CUNY SPS DATA 622 - Machine Learning and Big Data

Spring 2021 - Home Work #1

Samantha Deokinanan

February 15, 2021

Problem

1. Logistic Regression with a binary outcome.

- a. The penguin dataset has a 'species' column. Please check how many categories you have in the species column. Conduct whatever data manipulation you need to do to be able to build a logistic regression with a binary outcome. Please explain your reasoning behind your decision as you manipulate the outcome/dependent variable (species).
- b. Please make sure you are evaluating the independent variables appropriately in deciding which ones should be in the model.
- c. Provide variable interpretations in your model.

2. For your model from #1, please provide: AUC, Accuracy, TPR, FPR, TNR, FNR

3. Multinomial Logistic Regression.

- a. Please fit it into a multinomial logistic regression where your outcome variable is 'species'.
- b. Please be sure to evaluate the independent variables appropriately to fit your best parsimonious model.
- c. Please be sure to interpret your variables in the model.

4. What would be some of the fit statistics you would want to evaluate for your model in question #3? Feel free to share whatever you can provide.

R Packages

The statistical tool that will be used to fascinate in the modeling of the data was `R`. The main packages used for data wrangling, visualization, and graphics were listed below. Any other minor packages for analysis will be listed when needed.

```
# Required R packages
library(palmerpenguins)
library(tidyverse)
library(kableExtra)
library(summarytools)
library(psych)
library(GGally)
library(mice)
library(nnet)
library(effects)
library(caret)
library(precrec)
library(DescTools)
```

Data Exploration

The `palmerpenguins` data contains size measurements collected from 2007 - 2009 for three penguin species observed on three islands in the Palmer Archipelago, Antarctica. For more information about this data collection, refer to palmerpenguins website. (https://allisonhorst.github.io/palmerpenguins/articles/intro.html)

Penguins Data Column Definition

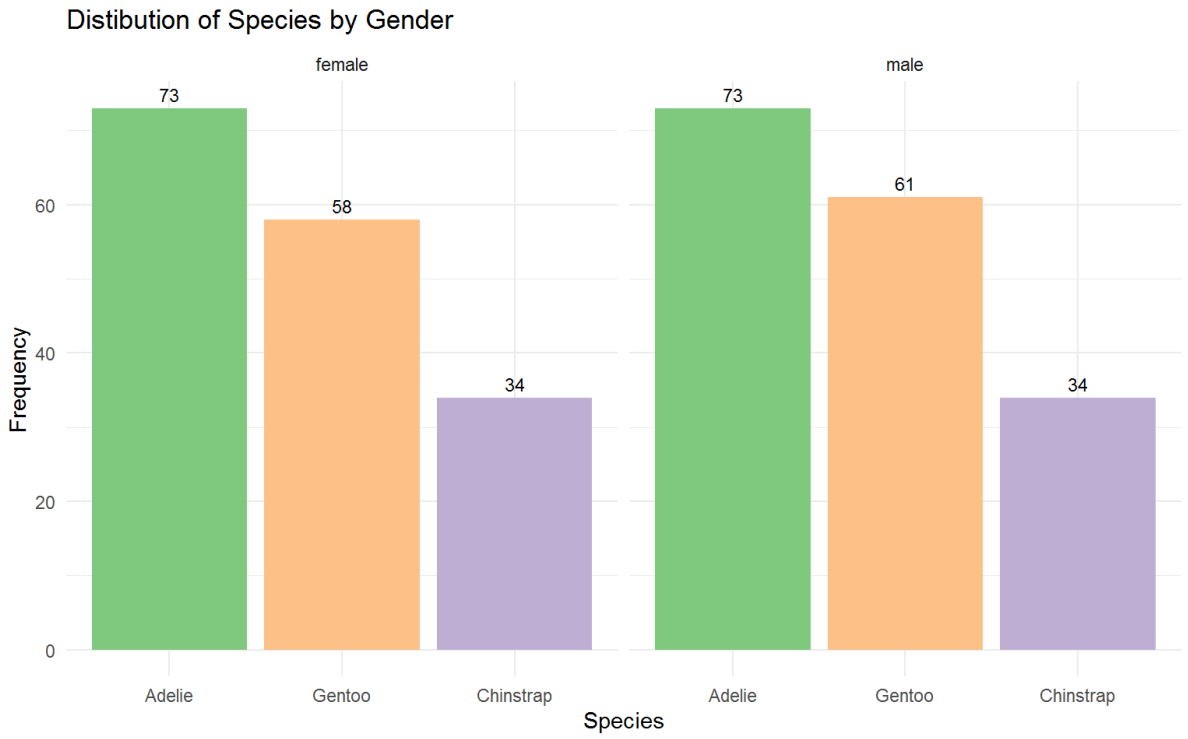
Variable	Description
species	penguin species (Adélie, Chinstrap, and Gentoo)
island	island in Palmer Archipelago, Antarctica (Biscoe, Dream or Torgersen)
bill_length_mm	bill length (millimeters)
bill_depth_mm	bill depth (millimeters)
flipper_length_mm	flipper length (millimeters)
body_mass_g	body mass (grams)
sex	penguin sex (female, male)
year	year data was collected

```
# Load dataset
penguins = penguins

# Number of observations
ntrobs = dim(penguins)[[1]]

# Converting Year to factor
penguins$year = as.factor(penguins$year)
```

Target Variable (Species)

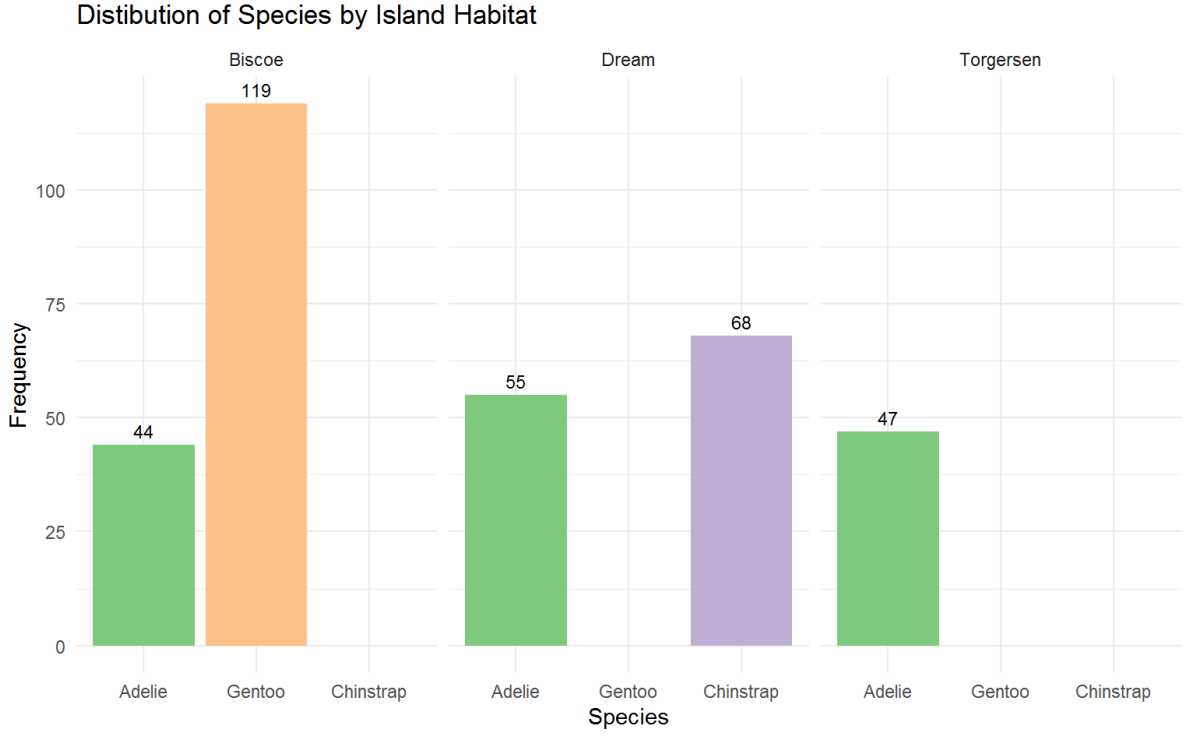
The response variable, `species` denotes one of three penguin species, namely Adélie, Chinstrap, and Gentoo. From the bar plot below, a majority of the penguins are Adelie (n = 153), followed by Gentoo (n = 124) and Chinstrap (n = 68). The distribution between gender is also nearly equally divided among the species.

```
reorder <- function(x){
  factor(x, levels = names(sort(table(x), decreasing = TRUE)))
}

ggplot(drop_na(penguins), aes(x = reorder(species), fill = species)) +
  geom_bar() +
  geom_text(stat = "count", aes(label =..count..), vjust=-0.5, size = 3) +
  facet_wrap(~sex) +
  scale_fill_brewer(palette = "Accent") +
  theme_minimal() +
  theme(legend.position = "none")+
  labs(title = "Distibution of Species by Gender", y = "Frequency", x = "Species")
```



However, there is not an equal distribution for their island habitat since it seems that some species do not reside on some islands. For instance, no Chinstrap and Gentoo were recorded from the island of Torgersen.

```
ggplot(drop_na(penguins), aes(x = reorder(species), fill = species)) +
  geom_bar() +
  geom_text(stat = "count", aes(label =..count..), vjust=-0.5, size = 3) +
  facet_wrap(~island) +
  scale_fill_brewer(palette = "Accent") +
  theme_minimal() +
  theme(legend.position = "none")+
  labs(title = "Distibution of Species by Island Habitat", y = "Frequency", x = "Species")
```



Predictive Variables

Summary Statistic

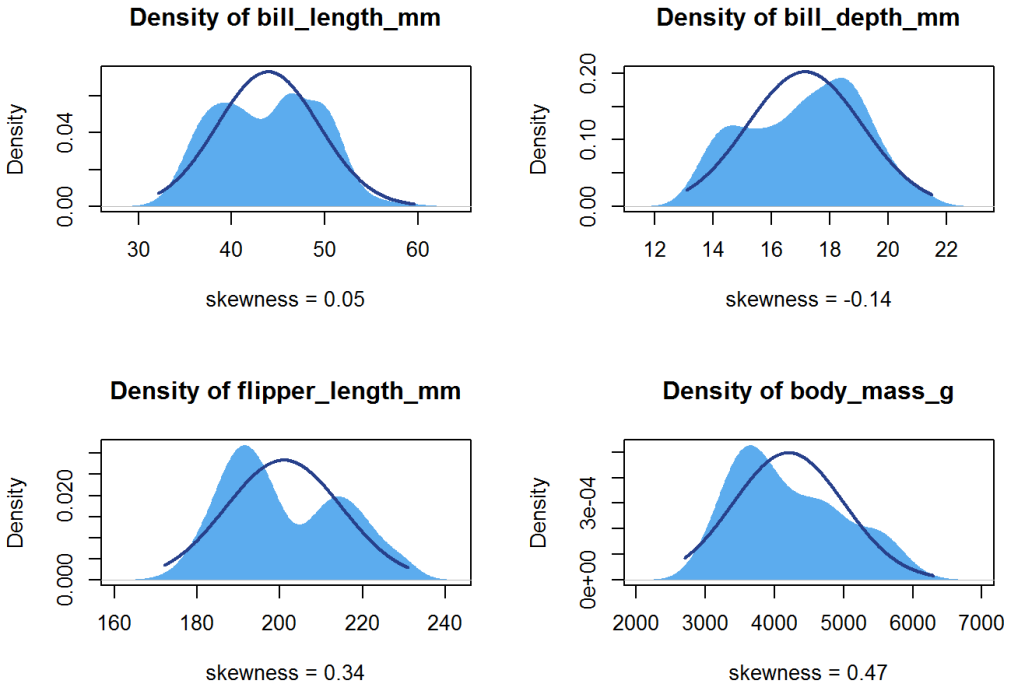
Based on the summary statistic for the species, some initial observations can be made. There were 344 observations of 4 numeric predictor variables and 2-factor predictor variables, namely `island` , and `sex` . There is also a `year` variable in which this analysis is denoted as a factor variable. The data set did not have complete cases, thus, there was a need for imputation.

```
dfSummary(penguins, plain.ascii = TRUE, style = "grid", graph.col = FALSE, footnote = NA)
```

```
## Data Frame Summary
## penguins
## Dimensions: 344 x 8
## Duplicates: 0
##
## +----+--------------------+---------------------------+---------------------+----------+---------+
## | No | Variable           | Stats / Values            | Freqs (% of Valid)  | Valid    | Missing |
## +====+====================+===========================+=====================+==========+=========+
## | 1  | species            | 1. Adelie                 | 152 (44.2%)         | 344      | 0       |
## |    | [factor]           | 2. Chinstrap              |  68 (19.8%)         | (100%)   | (0%)    |
## |    |                    | 3. Gentoo                 | 124 (36.0%)         |          |         |
## +----+--------------------+---------------------------+---------------------+----------+---------+
## | 2  | island             | 1. Biscoe                 | 168 (48.8%)         | 344      | 0       |
## |    | [factor]           | 2. Dream                  | 124 (36.0%)         | (100%)   | (0%)    |
## |    |                    | 3. Torgersen              |  52 (15.1%)         |          |         |
## +----+--------------------+---------------------------+---------------------+----------+---------+
## | 3  | bill_length_mm     | Mean (sd) : 43.9 (5.5)    | 164 distinct values | 342      | 2       |
## |    | [numeric]          | min < med < max:          |                     | (99.42%) | (0.58%) |
## |    |                    | 32.1 < 44.5 < 59.6        |                     |          |         |
## |    |                    | IQR (CV) : 9.3 (0.1)      |                     |          |         |
## +----+--------------------+---------------------------+---------------------+----------+---------+
## | 4  | bill_depth_mm      | Mean (sd) : 17.2 (2)      | 80 distinct values  | 342      | 2       |
## |    | [numeric]          | min < med < max:          |                     | (99.42%) | (0.58%) |
## |    |                    | 13.1 < 17.3 < 21.5        |                     |          |         |
## |    |                    | IQR (CV) : 3.1 (0.1)      |                     |          |         |
## +----+--------------------+---------------------------+---------------------+----------+---------+
## | 5  | flipper_length_mm  | Mean (sd) : 200.9 (14.1)  | 55 distinct values  | 342      | 2       |
## |    | [integer]          | min < med < max:          |                     | (99.42%) | (0.58%) |
## |    |                    | 172 < 197 < 231           |                     |          |         |
## |    |                    | IQR (CV) : 23 (0.1)       |                     |          |         |
## +----+--------------------+---------------------------+---------------------+----------+---------+
## | 6  | body_mass_g        | Mean (sd) : 4201.8 (802)  | 94 distinct values  | 342      | 2       |
## |    | [integer]          | min < med < max:          |                     | (99.42%) | (0.58%) |
## |    |                    | 2700 < 4050 < 6300        |                     |          |         |
## |    |                    | IQR (CV) : 1200 (0.2)     |                     |          |         |
## +----+--------------------+---------------------------+---------------------+----------+---------+
## | 7  | sex                | 1. female                 | 165 (49.5%)         | 333      | 11      |
## |    | [factor]           | 2. male                   | 168 (50.4%)         | (96.8%)  | (3.2%)  |
## +----+--------------------+---------------------------+---------------------+----------+---------+
## | 8  | year               | 1. 2007                   | 110 (32.0%)         | 344      | 0       |
## |    | [factor]           | 2. 2008                   | 114 (33.1%)         | (100%)   | (0%)    |
## |    |                    | 3. 2009                   | 120 (34.9%)         |          |         |
## +----+--------------------+---------------------------+---------------------+----------+---------+
```

Moreover, the plots below represent a density plot for a vector of values and a superimposed normal curve with the same mean and standard deviation. The plot can be used to quickly compare the distribution of data to a normal distribution. It is evident that no variables are truly normally distributed. The presence of bi- and tri-modal distributions suggest that there are differences among the penguin species.

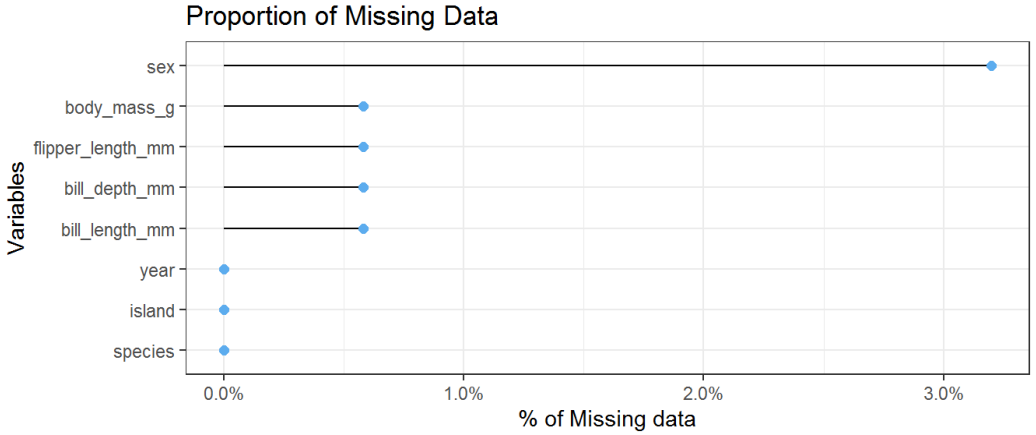
```
par(mfrow = c(2,2))
for (i in 3:6){
  rcompanion::plotNormalDensity(
    penguins[,i], main = sprintf("Density of %s", names(penguins)[i]),
    xlab = sprintf("skewness = %1.2f", psych::describe(penguins)[i,11]),
    col2 = "steelblue2", col3 = "royalblue4")
}
```



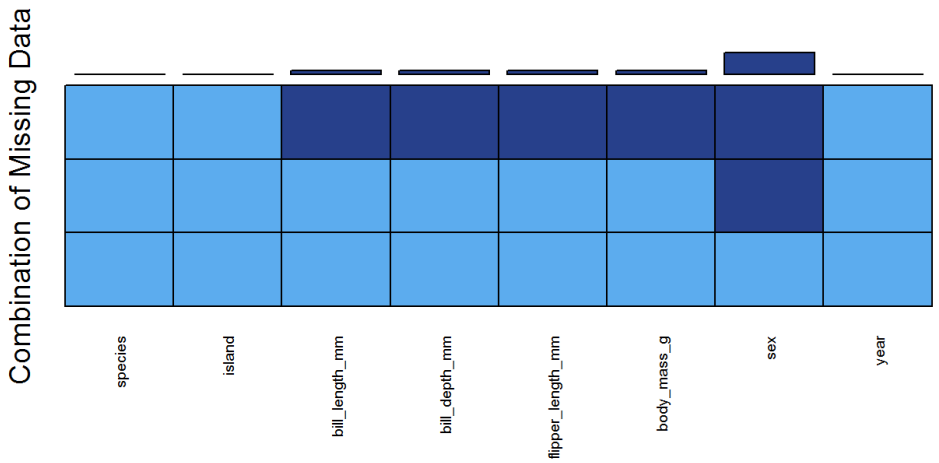
Missing Data

The graph below indicates the amount of missing data the penguin data contains. It appears that more than 3% of the missing data was from the `sex` variable. This further suggests that nearly 97% were complete. There were no missingness patterns, and their overall proportion was not very extreme. As a result, missingness can be corrected by imputation.

```
na.counts = as.data.frame(((sapply(penguins,
                                   function(x) sum(is.na(x))))/nrow(penguins))*100)
names(na.counts) = "counts"
na.counts = cbind(variables = rownames(na.counts),
                  data.frame(na.counts, row.names = NULL))

na.counts %>% arrange(counts) %>%
  mutate(name = factor(variables, levels = variables)) %>%
  ggplot(aes(x = name, y = counts)) + geom_segment( aes(xend = name, yend = 0)) +
  geom_point(size = 2, color = "steelblue2") + coord_flip() + theme_bw() +
  labs(title = "Proportion of Missing Data", x = "Variables", y = "% of Missing data") +
  scale_y_continuous(labels = scales::percent_format(scale = 1))
```



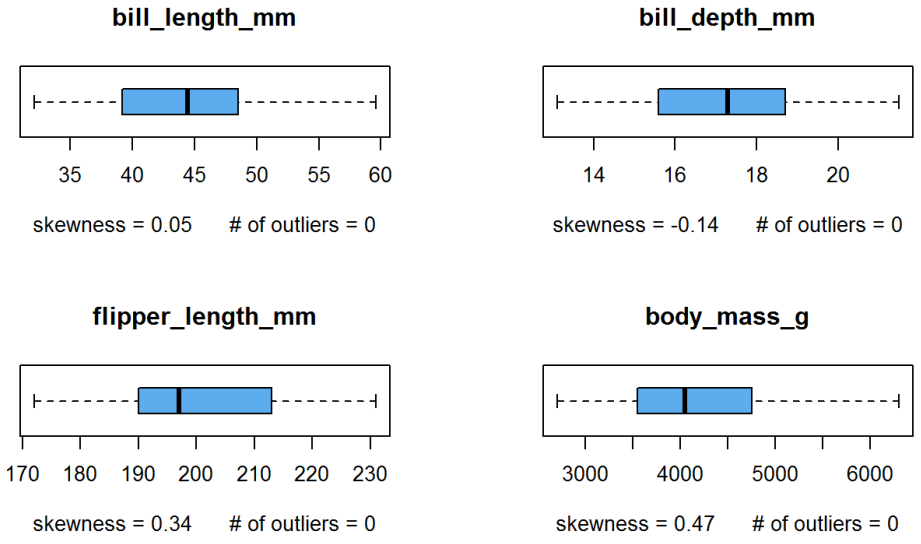

```
VIM::aggr(penguins, col = c('steelblue2','royalblue4'), numbers = FALSE,
          sortVars = FALSE, oma = c(6,4,3,2), labels = names(penguins),
          cex.axis = 0.6, axes = TRUE, bars = FALSE, combined = TRUE,
          Prop = TRUE, ylab = c("Combination of Missing Data"))
```



Outlier

An outlier is an observation that lies an abnormal distance from other values in a random sample. Outliers in the data could distort predictions and affect the accuracy, therefore, these would need to be corrected. However, further exploration revealed that no variable seems to be strongly influenced by any outliers.

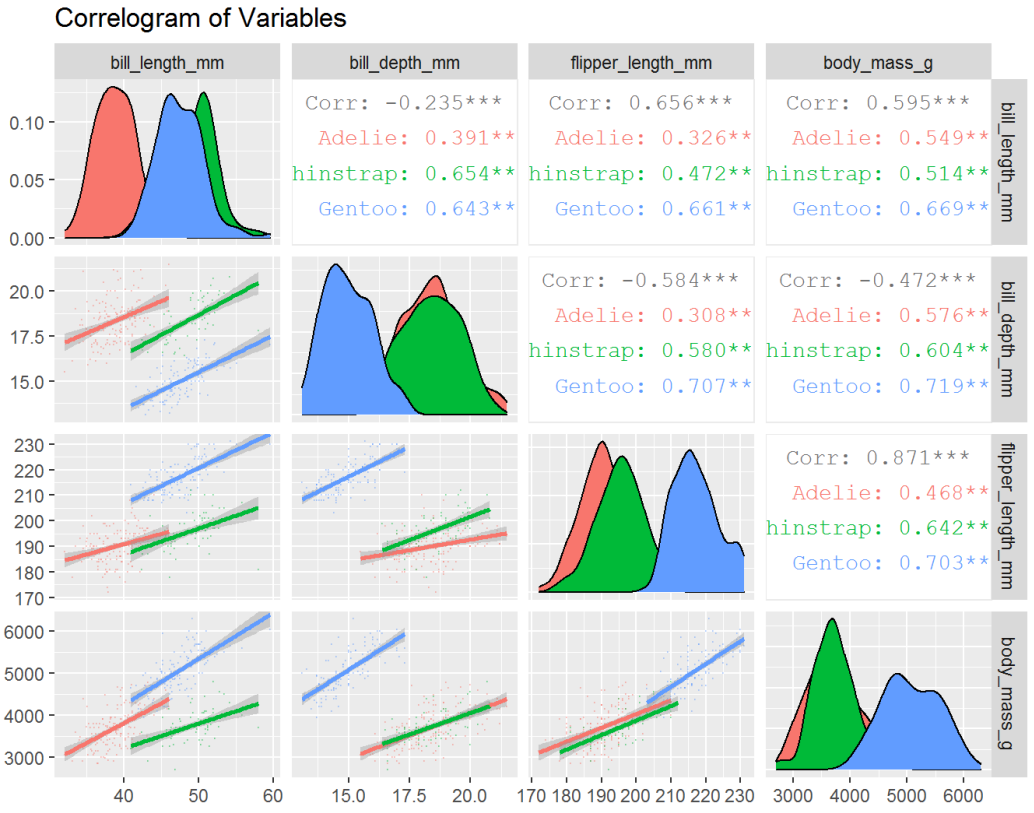
```
par(mfrow = c(2,2))
for (i in 3:6){
  boxplot(
    penguins[i], main = sprintf("%s", names(penguins)[i]),
    col = "steelblue2", horizontal = TRUE,
    xlab = sprintf("skewness = %1.2f      # of outliers = %d",
                   psych::describe(penguins)[i,11],
                   length(boxplot(penguins[i], plot = FALSE)$out)))
}
```



Correlation

The correlogram below graphically represents the correlations between the numeric predictor variables, when ignoring the missing variables. Most of the numeric variables were uncorrelated with one another, but there were a few highly correlated pairs. From the correlogram, the relationship between the `body_mass_g` and `flipper_length_mm` is a highly positive correlation, and within reason, as larger flippers would indicate an increase in body mass. There are some variables with moderate correlations, but their relationship is also intuitive.

```
ggpairs(penguins, columns = 3:6, title = "Correlogram of Variables",
        ggplot2::aes(color = species),
        progress = FALSE,
        lower = list(continuous = wrap("smooth", alpha = 0.3, size = 0.1)))
```



To build a smaller model without predictors with extremely high correlations, it is best to reduce the number of predictors such that there were no absolute pairwise correlations above 0.90. However, no relationship was too extreme, and instead, their interactions are analyzed. The graphic reveals how the predictor variables are distributed by species. Interestingly, Adelie and Chinstrap overlap for all variable measurements except bill length. This feature may be the definitive variable that produces complete separation among the penguin species into groups.

Data Preparation

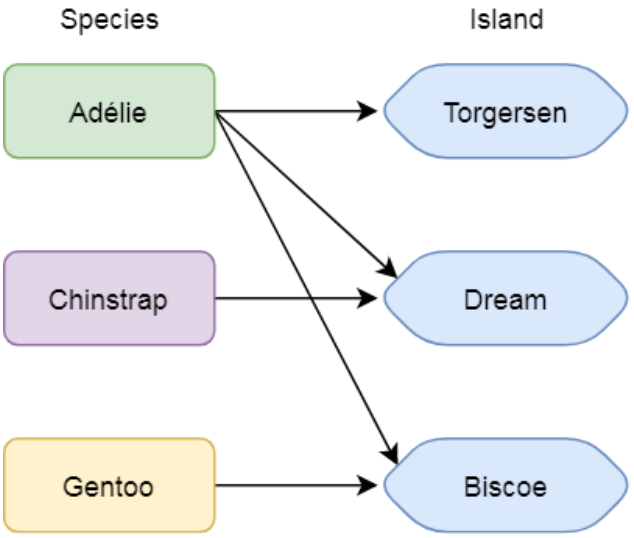
Binomial Target Variable

The first desired model is a logistic regression with the binary outcome for the target variable `species` (Problem 1a). However, this variable has three-factor levels and needs to be transformed reasonably into two levels. Considering the table below shows the frequency of penguin species based on their island habitat.

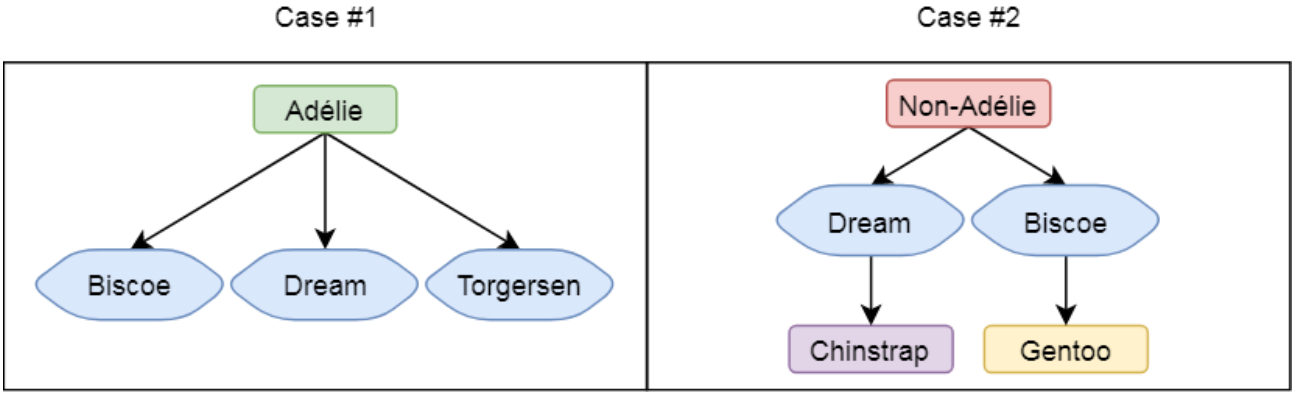
Proportion of Species by Island location

	Biscoe	Dream	Torgersen
Adelie	44	56	52
Chinstrap	0	68	0
Gentoo	124	0	0

Adelie penguins are found at the three islands, whereas Chinstrap is found only on Dream while Gentoo is found only on Biscoe. With this relationship in mind, a two-factor level of penguin species can be derived based on whether the species is Adelie or not.

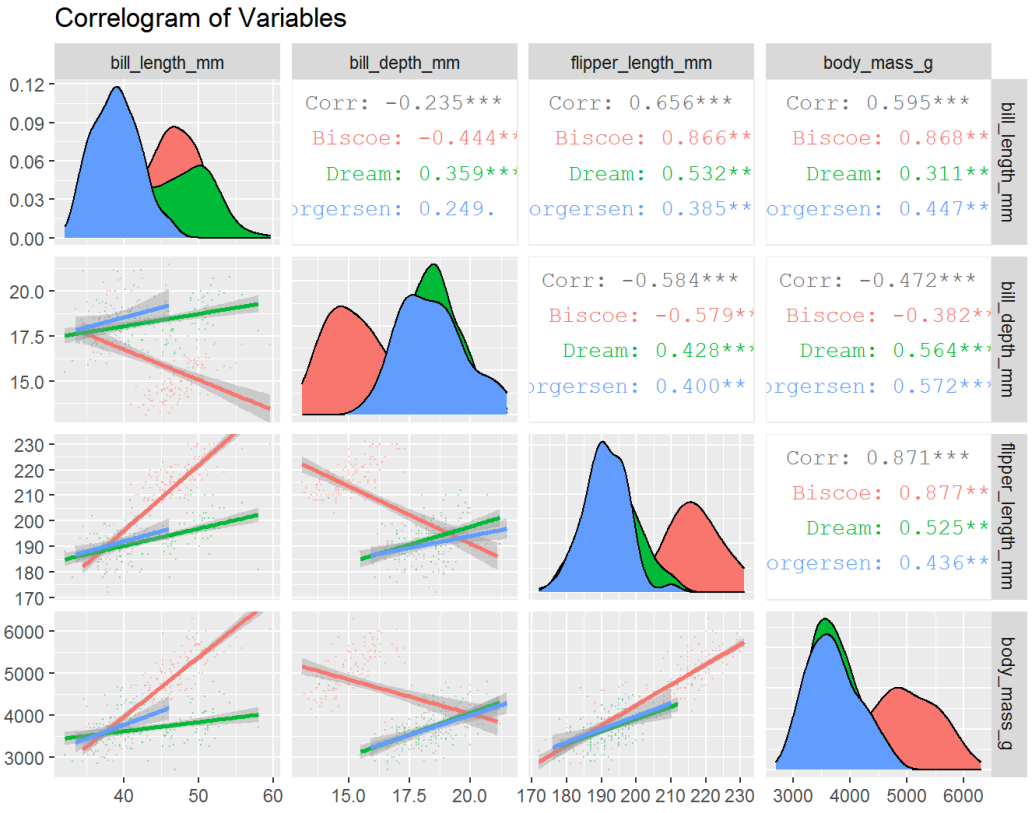


In other words, based on the island habitat of a penguin, the species can be deduced. For instance, if a penguin is from the island of Torgersen, it is more, if not absolutely, likely to be an Adelie. Whereas, if a penguin is from the island of Biscoe, it is either an Adelie or a Chinstrap, and if the penguin is non-Adelie, then it can be stated confidently that it is a Gentoo penguin.



Moreover, from the correlogram, distinguished by island, differences can already be seen in measurements particularly for bill length and depth. Thus, for these reasons, Problem #1 is solved where the target variable is transformed into two factors, namely `Adelie` and `NonAdelie`.

```
ggpairs(penguins, columns = 3:6, title = "Correlogram of Variables",
        ggplot2::aes(color = island),
        progress = FALSE,
        lower = list(continuous = wrap("smooth", alpha = 0.3, size = 0.1)))
```

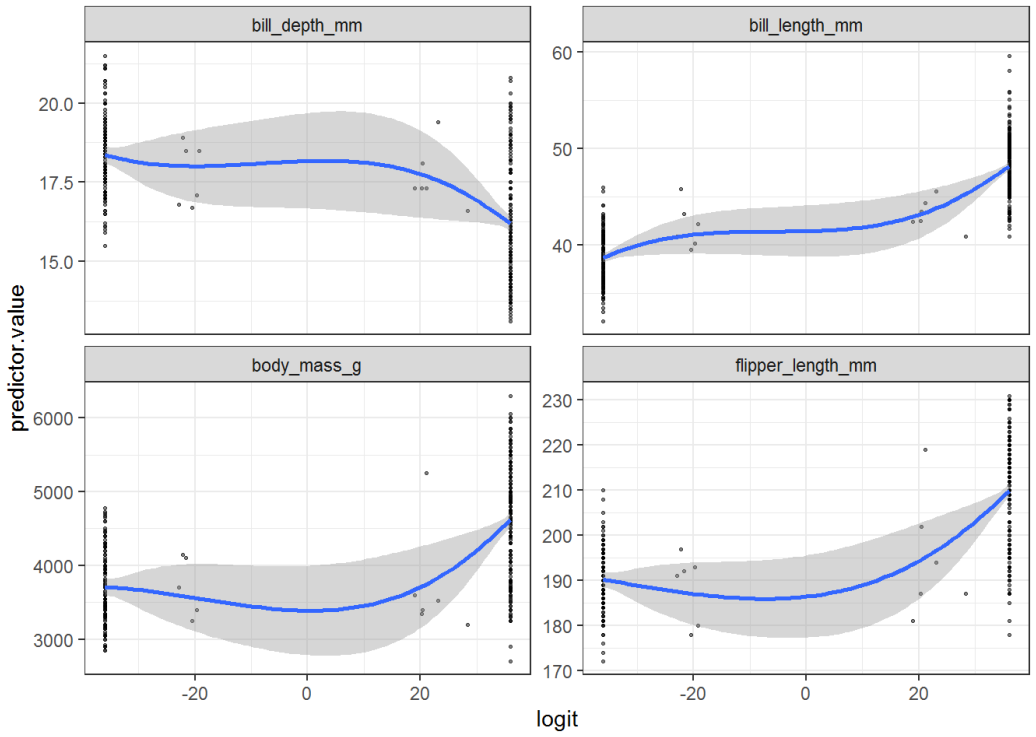



```
penguins$target_adelie = penguins$species
levels(penguins$target_adelie)[levels(penguins$target_adelie) != "Adelie"] <- "NonAdelie"
```

Normality & Linearity

Logistic regression does not assumptions regarding normality based on ordinary least squares algorithms. As such, it is not required, and the residuals do not need to be normally distributed. The smoothed scatter plots show that there is a separation relationship between continuous predictor variables and the logit of the target variable.

```
set.seed(525)
df = na.omit(penguins)

# Fit the logistic regression model
model = glm(species ~. -target_adelie, data = df, family = binomial)

# Predict the probability
probabilities = predict(model, type = "response")
predicted.classes = ifelse(probabilities > 0.5, "Adelie", "NonAdelie")

# Select numeric predictors
temp = df %>% select_if(is.numeric)
predictors = colnames(temp)

# Bind the logit and tidying the data for plot
df2 = temp %>% mutate(logit = log(probabilities/(1 - probabilities))) %>%
  gather(key = "predictors", value = "predictor.value", -logit)

ggplot(df2, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")
```



Pre-Processing of Predictors

Firstly, missing data are treated by imputation. The random forest (RF) missing data algorithm was implemented because this could handle mixed types of missing data, and adaptable to interactions and non-linearity. This would help to account for the uncertainty in the individual imputations.

```
set.seed(525)
temp = mice(penguins, method = 'rf', print = FALSE, m = 3, maxit = 3)
penguins_df = complete(temp)
```

Dummy Variables

The categorical variables are dummyfied by R as shown below. For instance, in the variable `sex`, the female will be used as the reference, whereas in the `target_adelie` variable, Adelie species will be used as the reference.

```
contrasts(penguins_df$sex)
```

```
##        male
## female    0
## male      1
```

```
contrasts(penguins_df$island)
```

```
##           Dream Torgersen
## Biscoe        0         0
## Dream         1         0
## Torgersen     0         1
```

```
contrasts(penguins_df$target_adelie)
```

```
##           NonAdelie
## Adelie            0
## NonAdelie         1
```

Training & Testing Split

The binomial and multinomial models were trained on the same approximately 70% of the data set, reserving 30% for validation of which model to select for the species class on the test set. This will allow for the test via cross-validation scheme of the models to tune parameters for optimal performance.

```
# Binomial Logistic Regression
# Create training and testing split
set.seed(525)
intrain = createDataPartition(penguins_df$species, p = 0.70, list = FALSE)

# Train & Test predictor variables
m1.train.p = penguins_df[intrain, ] %>% select(-c(species,target_adelie))
m1.test.p = penguins_df[-intrain, ] %>% select(-c(species,target_adelie))

# Train & Test response variable (Adelie or Non-Adelie)
m1.train.ra = penguins_df$target_adelie[intrain]
m1.test.ra = penguins_df$target_adelie[-intrain]
```

```
set.seed(525)
# Multinomial Logistic Regression
# Train & Test predictor variables
m2.train.p = penguins_df[intrain, ] %>% select(-c(species,target_adelie))
m2.test.p = penguins_df[-intrain, ] %>% select(-c(species,target_adelie))

# Train & Test response variable (species)
m2.train.r = penguins_df$species[intrain]
m2.test.r = penguins_df$species[-intrain]
```

Building the Models

Model 1: Binomial Logistic Regression (Adélie)

This model will be a binary logistic regression allowing for all variables and will be optimized by performing cross-validation. Given that `bill_length_mm` shows complete separation among the penguin species, this perfectly discriminating predictor will not be in this model (Problem 1b). Moreover, to work with complete separation in the logistic regression model, a Bayesian analysis is fitted. The Bayesian statistical model returns samples of the parameters of interest (the “posterior” distribution) based on some “prior” distribution which is then updated by the data. Here, the Cauchy distribution is accepted as a prior for parameters of the generalized linear model.

```
set.seed(525)
model1_adelie = train(x = m1.train.p[,-c(2)],
                      y = m1.train.ra,
                      method = "bayesglm",
                      trControl = trainControl(method = "repeatedcv",
                                               classProbs = TRUE,
                                               number = 10,
                                               summaryFunction = twoClassSummary),
                      family = binomial(link = "logit"),
                      trace = 0)
```

Binomial Logistic Regression Output

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-31.607	7.952	-3.975	0.000
islandDream	2.563	0.820	3.126	0.002
islandTorgersen	-2.733	1.590	-1.718	0.086
bill_depth_mm	-0.548	0.217	-2.523	0.012
flipper_length_mm	0.206	0.039	5.268	0.000
body_mass_g	0.000	0.001	0.282	0.778
sexmale	-0.860	0.649	-1.325	0.185
year2008	-1.144	0.597	-1.918	0.055
year2009	-1.120	0.580	-1.932	0.053

Performance Criteria

The confusion matrix is the most reliable metric commonly used to evaluate classification models (Problem 2). Following are the metrics that can be derived from a confusion matrix:

- Accuracy – the overall predicted accuracy of the model.
- True Positive Rate (TPR) – how many positive values, out of all the positive values, have been correctly predicted. It is also known as Sensitivity or Recall.
- False Positive Rate (FPR) – how many negative values, out of all the negative values, have been incorrectly predicted.
- True Negative Rate (TNR) – how many negative values, out of all the negative values, have been correctly predicted. It is also known as Specificity.
- False Negative Rate (FNR) – how many positive values, out of all the positive values, have been incorrectly predicted.
- Precision - how many values, out of all the predicted positive values, are positive.
- F-Score - the harmonic mean of precision and recall. The closer the value is to 1, the better the model.

```
set.seed(525)
# Model 1: Confusion Matrix
m1.pred.P = predict(model1_adelie, newdata = m1.test.p, type = "prob")
m1.pred.R = predict(model1_adelie, newdata = m1.test.p, type = "raw")

m1.confusion = confusionMatrix(m1.pred.R, m1.test.ra, mode = "everything")
m1.confusion
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Adelie NonAdelie
##   Adelie        41         5
##   NonAdelie      4        52
##
##                Accuracy : 0.9118
##                  95% CI : (0.8391, 0.9589)
##     No Information Rate : 0.5588
##     P-Value [Acc > NIR] : 5.166e-15
##
##                   Kappa : 0.8215
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9111
##             Specificity : 0.9123
##          Pos Pred Value : 0.8913
##          Neg Pred Value : 0.9286
##               Precision : 0.8913
##                  Recall : 0.9111
##                      F1 : 0.9011
##              Prevalence : 0.4412
##          Detection Rate : 0.4020
##    Detection Prevalence : 0.4510
##       Balanced Accuracy : 0.9117
##
##        'Positive' Class : Adelie
##
```

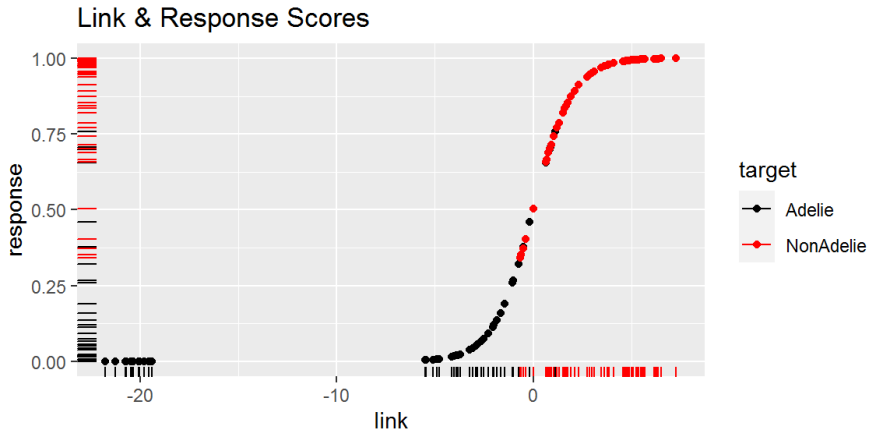
The confusion matrix results suggest that 91.2% of the predicted results seems to be correctly classified. This is impressive, even as the predictor that best explains the response, i.e. `bill_length_mm` was removed. The precision also suggests that 89.1% of the penguins belong to the actual `Adélie` species among all the penguins predicted to be `Adélie`. Moreover, the recall highlights that 91.1% of the `Adélie` species have been correctly classified as `Adélie`. These results represent that the model does a pretty good job classifying penguins into `Adélie` and `NonAdélie`. And lastly, the Kappa statistic, which is a measure of agreement between the predictions and the actual labels, suggests that the overall accuracy of this model is better than the expected random chance classifier's accuracy.

Next, using the fit from the logistic regression model, the reserved test set is used to generate scores and calculate the Receiver Operating Characteristic curve.

```
set.seed(525)
fit = glm(m1.train.ra ~ .,
          data = m1.train.p[,-c(2)],
          family = binomial(link = "logit"))

link.scores = predict(fit, newdata = m1.test.p, type = "link")
response.scores = predict(fit, newdata = m1.test.p, type = "response")

score.df = data.frame(link = link.scores,
                      response = response.scores,
                      target = m1.test.ra,
                      stringsAsFactors = FALSE)
```

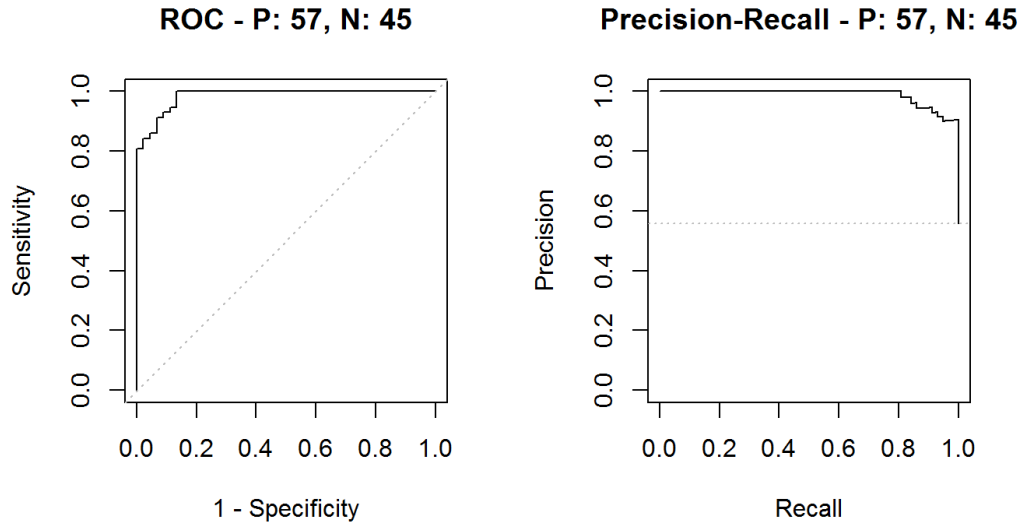
A plot of the link and response scores highlight that the classifications are the same. Also, it is apparent that there is some misclassification by the model in predicting the species using the test set.

```
ggplot(score.df, aes(x = link, y = response, col = target)) +
  scale_color_manual(values = c("black", "red")) +
  geom_point() + geom_rug() +
  ggtitle("Link & Response Scores")
```



Here the ROC curve for the response scores shows the broken line in the model as random choices (probability 50%) and the black solid line as the derived model. Already, the area under the curve (AUC) for the model is larger, highlighting that the accuracy is better than random choices. The AUC score is 0.984 which is quite good. Therefore, there is no need to further tune the current model to have a higher TPR.

```
precrec_obj = evalmod(scores = response.scores, labels = m1.test.ra)
plot(precrec_obj)
```



Coefficient Discussion

When interpreting the coefficient (Problem 1c), it is standard that a positive coefficient represents an increased probability that a penguin belongs to the `Adélie` species, since it is the reference factor. Whereas, a negative coefficient represents a decreased probability that a penguin belongs to the `Adélie` species.

```
varImp(model1_adelie)
```

```
## ROC curve variable importance
##
##                   Importance
## flipper_length_mm    100.000
## body_mass_g           80.728
## bill_depth_mm         77.993
## island                65.046
## year                   7.162
## sex                    0.000
```

From the intercept only, given no other information about the penguin measurements or where they are from, there is decreased chance that the penguin is `Adélie` . The predictor that statistically influence the classification is the flipper length. For reasons discussed behind the creation of the binary species level, the measurements based on island habitat is expected to influence the model. As a result, a penguin who resides on the island of Dream as opposed to Biscoe, the log odds of being Adelie species (versus non-Adelie species) increases by 2.56. For every one unit change in the flipper length of a penguin, the log odds of being Adelie species increases by 0.206. Adelie's species tend to have a smaller flipper length than compared to the Chinstrap and Gentoo, classified as Non-Adelie. Whereas, for a one-unit increase in bill depth, the log odds of being Adelie species decreases by 0.548. The Non-Adelie, particularly Gentoo, species have a larger bill depth than Adelie.

With a confidence level of 95%, the final binary logistic model for the probability that a penguin is Adelie or not is:

$$\hat{p}(X) = \frac{e^{-31.61+2.563\times Island_{Dream}-0.548\times BillDepth+0.206\times FlipperLength}}{1+e^{-31.61+2.563\times Island_{Dream}-0.548\times BillDepth+0.206\times FlipperLength}}$$

Model 2: Multinomial Logistic Regression

Multinomial logistic regression reports the odds of being in the different outcome categories about some base groups. In this assignment, a model is built to capture the odds of a penguin belonging to a specific species based on the independent variables (Problem 3a). Because there are three levels to `species` , the model will report two distinct sets of regression results corresponding to the following two models:

$$log(\frac{Pr(species = Chinstrap)}{Pr(species = Adélie)}) = \beta_0 + \beta_1(X_1) + \ldots + \beta_n(X_n) + \varepsilon$$

$$log(\frac{Pr(species = Gentoo)}{Pr(species = Adélie)}) = \beta_0 + \beta_1(X_1) + \ldots + \beta_n(X_n) + \varepsilon$$

In this case, the `Adélie` species is treated as the reference group of the three species. The multinomial log-linear model via neural networks is fitted. Moreover, to account for the complete separation, bias reduction is used to fit the multinomial regression model. There is an "only intercept" model to baseline the models as they are tested to find the best fit. To find the optimal fit and account for a parsimonious model, a variable selection will be implemented, and the optimal fit is found using both forward and backward stepwise-regression based on the Akaike information criterion. The stepwise AIC method will be used to select the best model from an information-criterion perspective, therefore cross-validation is not conducted, and this will further help to produce a parsimonious model (Problem 3b).

```
set.seed(525)
# Baseline model
base = multinom(m2.train.r ~ 1, data = m2.train.p[,-c(2)])
```

```
## # weights:  6 (2 variable)
## initial  value 265.864174
## final  value 253.978590
## converged
```

```
# Model 2: All variable first
model.full = multinom(m2.train.r ~ . , data = m2.train.p[,-c(2)])
```

```
## # weights:  30 (18 variable)
## initial  value 265.864174
## iter  10 value 67.802235
## iter  20 value 52.213692
## iter  30 value 51.997608
## iter  40 value 51.952516
## iter  50 value 51.938348
## iter  60 value 51.937445
## final  value 51.937442
## converged
```

```
# Variable selection
model2 = step(base, list(lower = formula(base),
                         upper = formula(model.full)),
              direction = "both", trace = 0)
```

```
## trying + island
## trying + bill_depth_mm
## trying + flipper_length_mm
## trying + body_mass_g
## trying + sex
## trying + year
## # weights:  9 (4 variable)
## initial  value 265.864174
## iter  10 value 107.857826
## iter  20 value 105.672285
## iter  30 value 105.242334
## iter  40 value 105.160945
## iter  50 value 105.151796
## final  value 105.150117
## converged
## trying - flipper_length_mm
## trying + island
## trying + bill_depth_mm
## trying + body_mass_g
## trying + sex
## trying + year
## # weights:  15 (8 variable)
## initial  value 265.864174
## iter  10 value 88.589062
## iter  20 value 58.864265
## iter  30 value 57.514511
## iter  40 value 56.915166
## final  value 56.914939
## converged
## trying - flipper_length_mm
## trying - island
## trying + bill_depth_mm
## trying + body_mass_g
## trying + sex
## trying + year
## # weights:  18 (10 variable)
## initial  value 265.864174
## iter  10 value 99.727968
## iter  20 value 55.215158
## iter  30 value 54.650178
## iter  40 value 54.204467
## iter  50 value 53.836797
## iter  60 value 53.834728
## final  value 53.834646
## converged
## trying - flipper_length_mm
## trying - island
## trying - sex
## trying + bill_depth_mm
## trying + body_mass_g
## trying + year
```

Multinomial Logistic Regression Output

	(Intercept)	flipper_length_mm	islandDream	islandTorgersen	sexmale
Chinstrap	-38.473	0.134	13.436	-3.459	-1.281
Gentoo	-272.000	1.357	-149.111	-37.087	-3.089

The summary results in the two models when `Adélie` is the reference point. In other words, the rows with `Chinstrap` are for the model comparing the probability of being a `Chinstrap` penguin versus an `Adélie` penguin. While the rows with `Gentoo` are for the model comparing the probability of being a `Gentoo` penguin versus an `Adélie` penguin.

Performance Criteria (Problem 4)

From the results below, a p-value calculation for the regression coefficients is used to determine whether coefficients are significant or not at $\alpha = 0.05$. It suggests that is the simplest model with great explanatory predictive power is one based on a penguin flipper length, island habitat, and gender. The model converged and the final negative log-likelihood is 53.83. The Akaike Information Criterion (AIC) is 127.6. More specifically, this model has the smallest AIC, suggesting that it is the best candidate among all other models in the step process.

```
# Model 2: Z-test
z.score = summary(model2)$coefficients/summary(model2)$standard.errors
p.values = (1 - pnorm(abs(z.score), 0, 1)) * 2
p.values
```

```
##           (Intercept) flipper_length_mm  islandDream islandTorgersen    sexmale
## Chinstrap           0      0.0006118135 0.0002869186               0 0.01717995
## Gentoo              0      0.0000000000 0.0000000000               0 0.00000000
```

The chi-square statistic measures the goodness of the fit between the observed values and the predicted values. The change result is significant, which means that the final model explains a significant amount of the original variability.

```
chisq.test(m2.train.r, predict(model2))
```

```
##
##  Pearson's Chi-squared test
##
## data:  m2.train.r and predict(model2)
## X-squared = 311.25, df = 4, p-value < 2.2e-16
```

The confusion matrix results suggest that 87.3% of the predicted results seems to be correctly classified. The precision for each type of species is high (Adelie = 82%, Chinstrap = 79%, and Gentoo = 97%), suggesting that the penguins belong to the actual species among all the penguins predicted to be that particular species. Moreover, the recall highlights that 91% of the Adelie species have been correctly classified accordingly, whereas 55% of the Chinstrap species have been correctly classified, and 100% of the Gentoo species have been correctly classified. In all, this model is capable of classifying penguins into one of the three species, particularly Adelie and Gentoo. And lastly, the Kappa statistic, which is a measure of agreement between the predictions and the actual labels, suggests that the overall accuracy of this model is better than the expected random chance classifier's accuracy.

```
# Model 2: Confusion Matrix
m2.pred.P = predict(model2, newdata = m2.test.p, type = "prob")
m2.pred.R = predict(model2, newdata = m2.test.p, type = "class")

m2.confusion = confusionMatrix(m2.pred.R, m2.test.r, mode = "everything")
m2.confusion
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  Adelie Chinstrap Gentoo
##   Adelie        41         9      0
##   Chinstrap      3        11      0
##   Gentoo         1         0     37
##
## Overall Statistics
##
##                Accuracy : 0.8725
##                  95% CI : (0.7919, 0.9304)
##     No Information Rate : 0.4412
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.795
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: Adelie Class: Chinstrap Class: Gentoo
## Sensitivity                 0.9111           0.5500        1.0000
## Specificity                 0.8421           0.9634        0.9846
## Pos Pred Value              0.8200           0.7857        0.9737
## Neg Pred Value              0.9231           0.8977        1.0000
## Precision                   0.8200           0.7857        0.9737
## Recall                      0.9111           0.5500        1.0000
## F1                          0.8632           0.6471        0.9867
## Prevalence                  0.4412           0.1961        0.3627
## Detection Rate              0.4020           0.1078        0.3627
## Detection Prevalence        0.4902           0.1373        0.3725
## Balanced Accuracy           0.8766           0.7567        0.9923
```

Lastly, below are some pseudo-R-squared statistics because logistic regression does not have an equivalent to the R^2 that is found in OLS regression. The goodness of fit of these pseudo R^2 statistics is mostly based on the deviance of the model. The Cox and Snell's R^2 replicates the R^2 based on 'likelihood', but its maximum can be less than 1.0, even for 'perfect' models. This makes it difficult to interpret. Below, the pseudo R^2 suggests that 80.97% of the variation in the dependent variable is explained by the model.

There is also the Nagelkerke modification to the Cox and Snell's R^2. It ranges from 0 to 1 and is considered a more reliable measure. In this case, it suggests that there is a relationship of 92.1% between the predictors and the prediction.

```
PseudoR2(model2, which = c("CoxSnell", "Nagelkerke", "AIC"))
```

```
##    CoxSnell  Nagelkerke         AIC
##   0.8087325   0.9217158 127.6692921
```

Coefficient Discussion

Again, when interpreting the coefficient (Problem 3c), a positive coefficient represents an increased probability, whereas, a negative coefficient represents a decreased probability that a penguin belongs to a specific species. The model indicates that the most important variable is the flipper length, followed by the island habitat, and gender.

```
varImp(model2)
```

```
##                     Overall
## flipper_length_mm   1.490262
## islandDream       162.546783
## islandTorgersen    40.545972
## sexmale             4.369114
```

Thus, the final model with the logit coefficients relative to the reference category, `Adelie`, becomes:

$$log(\frac{Pr(\hat{species} = Chinstrap)}{Pr(species = Adelie)}) = -38.47 + 0.13 \times FlipperLength + 13.43 \times Island_{Dream} - 3.45 \times Island_{Torgersen} - 1.28Sex_{mal}$$

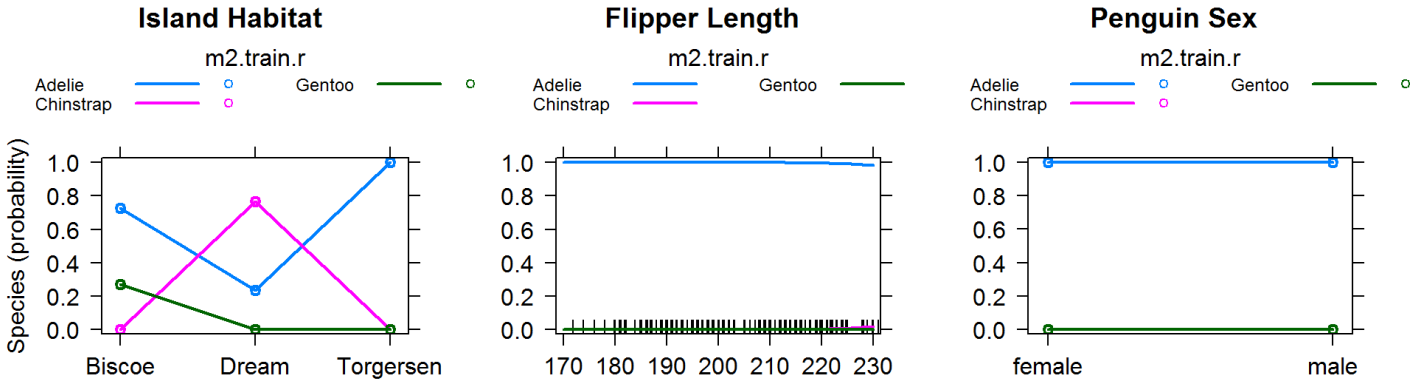
$$log(\frac{Pr(\hat{species} = Gentoo)}{Pr(species = Adelie)}) = -271.99 + 1.36 \times FlipperLength - 149.11 \times Island_{Dream} - 37.09 \times Island_{Torgersen} - 3.09Sex_{mal}$$

where

- The log odds for a penguin being a Chinstrap instead of an Adelie will have:
  - Flipper Length: increase by 0.13 in the length of the penguin flipper.
  - Island: increase by 13.43 if moving from “Biscoe” to “Dream”, and decrease by 3.45 if moving from “Biscoe” to “Torgersen” based on the island habitat in Antarctica.
  - Gender: decrease by 1.28 if moving from “female” to “male” as the penguin gender.
- The log odds for a penguin being a Gentoo instead of an Adelie will have:
  - Flipper Length: increase by 1.36 in the length of the penguin flipper.
  - Island: decrease by 149.11 if moving from “Biscoe” to “Dream”, and decrease by 37.09 if moving from “Biscoe” to “Torgersen” based on the island habitat in Antarctica.
  - Gender: decrease by 3.09 if moving from “female” to “male” as the penguin gender.

The plots below highlight the effect of each predictor according to their change in factors. For instance, sex and flipper length highlight large differences in the penguin species, whereas there is a noticeable difference in penguin species based on the island habitat. The probability of Chinstraps and Gentoos found on the island of Torgersen is 0.

```
p1 = plot(Effect("island", model2), multiline = TRUE,
          axes=list(x = list(island = list(lab = "")),
                    y = list(lab = "Species (probability)")),
          main = "Island Habitat")

p2 = plot(Effect("flipper_length_mm", model2), multiline = TRUE,
          axes=list(x = list(flipper_length_mm = list(lab = "")),
                    y = list(lab = "")),
          main = "Flipper Length")

p3 = plot(Effect("sex", model2), multiline = TRUE,
          axes=list(x = list(sex = list(lab = "")),
                    y = list(lab = "")),
          main = "Penguin Sex")

gridExtra::grid.arrange(p1,p2,p3,nrow=1, ncol=3)
```



Conclusion

Given the `palmerpenguins` dataset, which contains size measurements for three penguin species, a binary and a multinomial logistic regression was fitted. Upon exploratory analysis, some initial observations were be made about the dataset. These included working with missing data, bi- and tri-modal distributions, and definitive variable that produced complete separation. As a preparation, target variable was transformed into binary

outcomes based on island separations, data was processed for missing, and splited into train and test sets for model evaluations. As a result, the binary logistic regression model based on Bayesian analysis produced a model that is 91.2% accurate in correctly classifying penguins into `Adelie` and `NonAdelie`. The binary model is:

$$\hat{p}(X) = \frac{e^{-31.61+2.563\times Island_{Dream}-0.548\times BillDepth+0.206\times FlipperLength}}{1+e^{-31.61+2.563\times Island_{Dream}-0.548\times BillDepth+0.206\times FlipperLength}}$$

Moreover, a multinomial logistic regression was built to capture the odds of a penguin belonging to a specific species based on the predictors. For this model, bias reduction cause by complete separation and stepwise-regression based on the Akaike information criterion were implemented. The final multinomial models with 87.3% accuracy resulted in:

$$log(\frac{Pr(\hat{species} = Chinstrap)}{Pr(species = Adelie)}) = -38.47 + 0.13 \times FlipperLength + 13.43 \times Island_{Dream} - 3.45 \times Island_{Torgersen} - 1.28Sex_{mal}$$

$$log(\frac{Pr(\hat{species} = Gentoo)}{Pr(species = Adelie)}) = -271.99 + 1.36 \times FlipperLength - 149.11 \times Island_{Dream} - 37.09 \times Island_{Torgersen} - 3.09Sex_{mal}$$

In conclusion, both models perform well in classifying the penguins, even without the perfectly discriminating variable.

Works Cited

1. Horst AM, Hill AP, Gorman KB (2020). *palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0*. https://allisonhorst.github.io/palmerpenguins/ (https://allisonhorst.github.io/palmerpenguins/). doi:10.5281/zenodo.3960218 (doi:10.5281/zenodo.3960218).