Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

HOMEWORK 3: Logistic regression

WSCI 6390 – 002: Population Parameter Estimation

Due 11:59 PM Tuesday, February 20

Let’s practice building a logistic regression to evaluate what factors are associated with polar bear (*Ursus maritimus*) survival.

INSTRUCTIONS:

We will be using a dataset “PB\_data.csv”. In the late 1970s, Dr. Ian Stirling, the “Father” of polar bear research, began studying polar bears in Western Hudson Bay, Canada. He and colleagues were early adopters of satellite tracking and in 2000 began placing satellite tags on female polar bears in Western Hudson Bay. Tags are reliable up to about 1 year, although some go offline before a year is up and cannot be monitored further. For tags that stayed online, based on a mortality switch in the tags, biologists are able to determine whether the animal is alive or dead from the signal type. We have obtained a dataset of tags that were deployed in spring between 2000 and 2017 and were still online 1 year later. We have information on the state of the female one year after deployment (either alive or dead). We also have information on the age of the female at deployment (age 1, subadults; age 2, adults). Finally, we have information on the number of ice-free days (standardized, with a mean of 0 and standard deviation of 1) in Hudson Bay during the summer after tag deployment. THIS IS NOT A REAL DATASET.

Please install and load these packages before starting the assignment

install.packages("performance")

library(performance)

install.packages("see")

library(see)

install.packages("faraway")

library(faraway)

install.packages("caret")

library(caret)

install.packages("AICcmodavg")

library(AICcmodavg)

ASSIGNMENT:

1. First, read in the polar bear dataset using

polarbear <- read.csv(YOUR\_DIRECTORY/PB\_data.csv”, header=T)

Then, let’s make age a factor (rather than a continuous variable) using

polarbear$age <- as.factor(polarbear$age)

Then use

head(polarbear)

to see what the top six rows of the dataset look like.

**Why do the values for number of ice-free days (“ice”) have negative values? Clue is in the instructions.**

**Because we scaled and centered the data, with a mean of 0 and SD of 1.**

1. Our response variable in this case is survival, which takes values of 0 or 1 for each bear in a given year. **When we have a single trial where the outcome is either 0 or 1, what is the distribution that we use?**

Bernoulli distribution

1. When we are modeling binary outcomes from a single trial we use a type of modeling called logistic regression. **What is the link function that we use when modeling 0/1 outcomes?**
   1. Logit link
2. **What does this link function do?**
   1. **Connects the mean of our response and a linear model**
   2. **Allows us to take data without normally distributed errors and fit a linear model**
3. We suspect that there are effects of the number of ice-free days and age on annual survival. Build a model that reflects this hypothesis using:

survmod1 <- glm(survival ~ ice + age, data = polarbear, family = "binomial")

Use

summary(survmod1)

to see the coefficient table. **Paste the coefficient table here.**

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 2.3684 0.3005 7.882 3.23e-15 \*\*\*

ice -0.8021 0.2785 -2.880 0.00398 \*\*

age2 0.8497 0.4628 1.836 0.06638 .

1. **What are the relationships between polar bear survival and (1) number of ice-free days and (2) age? I am looking for general direction here (positive and negative associations, nothing more detailed). Remember that because “age” is a factor, “age2” represents the *difference* between age1 and age2.** 
   1. **Ice free days have a negative relationship on survival**
   2. **Age has a positive relationship on survival**
2. Remember that coefficients in logistic regression represent log-odds (essentially a “log of the odds”) and therefore aren’t so helpful in their current format. To convert these coefficients to something we can understand, like regular “odds,” use

exp(survmod1$coefficients["ice"])

exp(survmod1$coefficients["age2"])

**Can you tell me what these two coefficients represent in terms of odds? The interpretation for “ice” will be slightly different because it is standardized, so we are talking about a 1 SD increase in that covariate rather than a 1 unit increase.**

**Odds increase by 0.44 for every 1 SD increase in ice free days**

**Odds increase by 2.34 from age 1 to age 2**

1. Create binned residual plots using

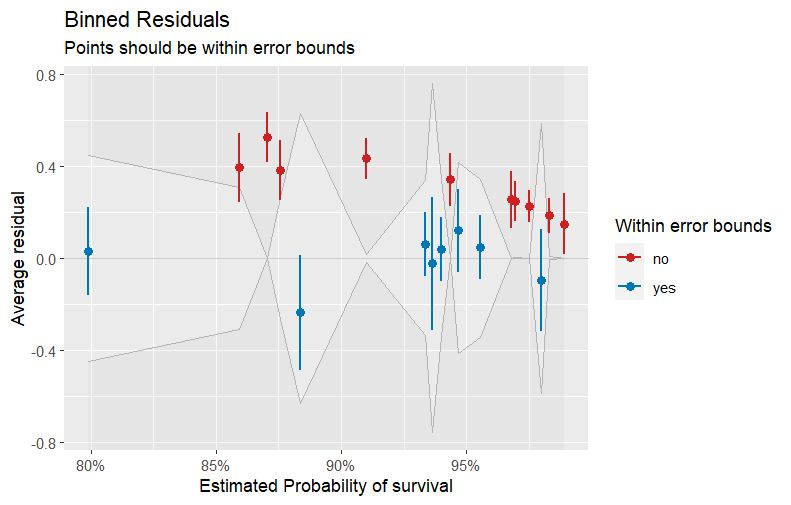
binned\_residuals(survmod1)

**What does the message say?**

|  |
| --- |
| Warning: Probably bad model fit. Only about 44% of the residuals are inside the error  bounds. |
|  |
| |  | | --- | |  | |

1. **Paste what the binned residual plot looks like using**

plot(binned\_residuals(survmod1))



1. Create half-normal plots of Cook’s distance (to look for unusual points or unexplained variation) using

halfnorm(cooks <- cooks.distance(survmod1))

**What do you notice here?**

**188 and 9 require further investigation**

1. Let’s see if this model did a good job of predicting the values in our dataset. In an ideal world we would have withheld 20% of our data to see how well the model predicts to the withheld data. In this case, we will use the original dataset to see what the model predicted.

#this line makes predictions from our survival model to our age and ice #data

prediction<-predict(survmod1, newdata=data.frame(age = polarbear$age, ice = polarbear$ice), type="response")

#we will say the model predicts survival if the predicted outcome is #>0.5 and predicts a death otherwise

binary\_outcome <- ifelse(prediction > 0.5, 1, 0)

#this line creates a confusion matrix. “Reference” refers to our #original data, and “Prediction” refers to our model predictions. We #really want to see all of our values in the diagonal, such that all 0s #and 1s in our dataset were correctly classified by the model

confusionMatrix(as.factor(binary\_outcome), as.factor(polarbear$survival))

**Ignore the warning message. What does the model tell us about how well our model predicts to our data?**

**Fairly accurate with 92%**

1. We think that number of ice-free days and age affect survival of polar bears, but we aren’t too sure. Let’s build four competing models

#model with intercept only

survmod1 <- glm(survival ~ 1 , data = polarbear, family = "binomial")

#model with age only

survmod2 <- glm(survival ~ age, data = polarbear, family = "binomial")

#model with ice-free days only

survmod3 <- glm(survival ~ ice , data = polarbear, family = "binomial")

#model with age and ice-free days

survmod4 <- glm(survival ~ age + ice , data = polarbear, family = "binomial")

#define list of models

models <- list(survmod1, survmod2, survmod3, survmod4)

#specify model names

mod.names <- c("intercept.only", "age", "ice", "age.ice")

#calculate AICc of each model

aictab(cand.set = models, modnames = mod.names, second.ord = T)

**Paste the model selection table below.**

Model selection based on AICc:

K AICc Delta\_AICc AICcWt Cum.Wt LL

age.ice 3 158.23 0.00 0.67 0.67 -76.07

ice 2 159.74 1.51 0.31 0.98 -77.85

age 2 166.22 7.99 0.01 0.99 -81.09

intercept.only 1 166.52 8.30 0.01 1.00 -82.26

1. **Which model(s) is/are best supported?**
   1. **Age.ice and ice**
2. Let’s pretend that the model with number of ice-free days only was our best model. Can you please make predictions from your best model?

#First, because we are interested in survival of polar bears across all #possible values of number of ice-free days, let’s calculate a series #of 100 values from the minimum to maximum number of ice-free days in #our data

ice\_values <- data.frame(ice = seq(min(polarbear$ice), max(polarbear$ice), length.out=100))

#then let’s make predictions on the logit scale (type="link”)

predictions <- predict(survmod3, newdata=ice\_values, type="link", se.fit=TRUE)

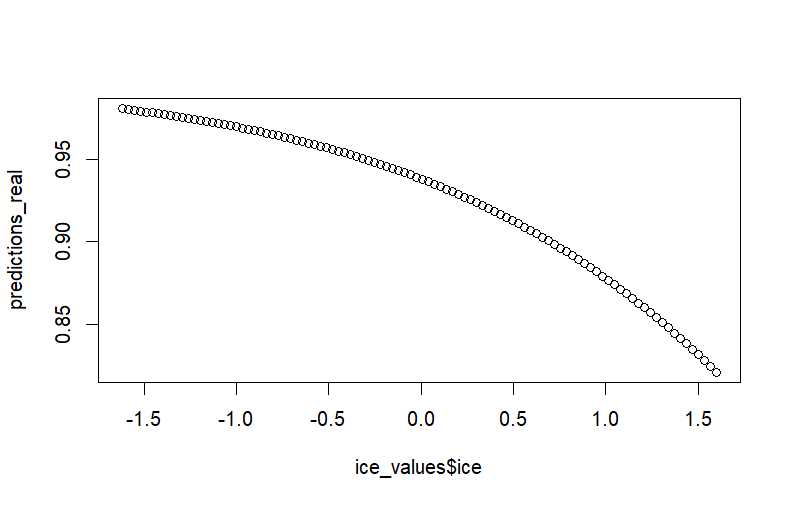
#now we can use the 1/1+exp(-PREDICTIONS) equation to convert these #values back to the 0-1 scale.

predictions\_real <- 1/(1+exp(-(predictions$fit)))

#then we can make a really basic plot of those predictions:

plot(x=ice\_values$ice, y=predictions\_real)

**Paste your prediction plot here.**



1. **From this graph, what can we broadly say about the impact of number of ice-free days on polar bear survival? (A note that you should always add the lower and upper confidence bounds. These code lines are in the slides. I didn’t do that here to keep things manageable for this homework assignment).**

**As the number of ice free days increases, survival will decline**