Influences on the Survival of Juvenile Smolt using Logistic Regression

Introduction

Salmon are spawned and spend their juvenile life in freshwater rivers or lakes before migrating out to sea where they spend their adult lives and gain most of their body mass. Once they reach sexual maturity they migrate from the sea back to freshwater to reproduce, after which they usually die and the salmon life cycle starts over again. In this analysis, we strive to understand the effect that migration date and weight of juvenile salmon, or smolt as they are commonly referred to, has on the probability that juvenile smolt return to their natal stream as reproductive adults (ie survive) and also how telomere length and age of juvenile smolt as they leave their natal stream for the sea influence survival.

Statistical Methods

Data was collected by researchers who caught and tagged several thousand smolt as they were leaving their natal stream for the sea, and then recaptured returning fish at the same natal stream over the following two years. When the fish were captured on their outward journey the date was recorded and each fish was weighed, aged (an estimate of the years each juvenile fish had spent in freshwater before their seaward migration), and a biopsy was taken and subsequently analysed to quantify telomere length using quantitative PCR. For this analysis we have 2 datasets. The first dataset contains 5 variables, smolt ID (Smolt.ID), the migration date expressed in the normal date format (Date) and Julian date (Julian.Date), weight in grams (Weight..g.) and whether or not the smolt survived (Survival.Binary) which is our response variable. Upon examination of the data the datapoint associated to smolt ID 730 was removed as it's weight was over 10 standard deviations away from the mean and almost 30g heavier than the next heaviest smolt signifying a significant outlier to the rest of the data. After the removal of this datapoint we are left with a sample size of n = 1805 to run our analysis with. The second dataset contains 4 variables, smolt ID (smolt.ID), the estimated freshwater age of the juvenile smolt (FW.age), telomere length (Smolt.RTL) and whether or not the smolt survived (Survival.Binary) which is our response variable. Upon examination of the data there were two datapoints with FW.age listed as "not established", these were removed before analysis. due to being incomplete. There was also a significant outlier within the smolt telomere length data with one observation being roughly five standard deviations away from the mean and over two standard deviations away from the next highest value, which was subsequently removed. This leaves us with a sample size of n = 66. Two models were fit using the glm() function, one model analyses the effect that migration date and weight have on survival while the other model tests the effect of age and telomere length on survival. The final models take the form

Model 1:

$$Survival.Binary_i \sim binomial(1805, p_i)$$

$$f(p_i) = \beta_0 + \beta_1 \times \text{Julian Date}_i + \beta_2 \times \text{Weight(g)}_i$$

$$f(p_i) = log(\frac{p_i}{1-p_i})$$

Model 2:

$$Survival.Binary_i \sim binomial(67, p_i)$$

$$f(p_i) = \beta_0 + \beta_1 \times \text{fresh water age}_i + \beta_2 \times \text{telomere length}_i$$

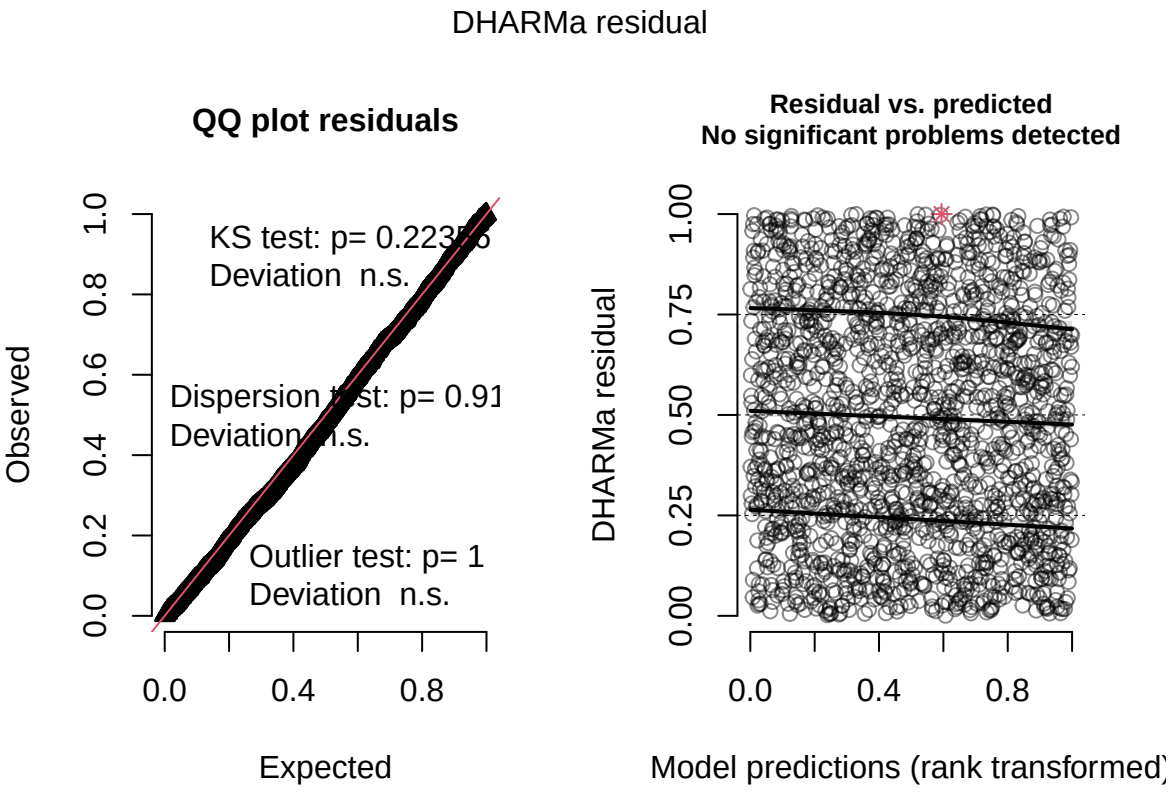
$$f(p_i) = log(\frac{p_i}{1-p_i})$$

Model fit was assessed using the pregibon test and model assumptions were checked using diagnostic plots from the DHARMa package.

Results

Model 1

Model 1 was fit using the glm() function and assessed using the pregibbon test which resulted in a test p-value of 0.2748. The model assumptions were assessed using the DHARMa package producing the following diagnostic plots




```
## Object of Class DHARMa with simulated residuals based on 250 simulations with refit = FALSE . See ?DH
##
## Scaled residual values: 0.6988884 0.6198124 0.3337293 0.6698115 0.4392463 0.3025449 0.8369241 0.11819
```

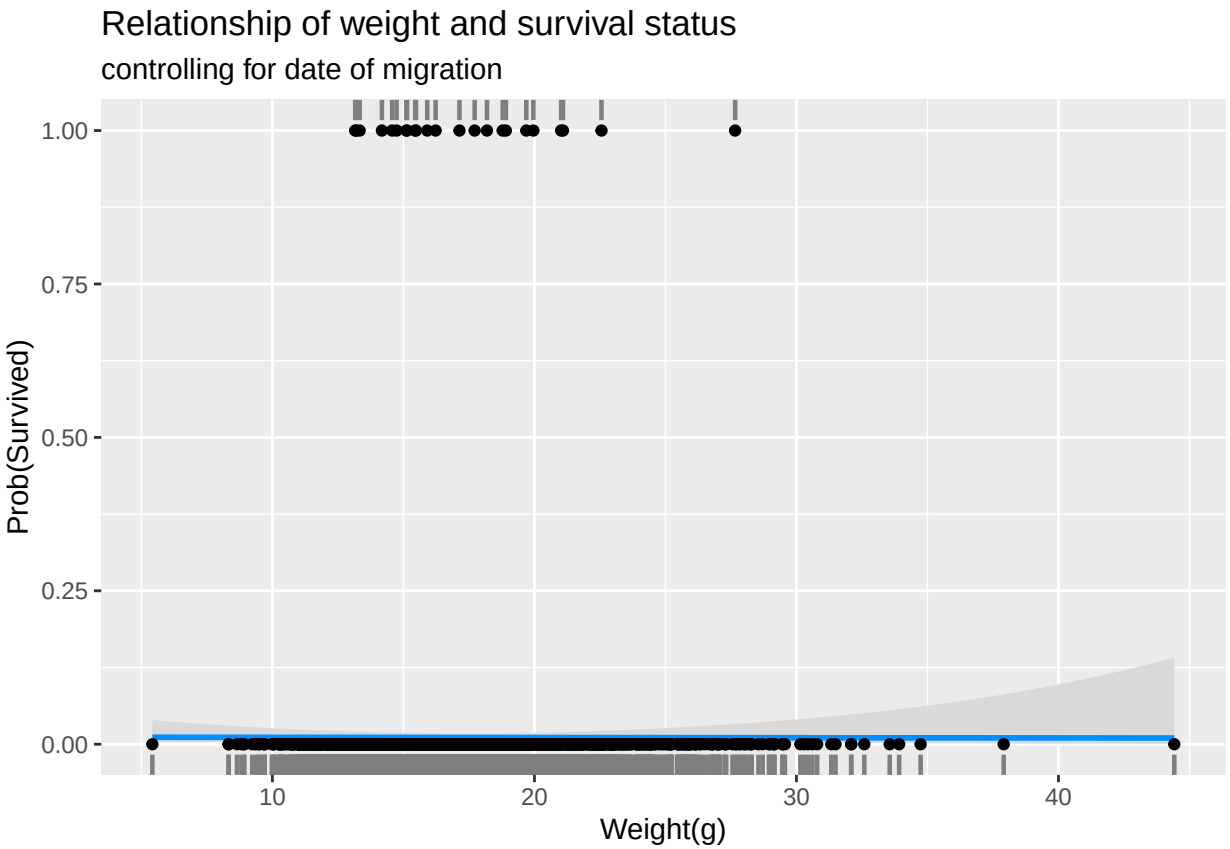
As you can see from the above diagnostic plots the model assumptions seem to be met and the model fits quite nicely.

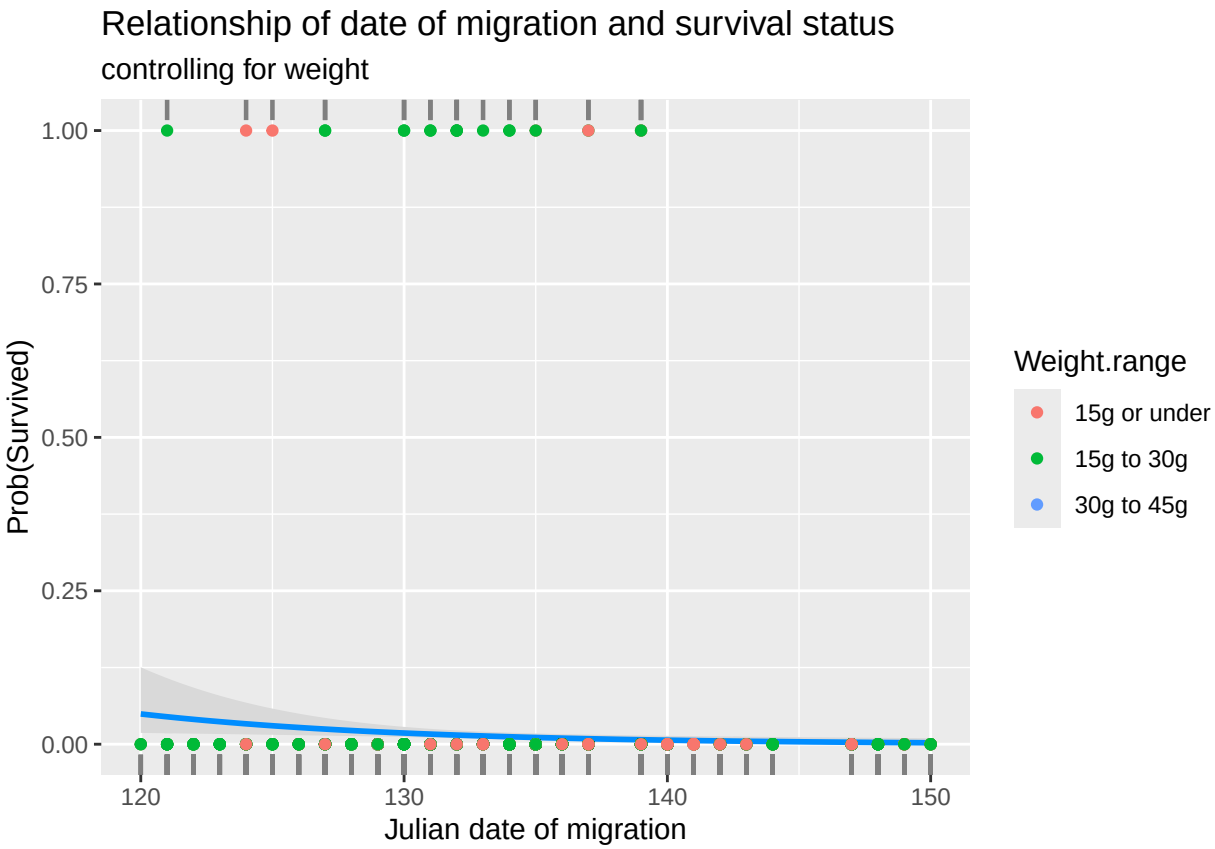
The summary information for model 1 is as follows

```
##
## Call:
## glm(formula = Survival.Binary ~ Julian.Date + Weight..g., family = "binomial",
##     data = Salmon1.clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3212  -0.1791  -0.1480  -0.1220   3.1469
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  9.420872   5.293401   1.780   0.0751 .
## Julian.Date -0.102747   0.038927  -2.639   0.0083 **
## Weight..g.  -0.003016   0.051313  -0.059   0.9531
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 246.40  on 1804  degrees of freedom
## Residual deviance: 239.33  on 1802  degrees of freedom
## AIC: 245.33
##
## Number of Fisher Scoring iterations: 7
```

From the above model summary output we have an intercept and two coefficients relating to our two explanatory variables. For every one day increase in our migration date (Julian.Date in this model) we have a 0.103 decrease in our log odds of survival and for every one gram increase increase in the weight of our juvenile smolt we see a 0.003 decrease in our log odds of survival. It makes sense that the larger smolt could be bigger targets for predatory fish but the association of migration date is surprising. Further analysis could be undertake to determine if the julian date of migration of juvenile smolt could coincide with migratory patterns of predatory fish.

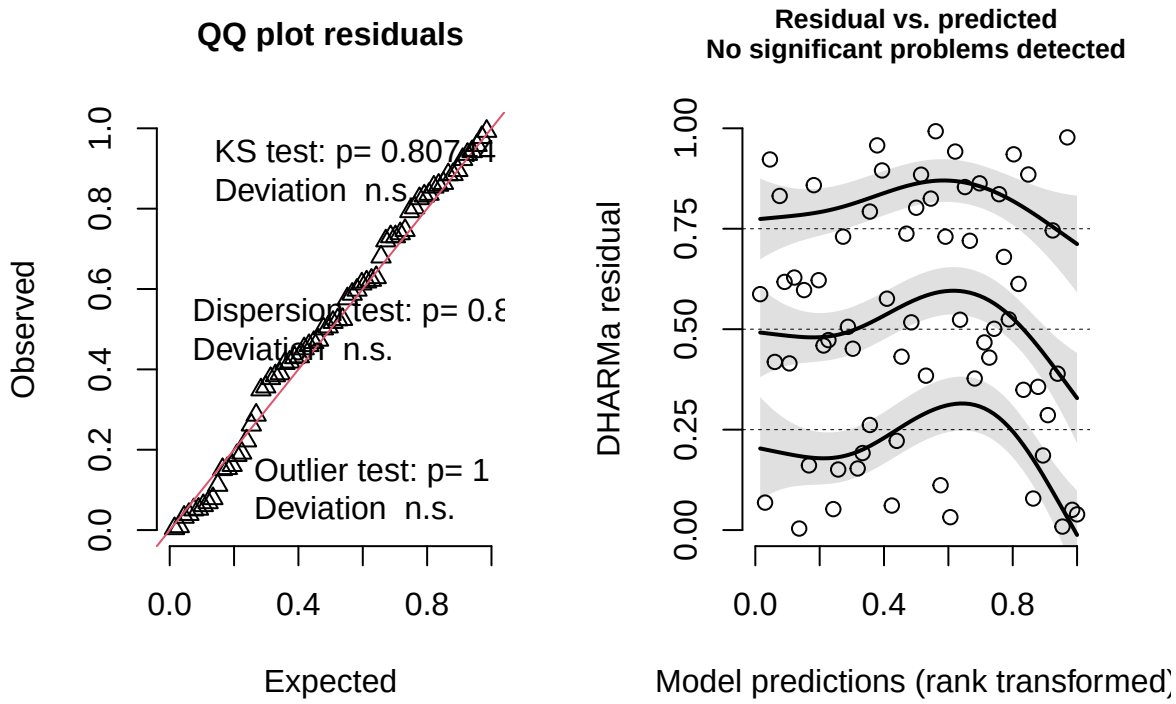
The following 2 plots visualise the effects of weight and migration date on survival of juvenile smolt.


Relationship of weight and survival status
controlling for date of migration
Prob(Survived)
1.00
0.75
0.50
0.25
0.00
10
20
30
40
Weight(g)



Model 2

Model 2 was fit using the glm() function and assessed using the pregibbon test which resulted in a test p-value of 0.1219. The model assumptions were assessed using the DHARMa package producing the following diagnostic plots




```
## Object of Class DHARMa with simulated residuals based on 250 simulations with refit = FALSE . See ?DH
##
## Scaled residual values: 0.6285682 0.4728647 0.505839 0.384296 0.3565195 0.003808071 0.06841269 0.587
```

As you can see from the diagnostic plots above, the QQ residuals plot seems to be quite good but the residuals vs predicted plot shows some curvature in the residuals, especially at he 0.25 level. This could signify a missing quadratic effect in the model.

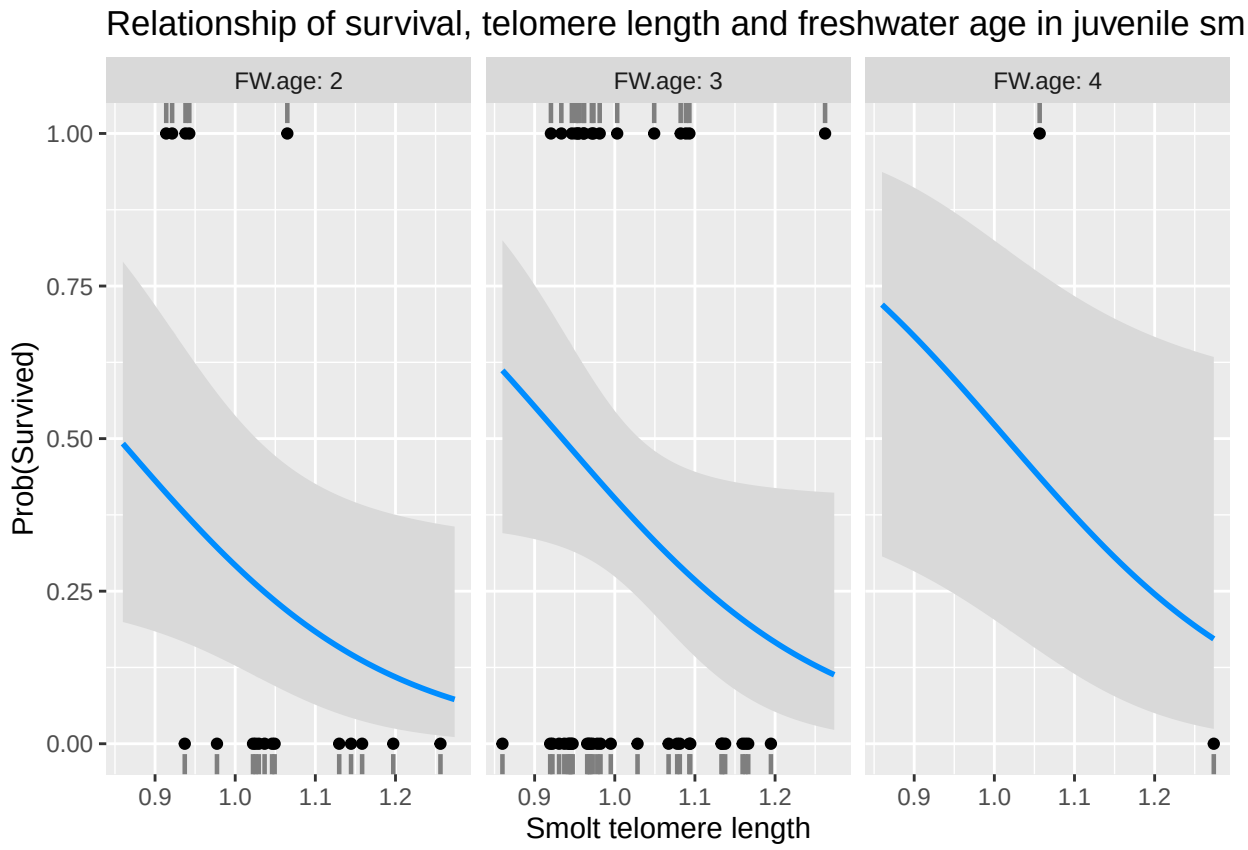
The summary for model 2 is as follows

```
##
## Call:
## glm(formula = Survival.Binary ~ FW.age + Smolt.RTL, family = "binomial",
##     data = Salmon2.clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3757  -0.9666  -0.6946   1.2416   2.0586
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.2173     3.4798   1.212   0.2255
## FW.age        0.4873     0.5734   0.850   0.3954
## Smolt.RTL    -6.0759     3.1247  -1.944   0.0518 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 85.338  on 65  degrees of freedom
## Residual deviance: 80.341  on 63  degrees of freedom
## AIC: 86.341
##
## Number of Fisher Scoring iterations: 3
```

Interpreting the above model summary output we can see that for every extra year spent in freshwater we see a 0.49 increase in our log odds of survival and for every unit increase in telomere length we see a decrease of 6.08 in our log odds of survival.

The following plots show the effect of telomere length on survival of juvenile smolt with freshwater ages 2, 3 and 4 years.



Problems Encountered

When first examining the data it was clear that there was some incomplete data contained in the second dataset which had to be removed from the analysis. This combined with the outlier that was removed reduced the sample size to 65 which was far less than the sample size of the first dataset, suggesting that the confidence in the results of model 1 would be greater. The smaller sample size also meant that there were only two fish with freshwater age of 4 years included in the observations which again would lower our confidence in the results of the model. As with the first assignment, there are small effect sizes associated with some of the model coefficients which makes results harder to interpret.

Appendix: All code for this analysis

```
knitr::opts_chunk$set(echo = TRUE)
#load packages
library(readr)
library(MASS)
library(LDdiag)
library(tidyr)
library(DHARMa)
library(dplyr)
library(ggplot2)
library(car)
library(visreg)

Salmon1 <- read_csv("Salmon1.csv")
Salmon2 <- read_csv("Salmon2.csv")

#explore dataset "Salmon1"
plot(Salmon1$Julian.Date, Salmon1$Survival.Binary)
plot(Salmon1$Weight..g., Salmon1$Survival.Binary)
hist(Salmon1$Julian.Date)
hist(Salmon1$Weight..g.)
boxplot(Salmon1$Weight..g.)
summary(Salmon1)

Salmon1.clean <- Salmon1[-715, ]
Salmon1.clean$Weight.range <- with(Salmon1.clean,
                                   ifelse(Salmon1.clean$Weight..g. > 30, "30g to 45g",
                                          ifelse(Salmon1.clean$Weight..g. <= 15, "15g or under", "15g t

#fit model 1
m1 <- glm(Survival.Binary ~ Julian.Date + Weight..g., data = Salmon1.clean, family = "binomial")
pregibon.glm(m1)
summary(m1)
simulateResiduals(m1,plot=T)

#confidence intervals for model 1
explanatory_data1 <- expand_grid(
  Julian.Date = seq(min(Salmon1.clean$Julian.Date), max(Salmon1.clean$Julian.Date), 1),
  Weight..g. = seq(5,45, 0.5)
)

prediction_data1 <- explanatory_data1 %>%
  mutate(
    Survival.Binary = predict.glm(m1, newdata = explanatory_data1, type = "response")
  )


coeffs1 <- coef(m1)

#explore dataset "Salmon2"
Salmon2.clean <- subset(Salmon2, Salmon2$FW.age != "not established")
Salmon2.clean$FW.age <- as.numeric(Salmon2.clean$FW.age)
```

```r
Salmon2.clean <- Salmon2.clean[-43,]
plot(Salmon2.clean$FW.age,Salmon2.clean$Survival.Binary)
plot(Salmon2.clean$Smolt.RTL,Salmon2.clean$Survival.Binary)


#fit model 2
m2 <- glm(Survival.Binary ~ FW.age + Smolt.RTL, data = Salmon2.clean, family = "binomial")
pregibon.glm(m2)
summary(m2)
simulateResiduals(m2,plot=T)

#confidence intervals for model 2
explanatory_data2 <- expand_grid(
  FW.age = seq(min(Salmon2.clean$FW.age), max(Salmon2.clean$FW.age), 1),
  Smolt.RTL = seq(0.85,1.3, 0.05)
)

prediction_data2 <- explanatory_data2 %>%
  mutate(
    Survival.Binary = predict.glm(m2, newdata = explanatory_data2, type = "response")
  )

# Visuals
visreg(m1, "Weight..g.",
       gg = TRUE,
       scale="response") +
  labs(y = "Prob(Survived)",
       x = "Weight(g)",
       title = "Relationship of weight and survival status",
       subtitle = "controlling for date of migration") +
  geom_point(data = Salmon1.clean, aes(Weight..g., Survival.Binary))

visreg(m1, "Julian.Date",
       gg = TRUE,
       scale="response") +
  labs(y = "Prob(Survived)",
       x = "Julian date of migration",
       title = "Relationship of date of migration and survival status",
       subtitle = "controlling for weight") +
  geom_point(data = Salmon1.clean, aes(Julian.Date, Survival.Binary, color = Weight.range))


visreg(m2, "Smolt.RTL",
       by = "FW.age",
       gg = TRUE,
       scale="response") +
  labs(y = "Prob(Survived)",
       x = "Smolt telomere length",
       title = "Relationship of survival, telomere length and freshwater age in juvenile smolt ") +
  geom_point(data = Salmon2.clean, aes(Smolt.RTL, Survival.Binary))
```

```r
simulateResiduals(m1,plot=T)
summary(m1)
visreg(m1, "Weight..g.",
       gg = TRUE,
       scale="response") +
  labs(y = "Prob(Survived)",
       x = "Weight(g)",
       title = "Relationship of weight and survival status",
       subtitle = "controlling for date of migration") +
  geom_point(data = Salmon1.clean, aes(Weight..g., Survival.Binary))

visreg(m1, "Julian.Date",
       gg = TRUE,
       scale="response") +
  labs(y = "Prob(Survived)",
       x = "Julian date of migration",
       title = "Relationship of date of migration and survival status",
       subtitle = "controlling for weight") +
  geom_point(data = Salmon1.clean, aes(Julian.Date, Survival.Binary, color = Weight.range))
simulateResiduals(m2,plot=T)
summary(m2)
visreg(m2, "Smolt.RTL",
       by = "FW.age",
       gg = TRUE,
       scale="response") +
  labs(y = "Prob(Survived)",
       x = "Smolt telomere length",
       title = "Relationship of survival, telomere length and freshwater age in juvenile smolt ") +
  geom_point(data = Salmon2.clean, aes(Smolt.RTL, Survival.Binary))
```