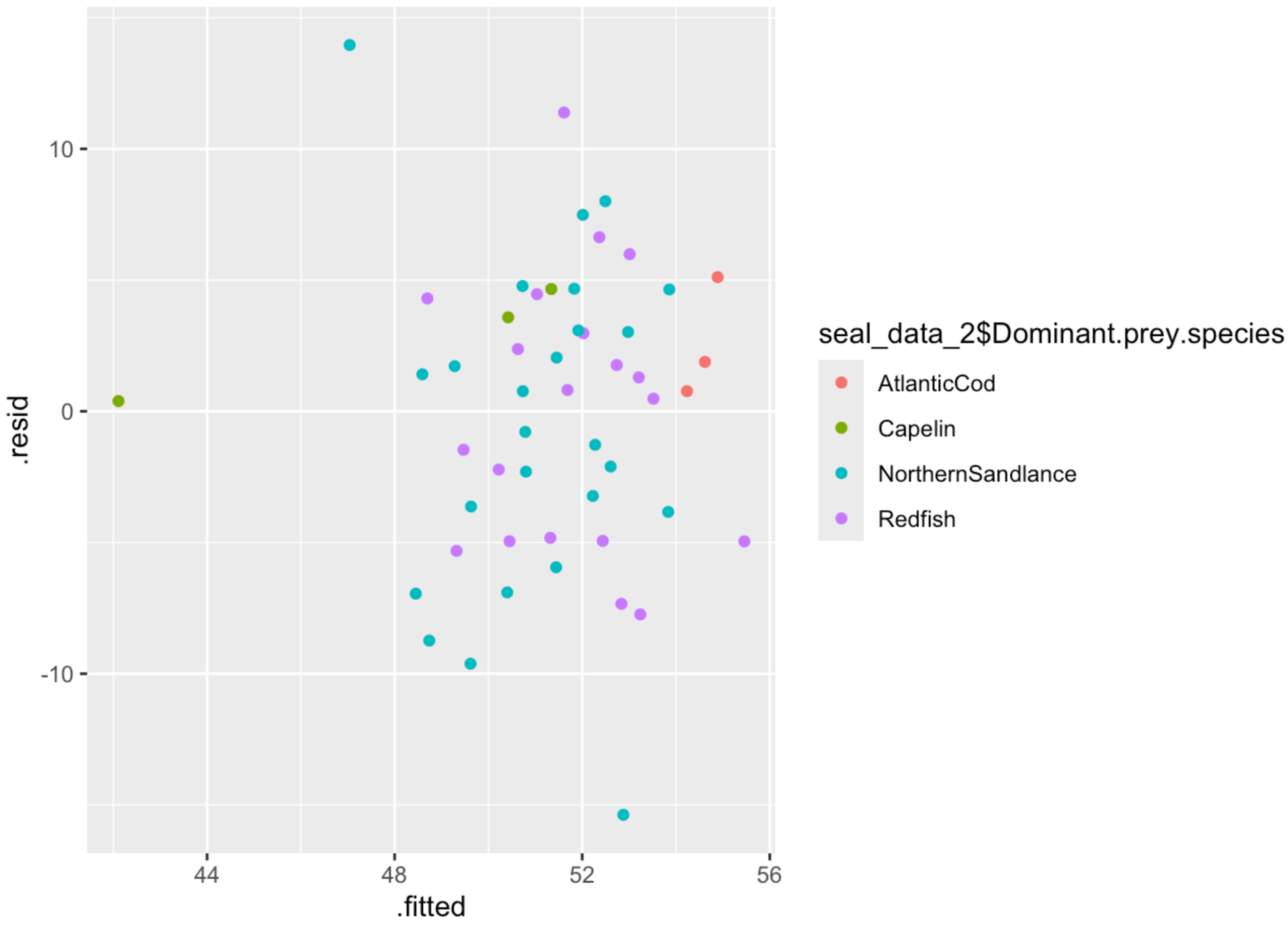


Figure 13 - Residual for Reduced Dataset (White Hake and Pollock removed).

The lowest value of Caplin is now an outlier and has very high leverage. It is most likely best to remove the lowest value of Capelin or remove the Capelin coefficient all together. Below we load a new dataset with the lowest value of Caplelin removed.

```
seal_data_3 <- read.csv("/Users/peterbraithwaite/Desktop/IDPhD Classes/Stat 5620_Updated/
summary(seal_data_2)
```

```
     MomID              Year      Dietary.energy.density Diet.diversity  
 Min.   :   19.0   Min.   :1996   Min.   :5.207          Min.   :0.1788  
 1st Qu.:  701.5   1st Qu.:2002   1st Qu.:5.686          1st Qu.:0.3061  
 Median : 4269.0   Median :2011   Median :5.805          Median :0.3518  
 Mean   : 4897.8   Mean   :2008   Mean   :5.782          Mean   :0.3531  
 3rd Qu.: 9020.5   3rd Qu.:2013   3rd Qu.:5.904          3rd Qu.:0.3959  
 Max.   :10690.0   Max.   :2015   Max.   :6.517          Max.   :0.5290  
 Dominant.prey.species    Mom.Age         Pup.sex      Pup.Wean.Mass  
 Length:51             Min.   : 9.00   Min.   :1.000   Min.   :37.50  
 Class :character      1st Qu.:23.00   1st Qu.:1.000   1st Qu.:47.00  
 Mode  :character      Median :25.00   Median :1.000   Median :51.50  
                       Mean   :23.22   Mean   :1.431   Mean   :51.36  
                       3rd Qu.:26.00   3rd Qu.:2.000   3rd Qu.:55.50  
                       Max.   :31.00   Max.   :2.000   Max.   :63.00  
```

Fit Full Model with Reduced Data Set

Removed Lowest Value of Capelin species from data.

```
Q4_Full= lm(Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.densi

summary(Q4_Full)
```

```
Call:
lm(formula = Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + 
    Dietary.energy.density + Year + Mom.Age + Pup.sex, data = seal_data_3)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.7728  -3.0865   0.7264   3.0387  12.9970 

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            -311.0480   401.3625  -0.775    0.443
Dominant.prey.speciesCapelin             -1.3014     7.0288  -0.185    0.854
Dominant.prey.speciesNorthernSandlance   -5.1433     5.6712  -0.907    0.370
Dominant.prey.speciesRedfish             -4.6264     5.3875  -0.859    0.395
Diet.diversity                          -20.3147    13.8136  -1.471    0.149
Dietary.energy.density                   -3.6124     7.2924  -0.495    0.623
Year                                      0.1990     0.2056   0.968    0.339
Mom.Age                                  -0.1911     0.2445  -0.782    0.439
Pup.sex                                  -0.1885     1.7953  -0.105    0.917

Residual standard error: 5.962 on 41 degrees of freedom
Multiple R-squared:  0.1651,    Adjusted R-squared:  0.002236 
F-statistic: 1.014 on 8 and 41 DF,  p-value: 0.441
```

```
# p-value: 0.002236
# Adjusted R-squared:0.004332
```

Step Function (Forward)

We now used the Step Function to determine the best fit model using AIC.

```
fwd.model = step (Q4_Full, direction='forward')
```

```
Start:  AIC=186.61
Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.density + 
    Year + Mom.Age + Pup.sex
```

The forward step retained all variables with AIC = 186.61.

Step Function (Backward)

```r
backward.model = step(Q4_Full, direction='backward')
```

```
Start:  AIC=186.61
Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.density +
    Year + Mom.Age + Pup.sex

                         Df Sum of Sq    RSS    AIC
- Dominant.prey.species   3    52.481 1509.7 182.38
- Pup.sex                 1     0.392 1457.6 184.62
- Dietary.energy.density  1     8.721 1465.9 184.91
- Mom.Age                 1    21.729 1478.9 185.35
- Year                    1    33.316 1490.5 185.74
<none>                                1457.2 186.61
- Diet.diversity          1    76.867 1534.1 187.18

Step:  AIC=182.38
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + Year +
    Mom.Age + Pup.sex

                         Df Sum of Sq    RSS    AIC
- Pup.sex                 1     1.992 1511.7 180.45
- Year                    1    39.426 1549.1 181.67
- Mom.Age                 1    56.342 1566.0 182.21
<none>                                1509.7 182.38
- Dietary.energy.density  1    77.795 1587.5 182.89
- Diet.diversity          1   116.741 1626.4 184.10

Step:  AIC=180.45
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + Year +
    Mom.Age

                         Df Sum of Sq    RSS    AIC
- Year                    1    45.154 1556.8 179.92
- Mom.Age                 1    56.672 1568.3 180.29
<none>                                1511.7 180.45
- Dietary.energy.density  1    78.067 1589.7 180.97
- Diet.diversity          1   117.958 1629.6 182.20

Step:  AIC=179.92
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + Mom.Age

                         Df Sum of Sq    RSS    AIC
- Mom.Age                 1    18.078 1574.9 178.50
- Dietary.energy.density  1    48.992 1605.8 179.47
<none>                                1556.8 179.92
- Diet.diversity          1   124.319 1681.1 181.76

Step:  AIC=178.5
```

```
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density

                         Df Sum of Sq    RSS    AIC
<none>                                1574.9 178.50
- Dietary.energy.density  1    66.647 1641.5 178.57
- Diet.diversity          1   159.201 1734.1 181.31
```

The backward step retained the Diet Diversity and Dietary.energy.density variables with AIC=178.5. It should be noted that this the best AIC value so far.

Explore Best Backward Step Linear Model

```
Q4_Reduced = lm(Pup.Wean.Mass ~ Dietary.energy.density + Diet.diversity,
                data = seal_data_3)
summary(Q4_Reduced)
```

```

Call:
lm(formula = Pup.Wean.Mass ~ Dietary.energy.density + Diet.diversity, 
    data = seal_data_3)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.3747  -4.5794   0.8175   4.1879  14.0677 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)   
(Intercept)              98.933     29.029   3.408  0.00135 **
Dietary.energy.density   -6.614      4.689  -1.410  0.16503   
Diet.diversity          -26.461     12.140  -2.180  0.03432 * 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.789 on 47 degrees of freedom
Multiple R-squared:  0.0977,	Adjusted R-squared:  0.0593 
F-statistic: 2.545 on 2 and 47 DF,  p-value: 0.08928
```

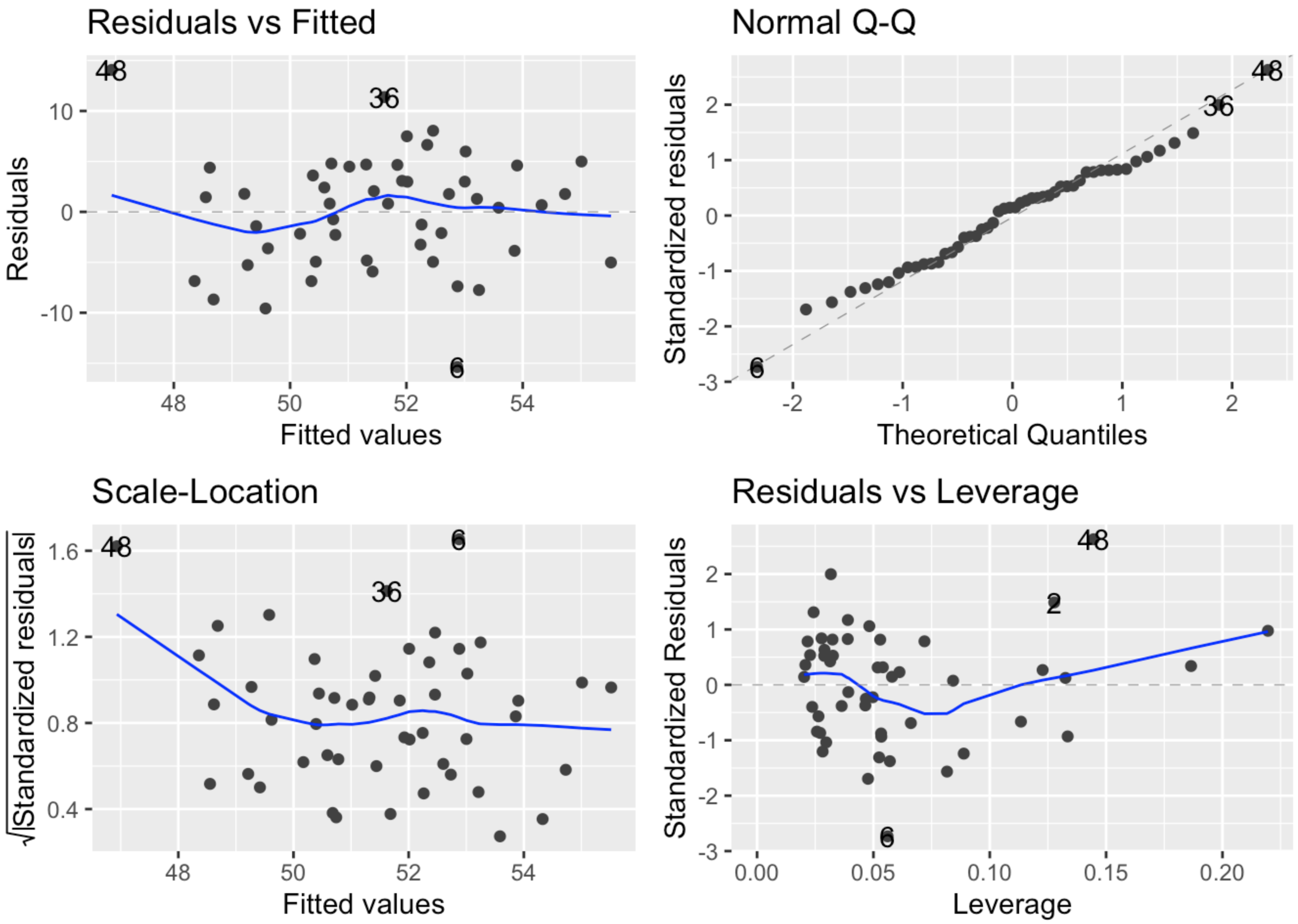
```
extractAIC(Q4_Reduced)
```

```
[1]   3.000 178.496
```

```
# p-value: 0.08928
# Adjusted R-squared:  0.0593
# AIC=178.496
```

Explore Residuals

```
library(ggfortify)
autoplot(Q4_Reduced)
```



```
ggplot(Q4_Reduced, aes(x = .fitted, y = .resid, color = seal_data_3$Dominant.prey.species
```

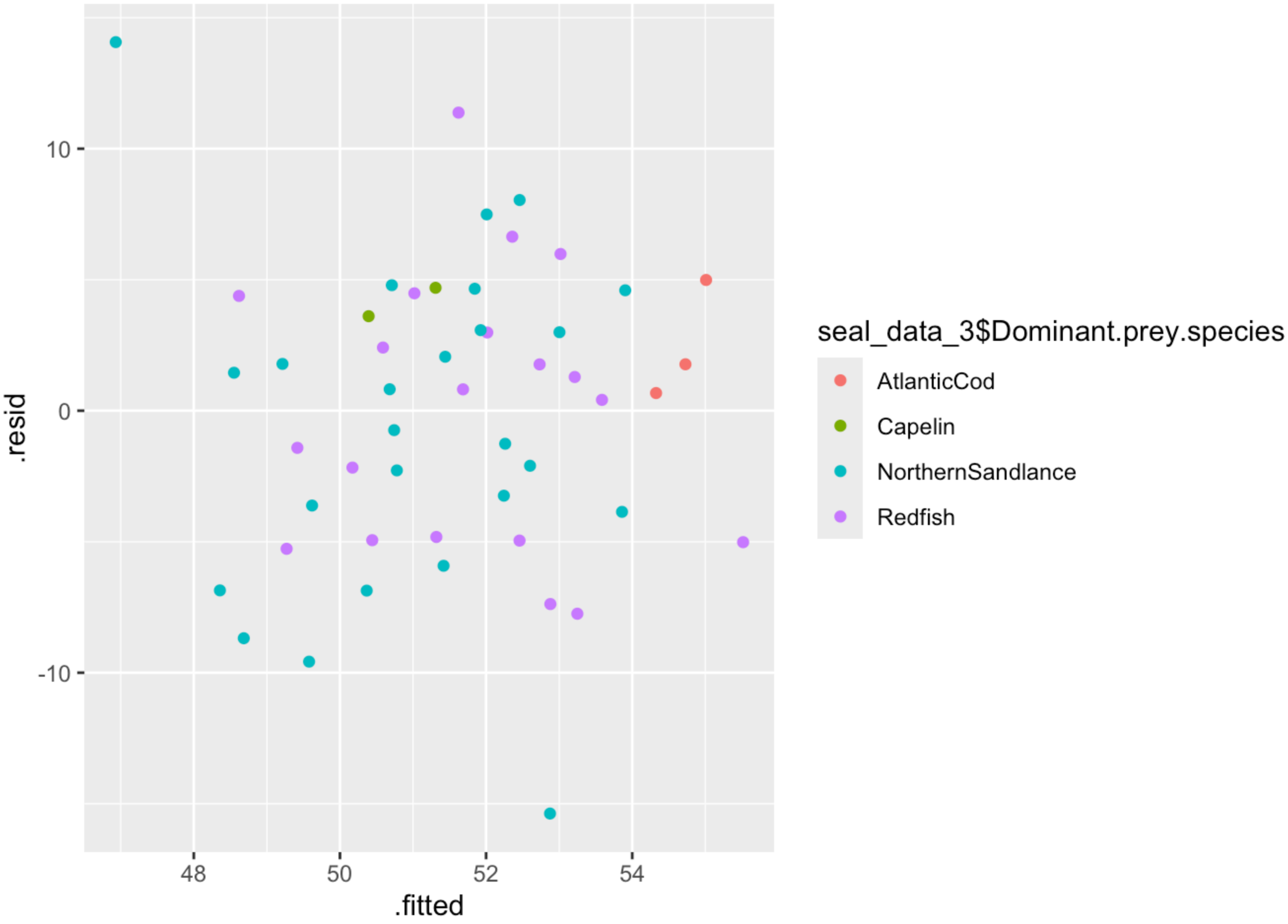


Figure 16 - Residual for Reduced Dataset (White Hake and Pollock removed completely and lowest value of Capeline).

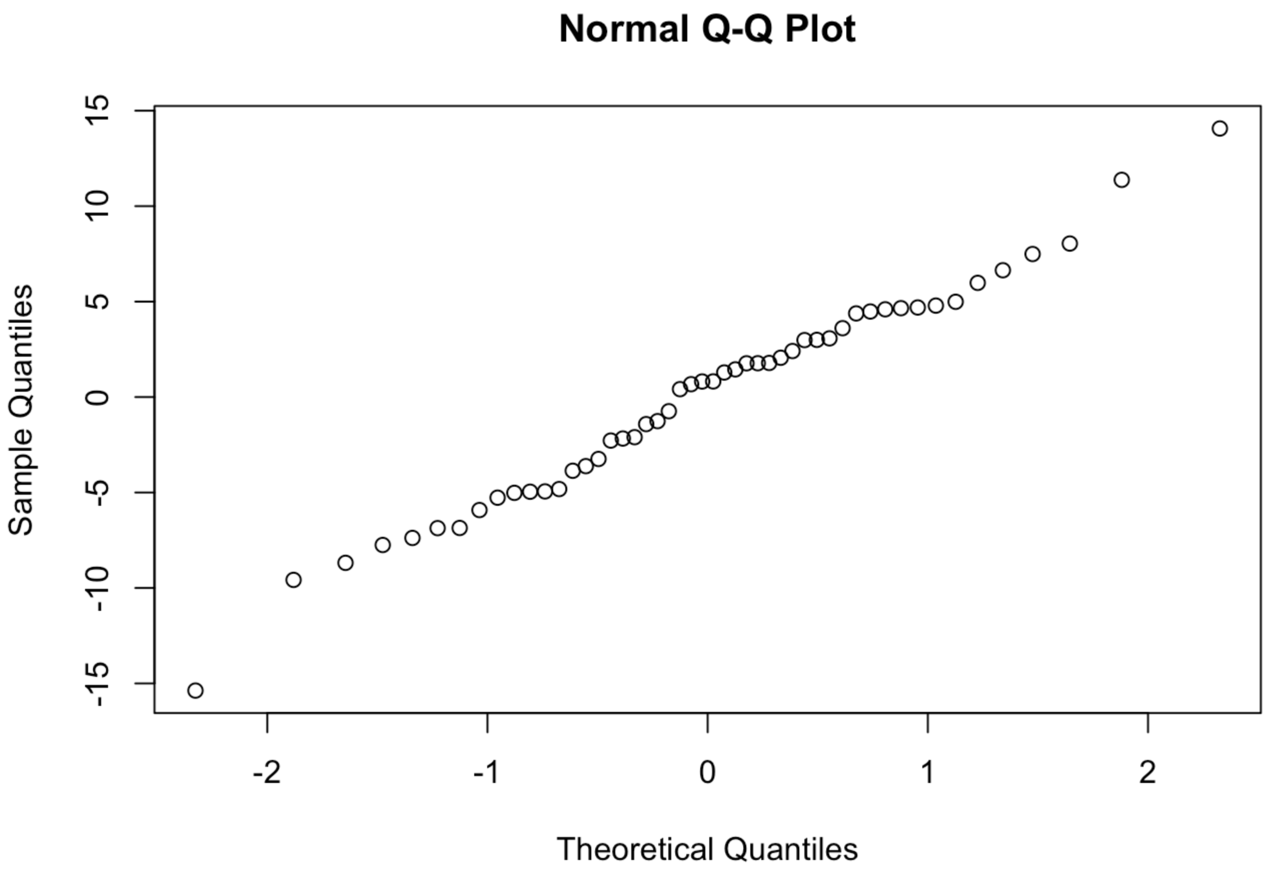
From this figure it seems we have a fairly balanced residuals now. Although Northerbn Sandlance is show some extreme values (both large and small) for risk of overfitting the model we will stop removing variables and observations.

Now, since this is a General Linear Model and not a Generalized Linear Model we will need to check for the assumptions including linearity, independence of errors, homoscedasticity, and normality.

Testing Assumptions

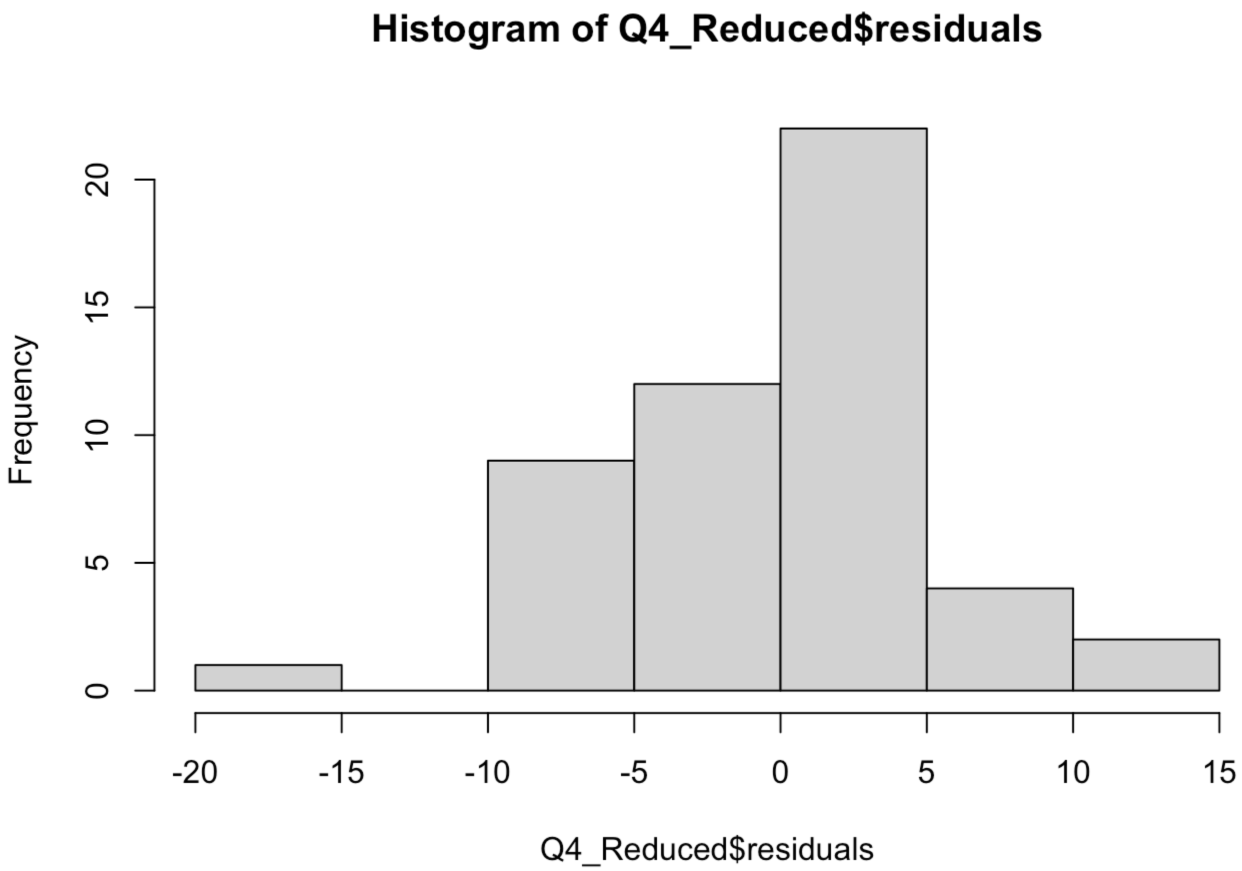
Checking for Normality

```
qqnorm(Q4_Reduced$residuals)
```



Q-Q plot appears to show normality

```
hist(Q4_Reduced$residuals)
```



Histogram appears to show normality

Shapiro-Wilk’s Test

```
shapiro.test(Q4_Reduced$residuals)
```

```

	Shapiro-Wilk normality test

data:  Q4_Reduced$residuals
W = 0.98681, p-value = 0.8456
```

From this W = 0.98681 and p-value = 0.8456. Given the p-value is not significant their is not a variation of the assumption of normality.

From these 3 explorations we are now confident in normality of the data.

Checking for Homogeneity of Variance

Using Breusch_Pagan Test

```
require (lmtest)
```

```
Loading required package: lmtest

Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```
bptest(Q4_Reduced)
```

```

    studentized Breusch-Pagan test

data:  Q4_Reduced
BP = 2.0797, df = 2, p-value = 0.3535
```

```
# BP = 2.0797
# df = 2
# p-value = 0.3535
```

Thus the p-value is greater then 0.05 so we accept the null hypothesis and assume we have homoscedasticity.

Checking for Independence of Predictor Variables

Using the Dublin-Watson

```
dwtest(Q4_Reduced)
```

```

    Durbin-Watson test

data:  Q4_Reduced
DW = 2.2815, p-value = 0.792
alternative hypothesis: true autocorrelation is greater than 0
```

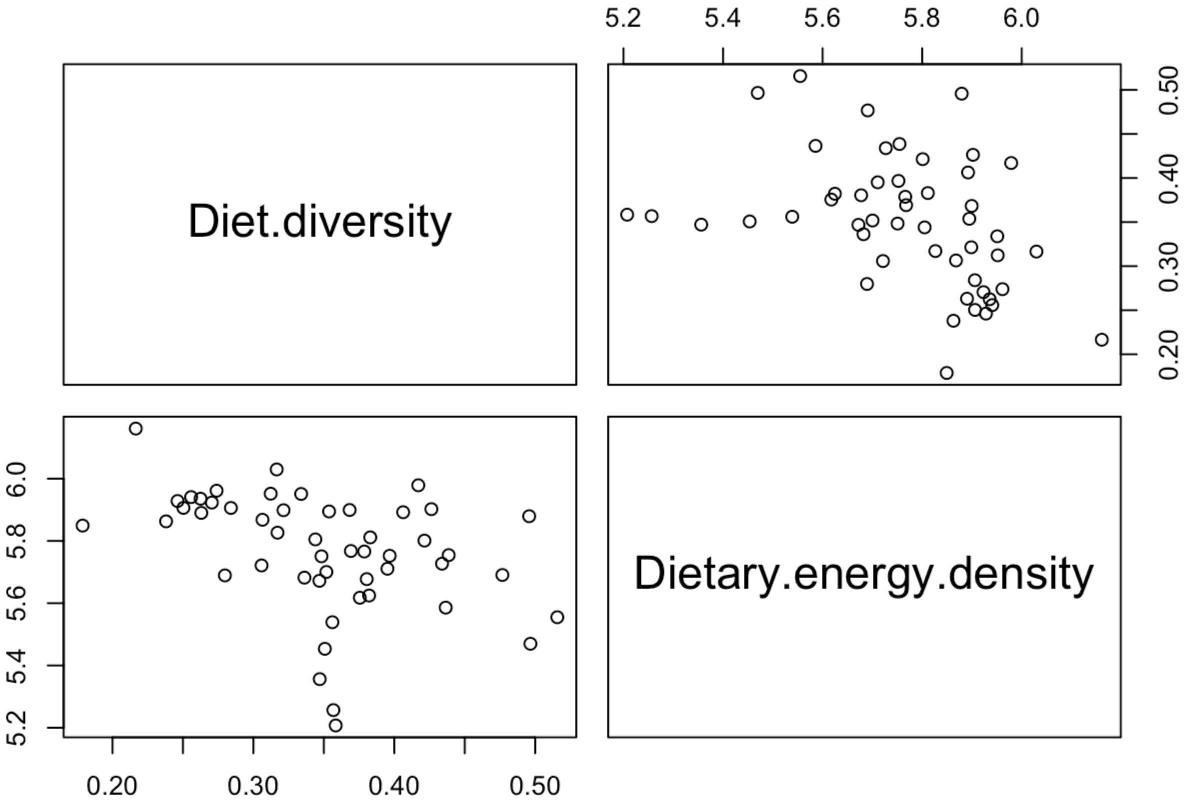
```
# DW = 2.2815
# p-value = 0.792
```

This p-value suggests very little to no autocorrelation.

Checking for Correlation of the Variables

Graphically Explore for Correlation

```
pairs( ~Diet.diversity + Dietary.energy.density, data = seal_data_3)
```



Data points look random and not correlated.

Numerically Explore for Correlation

```
cor.test( ~Diet.diversity + Dietary.energy.density, data = seal_data_3)
```

```

	Pearson's product-moment correlation

data:  Diet.diversity and Dietary.energy.density
t = -3.0499, df = 48, p-value = 0.00372
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6125589 -0.1402952
sample estimates:
      cor 
-0.402909 
```

```
# t = -0.35989
# df = 74
# p-value = 0.72
```

From this we can conclude that the variables are not correlated.

Thus, all our assumptions are met for the Linear Model and our findings can be accepted.

Generalized Linear Model

Given we will be fitting a Gaussian (normal) distribution Generalized Linear Model it seeem safe to assume that the model will demonstrate the same data issues as the Linear Model with regard to residual outliers. For this reason we start with the reduced dataset that eliminated the the Dominant Prey Species White Hake, Pollock, and the lowest value of Capelin.

Fitting Full Generalized Linear Model (Gaussian Family)

All Predictor Variables

```
GLM_1 = glm (Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.dens

summary (GLM_1)
```

```

Call:
glm(formula = Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + 
    Dietary.energy.density + Year + Mom.Age + Pup.sex + MomID, 
    family = gaussian, data = seal_data_3)

Coefficients:
                                         Estimate Std. Error t value Pr(>|t|)  
(Intercept)                            -1.076e+02  4.092e+02  -0.263   0.7940  
Dominant.prey.speciesCapelin           -4.464e+00  7.103e+00  -0.629   0.5332  
Dominant.prey.speciesNorthernSandlance -6.486e+00  5.592e+00  -1.160   0.2530  
Dominant.prey.speciesRedfish           -6.424e+00  5.363e+00  -1.198   0.2380  
Diet.diversity                         -2.408e+01  1.366e+01  -1.762   0.0856 .
Dietary.energy.density                 -3.922e+00  7.123e+00  -0.551   0.5850  
Year                                    9.835e-02  2.090e-01   0.471   0.6405  
Mom.Age                                -1.175e-01  2.425e-01  -0.484   0.6307  
Pup.sex                                -4.899e-01  1.762e+00  -0.278   0.7824  
MomID                                   4.208e-04  2.432e-04   1.730   0.0913 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 33.89238)

    Null deviance: 1745.4  on 49  degrees of freedom
Residual deviance: 1355.7  on 40  degrees of freedom
AIC: 328.9
```

```
Number of Fisher Scoring iterations: 2
```

```
# AIC = 328.9
```

From here we will use the GLM_1 model for the Step Funtion to reduce the number of predictor variables and increase the accuracy and efficiency of the model.

Step Function (Backward)

```
backward.model = step(GLM_1, direction='backward')
```

```
Start:  AIC=328.9
Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity + Dietary.energy.density + 
    Year + Mom.Age + Pup.sex + MomID

                         Df Deviance    AIC
- Dominant.prey.species   3   1412.7 324.96
- Pup.sex                 1   1358.3 326.99
- Year                    1   1363.2 327.17
- Mom.Age                 1   1363.7 327.19
- Dietary.energy.density  1   1366.0 327.27
<none>                        1355.7 328.90
- MomID                   1   1457.2 330.51
- Diet.diversity          1   1461.0 330.64

Step:  AIC=324.96
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + Year + 
    Mom.Age + Pup.sex + MomID

                         Df Deviance    AIC
- Pup.sex                 1   1416.7 323.10
- Year                    1   1422.6 323.30
- Mom.Age                 1   1444.5 324.07
<none>                        1412.7 324.96
- MomID                   1   1509.7 326.28
- Dietary.energy.density  1   1521.1 326.65
- Diet.diversity          1   1571.7 328.29

Step:  AIC=323.1
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + Year + 
    Mom.Age + MomID

                         Df Deviance    AIC
- Year                    1   1430.0 321.57
- Mom.Age                 1   1449.0 322.23
<none>                        1416.7 323.10
- MomID                   1   1511.7 324.34
- Dietary.energy.density  1   1525.0 324.78
- Diet.diversity          1   1576.8 326.45
```

```
Step:  AIC=321.57
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + Mom.Age +
    MomID

                         Df Deviance    AIC
- Mom.Age                 1   1449.0 320.23
<none>                        1430.0 321.57
- Dietary.energy.density  1   1525.9 322.81
- MomID                   1   1556.8 323.81
- Diet.diversity          1   1601.1 325.21

Step:  AIC=320.23
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + MomID

                         Df Deviance    AIC
<none>                        1449.0 320.23
- Dietary.energy.density  1   1569.6 322.22
- MomID                   1   1574.9 322.39
- Diet.diversity          1   1661.7 325.07
```

The backward step retained the Diet Diversity and Dietary.energy.density variables with AIC AIC = 320.23. It should be noted this is the same predictor variables that were selected in the Step Function as the Linear Model but the AIC is much worse then the reduced Linear Model with AIC = 178.5. This is because the Linear Model is a less complex model and thus has a lower AIC.

Explore Backward Step Linear Model

```r
GLM_2 = glm(Pup.Wean.Mass ~ Dietary.energy.density + Diet.diversity,
                data = seal_data_3, family = gaussian)
summary(GLM_2)
```

```

Call:
glm(formula = Pup.Wean.Mass ~ Dietary.energy.density + Diet.diversity,
    family = gaussian, data = seal_data_3)

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              98.933     29.029   3.408  0.00135 **
Dietary.energy.density   -6.614      4.689  -1.410  0.16503
Diet.diversity          -26.461     12.140  -2.180  0.03432 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 33.50838)

    Null deviance: 1745.4  on 49  degrees of freedom
Residual deviance: 1574.9  on 47  degrees of freedom
AIC: 322.39
```

```
Number of Fisher Scoring iterations: 2
```

```
# AIC = 322.39
```

Explore Residuals

```
library(ggfortify)
autoplot(GLM_2)
```

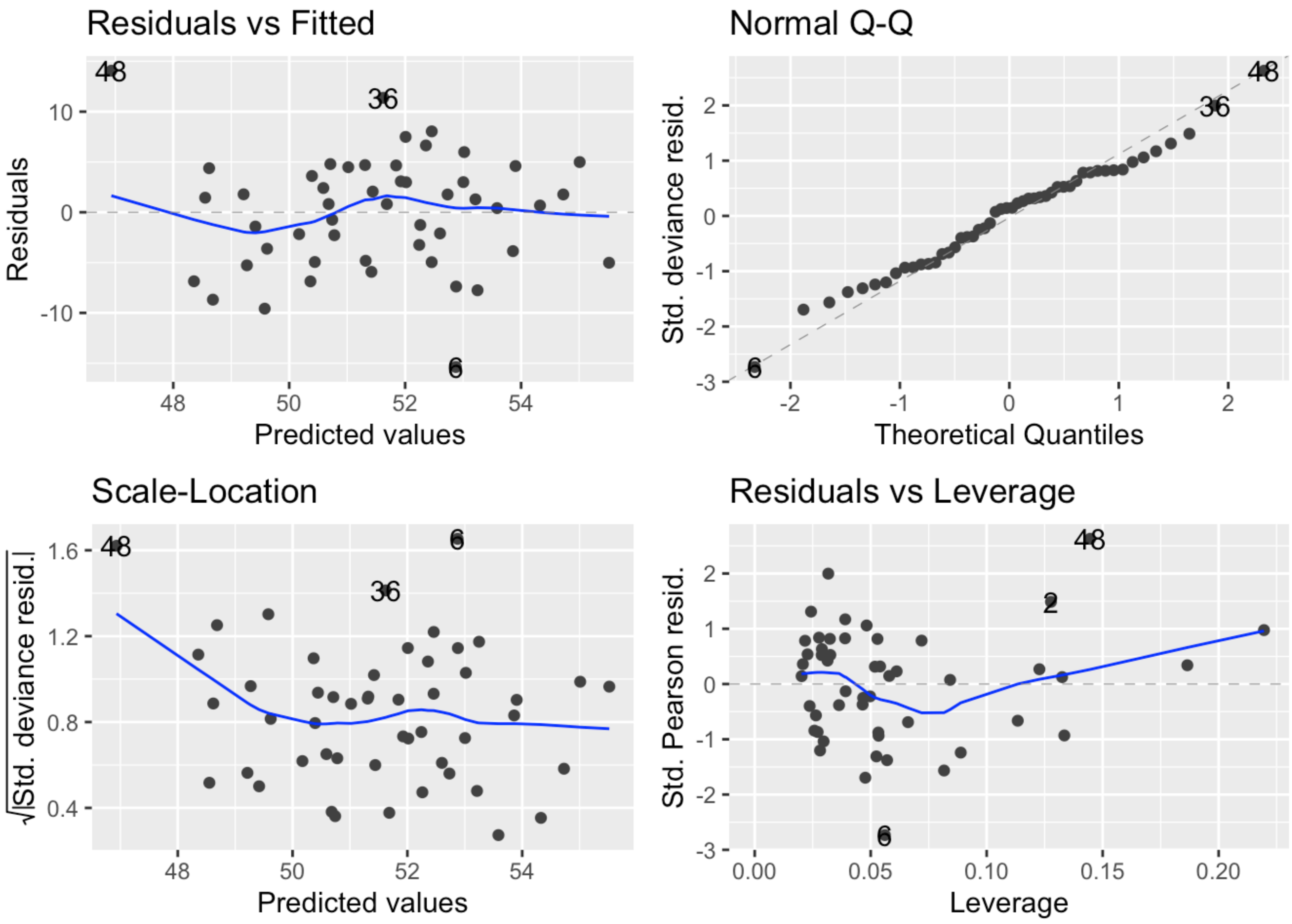


Figure 17 - Residual for Generalized Linear Model With Reduced Dataset

As we might expect the residuals for this Gaussian Generalized Linear Model are very similar to the residuals for the Linear Model above.

```
ggplot(GLM_2, aes(x = .fitted, y = .resid, color = seal_data_3$Dominant.prey.species )) +
```

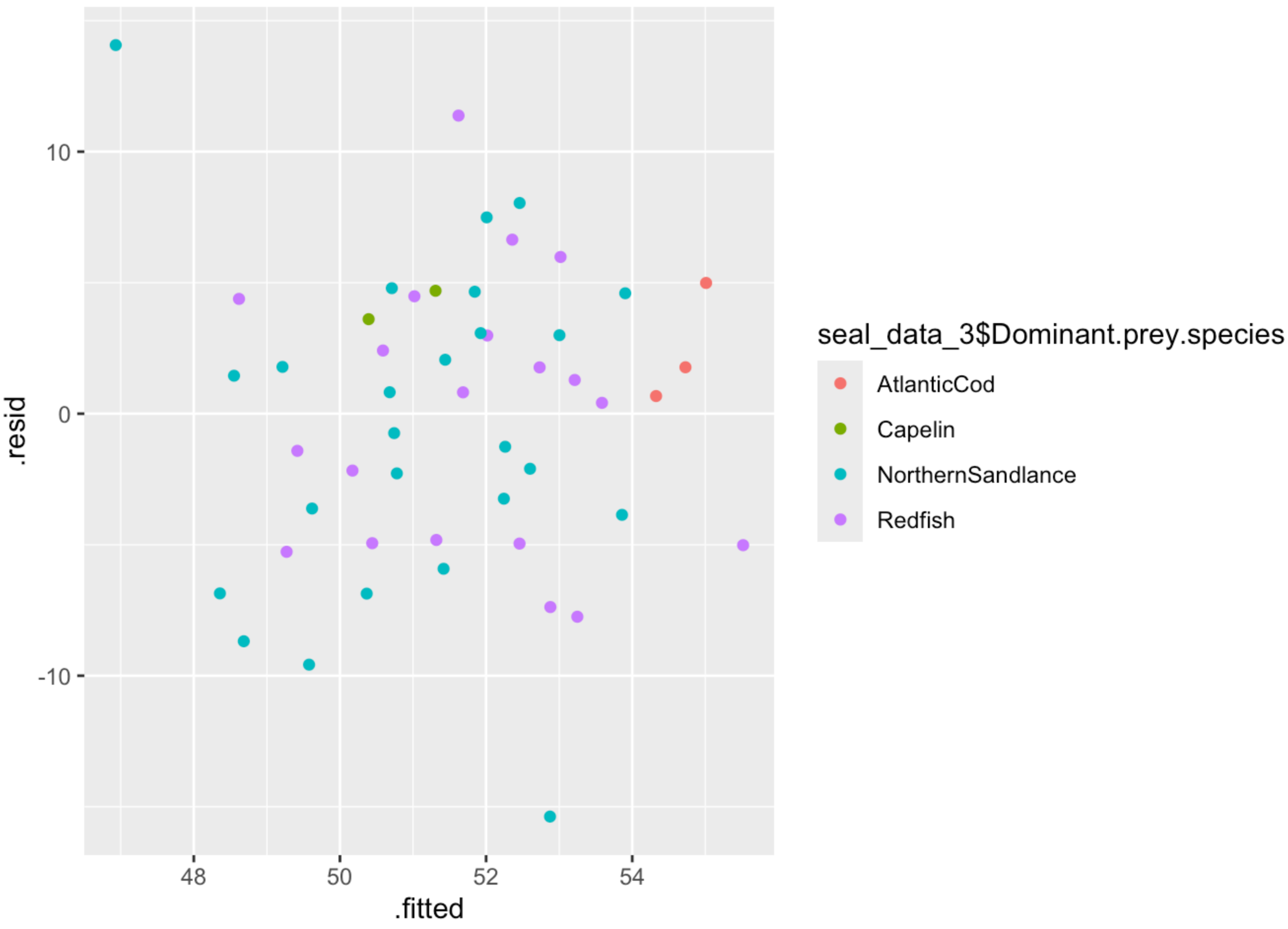


Figure 18 - Residual for Generalized Linear Model With Reduced Dataset

Again, as we might expect the residuals for this Gaussian Generalized Linear Model are very similar to the residuals for the Linear Model above. Given the Generalized Linear Model is creating similar residuals with the a same predictor variables as the Linear Model but with a less desirable AIC (GLM = 322.39 vs. LM = 178.5), it
seems the more simple Linear Model should be used.

Fit Linear Mixed Model

The Linear Model may benefit from Random Effects on the Dominant Prey Species variable given the different prey species would have an effect on the other two remaining variables, Dietary Energy Density and Diet Diversity.

Linear Mixed Model With Both Diet Diversity and Dietary Energy Loss Random Effects Based on Cluster Variable Dominant Prey Species

```
require(lme4)

GLMM_1 = lmer(Pup.Wean.Mass ~  Diet.diversity + Dietary.energy.density +  (Dietary.energy
boundary (singular) fit: see help('isSingular')
```

```
summary(GLMM_1)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula:
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + (Dietary.energy.density |
    Dominant.prey.species) + (Diet.diversity | Dominant.prey.species)
   Data: seal_data_3

REML criterion at convergence: 301.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6560 -0.7911  0.1412  0.7235  2.4302

Random effects:
 Groups                  Name                   Variance  Std.Dev. Corr
 Dominant.prey.species   (Intercept)            1.228e-02 0.11081
                         Dietary.energy.density 3.838e-04 0.01959  -1.00
 Dominant.prey.species.1 (Intercept)            0.000e+00 0.00000
                         Diet.diversity         2.394e-04 0.01547   NaN
 Residual                                       3.351e+01 5.78864
Number of obs: 50, groups:  Dominant.prey.species, 4

Fixed effects:
                       Estimate Std. Error t value
(Intercept)              98.933     29.029   3.408
Diet.diversity          -26.461     12.140  -2.180
Dietary.energy.density   -6.613      4.689  -1.410

Correlation of Fixed Effects:
            (Intr) Dt.dvr
Diet.dvrsty -0.522
Dtry.nrgy.d -0.991  0.403
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

```
extractAIC(GLMM_1)
```

```
[1]  10.0000 334.3899
```

```
# AIC = 337.7947
```

Linear Mixed Model With Dietary Energy Loss Random Effects Based on Cluster Variable Dominant Prey Species

```
GLMM_2 = lmer(Pup.Wean.Mass ~  Diet.diversity + Dietary.energy.density +  (Dietary.energy
```

```
boundary (singular) fit: see help('isSingular')
```

```
summary(GLMM_2)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula:
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + (Dietary.energy.density |
    Dominant.prey.species)
   Data: seal_data_3

REML criterion at convergence: 301.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6560 -0.7911  0.1412  0.7235  2.4302

Random effects:
 Groups                Name                   Variance  Std.Dev. Corr
 Dominant.prey.species (Intercept)            1.420e-03 0.037686
                       Dietary.energy.density 4.228e-05 0.006502 -1.00
 Residual                                     3.351e+01 5.788642
Number of obs: 50, groups:  Dominant.prey.species, 4

Fixed effects:
                       Estimate Std. Error t value
(Intercept)              98.933     29.029   3.408
Diet.diversity          -26.461     12.140  -2.180
Dietary.energy.density   -6.614      4.689  -1.410

Correlation of Fixed Effects:
            (Intr) Dt.dvr
Diet.dvrsty -0.522
Dtry.nrgy.d -0.991  0.403
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

```
extractAIC(GLMM_2)
```

```
[1]   7.0000 328.3899
```

```
# AIC = 328.3899
```

Linear Mixed Model With Diet Diversity Random Effects Based on Cluster Variable Dominant Prey Species

```
GLMM_3 = lmer(Pup.Wean.Mass ~  Diet.diversity + Dietary.energy.density + (Diet.diversity
```

```
boundary (singular) fit: see help('isSingular')
```

```
summary(GLMM_3)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula:
Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + (Diet.diversity |
    Dominant.prey.species)
   Data: seal_data_3

REML criterion at convergence: 301.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6560 -0.7911  0.1412  0.7235  2.4302

Random effects:
 Groups                Name           Variance  Std.Dev.  Corr
 Dominant.prey.species (Intercept)    0.000e+00 0.000e+00
                       Diet.diversity 2.911e-17 5.395e-09  NaN
 Residual                             3.351e+01 5.789e+00
Number of obs: 50, groups:  Dominant.prey.species, 4

Fixed effects:
                       Estimate Std. Error t value
(Intercept)              98.933     29.029   3.408
Diet.diversity          -26.461     12.140  -2.180
Dietary.energy.density   -6.614      4.689  -1.410

Correlation of Fixed Effects:
            (Intr) Dt.dvr
Diet.dvrsty -0.522
Dtry.nrgy.d -0.991  0.403
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

```
extractAIC(GLMM_3)
```

```
[1]   7.0000 328.3899
```

```
# AIC = 328.3899
```

Linear Mixed Model With Intercept Random Effects Based on Cluster Variable Dominant Prey Species

```
GLMM_4 = lmer(Pup.Wean.Mass ~  Diet.diversity + Dietary.energy.density + (1 | Dominant.pr
```

```
boundary (singular) fit: see help('isSingular')
```

```
summary(GLMM_4)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Pup.Wean.Mass ~ Diet.diversity + Dietary.energy.density + (1 |
    Dominant.prey.species)
   Data: seal_data_3

REML criterion at convergence: 301.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6560 -0.7911  0.1412  0.7235  2.4302

Random effects:
 Groups                Name        Variance Std.Dev.
 Dominant.prey.species (Intercept)  0.00    0.000
 Residual                          33.51    5.789
Number of obs: 50, groups:  Dominant.prey.species, 4

Fixed effects:
                       Estimate Std. Error t value
(Intercept)              98.933     29.029   3.408
Diet.diversity          -26.461     12.140  -2.180
Dietary.energy.density   -6.614      4.689  -1.410

Correlation of Fixed Effects:
            (Intr) Dt.dvr
Diet.dvrsty -0.522
Dtry.nrgy.d -0.991  0.403
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

```
extractAIC(GLMM_4)
```

```
Warning in optwrap(optimizer, devfun, x@theta, lower = x@lower, calc.derivs =
TRUE, : convergence code 3 from bobyqa: bobyqa -- a trust region step failed to
reduce q
```

```
[1]   5.0000 324.3899
```

```
# AIC = 324.3899
```

So, in summary, none of the Linear Mixed Models with Random Effect on the cluster variavble Dominant Prey Species resulted in a more desirable AIC. The best Mixed Model model had an AIC = 324.3899 but the best Linear Model had an AIC=178.496. Thus the less complex Linear Model is the most efficient and most accurate model.

Cross Validation

It has been determined that a simple Linear Model using a maternal mother seal's Dietary Energy Density and Diet Diversity are the best predictor variables for Pup Wean Mass. The model lm(Pup.Wean.Mass ~ Dietary.energy.density + Diet.diversity, data = seal_data_3) has shown an AIC = 178.496.

We will now use Cross Validation to assess how well we expect this model to preform when predicting future Pup Wean Mass changes.

```r
require(modelr)
```

```
Loading required package: modelr
```

```r
require(caret)
```

```
Loading required package: caret
```

```
Loading required package: lattice
```

```r
require(readr)

# Data
attach(seal_data_3)

# Set Random Seed
set.seed(1980)

# Create Index Matrix (80% train data and 20% test data )
index = createDataPartition(seal_data_3$Pup.Wean.Mass, p = .8, list = FALSE, times = 1)

summary(seal_data_3$Pup.Wean.Mass)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  37.50   47.62   52.00   51.54   55.50   63.00 
```

```r
print(seal_data_3$Pup.Wean.Mass)
```

```
 [1] 50.5 60.5 50.0 54.5 45.5 37.5 54.5 59.0 50.5 51.0 58.5 59.0 56.0 55.0 56.0
[16] 51.5 59.5 45.5 50.0 47.5 53.5 49.0 55.0 52.5 54.0 55.0 53.0 45.5 56.5 60.0
[31] 48.0 55.5 56.5 48.5 46.5 63.0 43.5 55.5 54.0 51.0 41.5 48.0 53.0 40.0 45.5
[46] 44.0 40.0 61.0 46.0 50.0
```

```r
# Convert data frame
data_frame = as.data.frame(seal_data_3)

# Create a data frame for the train data that is 80%
train_data_frame = data_frame[index,]

# Create a data frame for the test data that is 20% ("-" before index mean everything but
test_data_frame = data_frame[-index,]

# Convert response variable in both train and test data frame to a factor
train_data_frame$Pup.Wean.Mass = as.factor (train_data_frame$Pup.Wean.Mass)
test_data_frame$Pup.Wean.Mass = as.factor (test_data_frame$Pup.Wean.Mass)


# Ensure the response variables classes are factors
class (train_data_frame$Pup.Wean.Mass)
```

```
[1] "factor"
```

```r
# "factor"
class (test_data_frame$Pup.Wean.Mass)
```

```
[1] "factor"
```

```r
# "factor"

# Specify type of training method used and the number of folds
control_specs = trainControl(method = "cv", number = 11 , savePredictions = "all")

# Set Random Seed
set.seed(1980)

require(randomForest)
```

```
Loading required package: randomForest

randomForest 4.7-1.2

Type rfNews() to see new features/changes/bug fixes.


Attaching package: 'randomForest'

The following object is masked from 'package:ggplot2':

    margin

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

```
# now we train the model with the best fit GLM above
# GLM_Final = glm (seal_data$Pup.Wean.Mass ~ seal_data$Dominant.prey.species + seal_data$


model_cross = train(Pup.Wean.Mass ~ Dominant.prey.species + Diet.diversity,
                    method = "lm", data = seal_data_3, trControl = control_specs)

print (model_cross)
```

```
Linear Regression

50 samples
 2 predictor

No pre-processing
Resampling: Cross-Validated (11 fold)
Summary of sample sizes: 45, 46, 46, 44, 46, 46, ...
Resampling results:

  RMSE      Rsquared   MAE
  5.434521  0.2021304  4.563434

Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
# 50 samples
#  2 predictor

#  RMSE      Rsquared    MAE
# 5.434521  0.20213048  4.563434


# We now apply the model to the test_data_frame we created from 20%
# of the data that the new model created from the 11 folds has not yet seen

# Predict outcome using model from train_data_frame applied to test_data_frame
Predict_Data = predict (model_cross, newdata = test_data_frame)

# predictions results

#. 52.74225 51.24721 50.67234 57.30204 51.49974 49.31590 50.08707 48.16223

test_data_frame$Pup.Wean.Mass
```

```
[1] 59   59.5 49   55   52.5 41.5 48   40
Levels: 40 41.5 48 49 52.5 55 59 59.5
```

```
# 59   59.5 49   55   52.5 41.5 48   40

#Creates vectors having data points
predicted_value <- (c(52.74225, 51.24721, 50.67234, 57.30204, 51.49974, 49.31590, 50.0870
expected_value <- (c(59, 59.5, 49, 55, 52.5, 41.5, 48, 40))


Cross_Validation_Corrolation = cor (predicted_value, expected_value)

### 0.6596914

R2 = (Cross_Validation_Corrolation^2 )

### 0.4351927 (43.52% of variance)
```

From this cross validation procedure we can see that the Linear Model has a validation score of about 0.66 which mean the Linear Model would do a fair job at predicting Pup Mean Mass based on the maternal mother seal's Dietary Energy Density and Diet Diversity. The estimated R-squared value suggests the model would account for about 43.5% of the variance of Pup Mean Mass.

Results

Although it was proposed that many predictor variables (2 continuous + 3 categorical for question one and 2 continuous + 5 categorical for question two) would allow greater computational potential to establish a well fit model, the added complexity of these variables was observed to significantly increase the AIC and potentially overfit the model. It was found, particularly in question two, that the simplicity of a Linear Model with a Least Square approach resulted in a simpler linear model with a significantly lower AIC.

In question 2 it was found that best fit Linear Model for Pup Wean Mass included the predictor variables Dietary Energy Density and Diet Diversity with an AIC = 178.496. When a Generalized Linear Model was explored with a Gaussian Family and Identity Link Function it was also found that the best fit model included Dietary Energy Density and Diet Diversity predictor variable but the AIC was much less preferable at 322.39. The reasons for this decreased predicted fit is likely based on the more complex Maximum Likelihood Estimation used in Generalized Linear Models rather than the Linear Least Square method used in the Linear Models. Generalized Linear Mixed Models (GLMM) were also explored using the Dominant Prey Species as the Cluster Variable because it was assumed that this variable would likely have a effect on the prey energy density as well as the diet diversity based on feeding location ecosystems. The best fit GLMM was found to include Diet Diversity and Dietary Energy Density as fixed effects predictor variables and Dominant Prey Species as an intercept random effect variable which had a AIC of 324.3899. As a result it was determined that the best fit model to determine Pup Wean Mass was the Linear Model and this model was tested using Cross Validation. The correlation of predicted values to actual values showed a value of 0.6596914 meaning the Linear Model could be expected to predict Pup Wean Mass with about a 66 percent accuracy.

Conclusions

References

Data source : Sara Iverson laboratory and DFO Grey Seal Program.