Using logistic regression to predict species of penguin

C. Rosemond

February 19, 2021

1. Logistic regression with a binary outcome (40)

The goal of this analysis is to build a binary logistic regression model that predicts the species of penguin given some combination of seven initial features. The full data set contains 344 records, each one representing a penguin observed on three islands in the Palmer Archipelago, Antarctica. Each record has a response `species`, which consists of three different species of penguin. The seven possible model features describe various characteristics of the observed penguins. They range from island where observed (`island`) to length of bill in millimeters (`bill_length_mm`) to body mass in grams (`body_mass_g`). There is a `year` feature, but there is zero indication that the records are panel data describing the same penguins over time.

Table 1: Data summary

Name	penguins
Number of rows	344
Number of columns	8
Column type frequency:	
factor	3
numeric	5
Group variables	None

Variable type: factor

skim_variable	complete_rate	ordered	n_unique	top_counts
species	1.00	FALSE	3	Ade: 152, Gen: 124, Chi: 68
island	1.00	FALSE	3	Bis: 168, Dre: 124, Tor: 52
sex	0.97	FALSE	2	mal: 168, fem: 165

Variable type: numeric

skim_variable	complete_rate	mean	sd	p0	p50	p100
bill_length_mm	0.99	43.92	5.46	32.1	44.45	59.6
bill_depth_mm	0.99	17.15	1.97	13.1	17.30	21.5
flipper_length_mm	0.99	200.92	14.06	172.0	197.00	231.0
body_mass_g	0.99	4201.75	801.95	2700.0	4050.00	6300.0
year	1.00	2008.03	0.82	2007.0	2008.00	2009.0

The `species` response consists of three categories: Adelie, with 152 observations; Gentoo, with 124 observations; and Chinstrap, with 68 observations. Zero observations have a missing response.

Personal domain expertise with other wildlife suggests that `island`, or geographic location, could share a close relationship with `species`. A cross-tabulation reveals that Chinstrap and Gentoo penguins are solely found on Dream and Biscoe islands, respectively, while solely Adelie penguins are found on Torgersen island. This pattern is identified as statistically significant ($\alpha = 0.05$) by a Pearson's Chi-squared test ($\chi^2 = 299.55$ on 4 degrees of freedom, p-value ~ 0). The same test applied to combinations of `species` and each of `sex` and `year` returns results that are not statistically significant ($\alpha = 0.05$).

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

```
    Pearson's Chi-squared test

data:  table(penguins$species, penguins$island)
X-squared = 299.55, df = 4, p-value < 2.2e-16
```

```
    Pearson's Chi-squared test

data:  table(penguins$species, penguins$sex)
X-squared = 0.048607, df = 2, p-value = 0.976
```

```
    Pearson's Chi-squared test

data:  table(penguins$species, penguins$year)
X-squared = 3.2156, df = 4, p-value = 0.5224
```

This perfect or near-perfect separation between `species` and `island` (or possibly other features) is not a bad thing, practically speaking, but it could pose problems for logistic regression. Regression typically uses maximum-likelihood estimation to estimate model parameters. Considering [near-]separation between values of the response and a feature, the resulting [near-]certainty around the data could prevent convergence to a maximum-likelihood estimate and thus result in unusually high parameter estimates with even higher standard errors. The feature `island` or its dummies may need to be removed from modeling given a response derived from `species`.

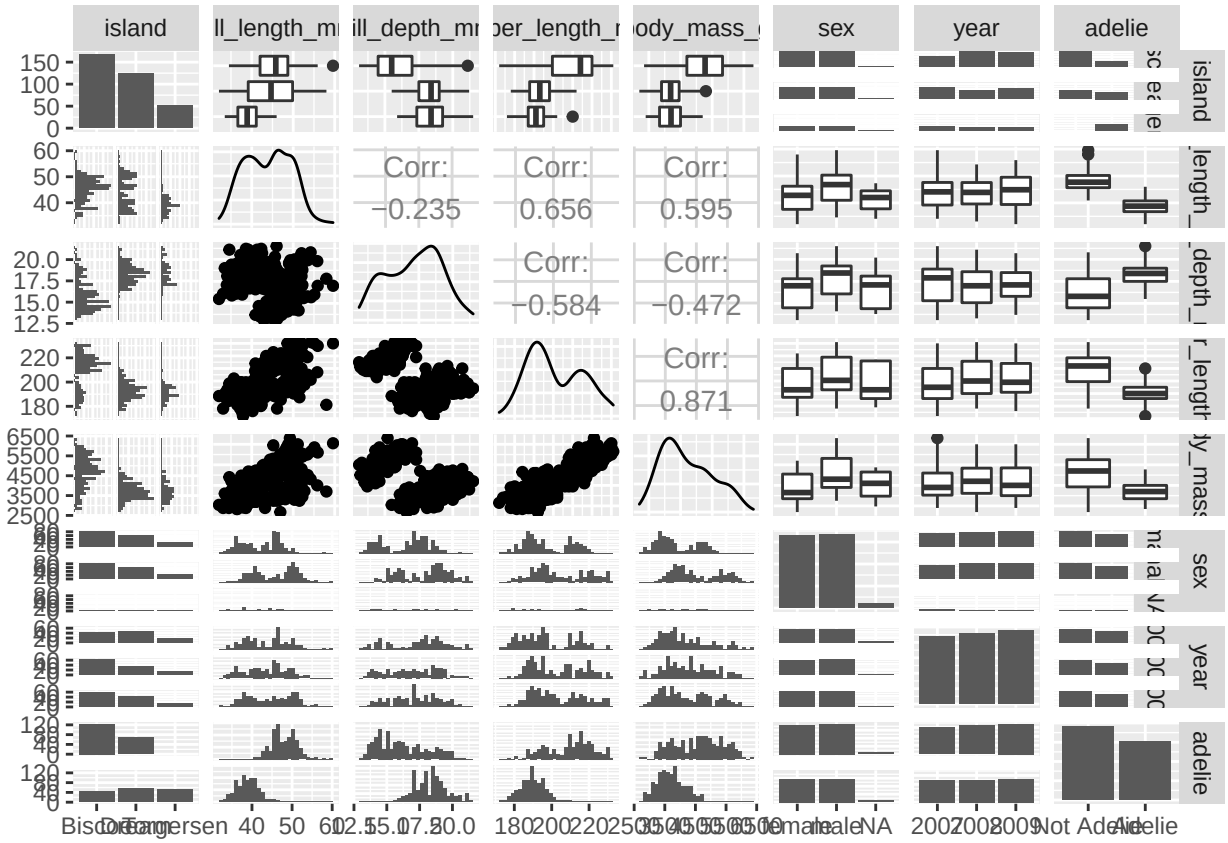
A new binary response `adelie` is created from `species`. The ultimate model will predict whether a penguin is an Adelie penguin (`adelie` == "Adelie") or is not an Adelie penguin (`adelie` == "Not Adelie"). One-hundred-fifty-two, or approximately 44.2 percent of observations have a value of "1" for `adelie`. Choosing `adelie`, and collapsing "Gentoo" and "Chinstrap" of `species`, creates relative balance in the binary response, which will facilitate modeling.

```
[1] 152
```

Another round of chi-squared tests, this time with `adelie`, is similar in results to the first. The relationship between `adelie` and `island` is statistically significant ($\alpha = 0.05$) with a p-value of approximately zero.

```
             Biscoe Dream Torgersen
  Not Adelie    124    68         0
  Adelie         44    56        52


    Pearson's Chi-squared test

data:  table(df$adelie, df$island)
X-squared = 87.792, df = 2, p-value < 2.2e-16


    Pearson's Chi-squared test with Yates' continuity correction

data:  table(df$adelie, df$sex)
X-squared = 0.0012128, df = 1, p-value = 0.9722


    Pearson's Chi-squared test

data:  table(df$adelie, df$year)
X-squared = 0.11208, df = 2, p-value = 0.9455
```

Beyond the response, the initial exploratory data analysis reveals several characteristics about the features. First, the data set consists generally of complete observations, and all features are complete at rates greater than approximately 0.97. Second, the numeric body characteristic features do not appear to show significant skewness, though they may yet benefit from transformation prior to modeling. And third, `year` is currently numeric but should arguably be categorical in this context; it is converted to a factor. All of the factors will be converted to dummy features, minus one level dummy for each original factor, prior to modeling.



Visualizing the distributions of the response and features along with the relationships between them reveals additional information to inform modeling. First, there are clear differences in body characteristics between "Adelie" and "Not Adelie" values of response `adelie`. These features could prove helpful for modeling. Second, considering the factors and building upon the prior chi-square tests, `sex` and `year` appear relatively balanced across levels and values of `adelie`; `island`, again, does not. And third, `flipper_length_mm` and `body_mass_g` share a strong positive relationship that may need to be accounted for prior to modeling. There also appear to be possible gender differences in the body characteristics.

Data pre-processing starts with addressing missingness. Imputing meaning for missing values–meaning where it may not exist–can be problematic, particularly with limited domain expertise. For ease of analysis and in response to the relatively few missing values as well as minimal information available to assess the data set's ignorability, the eleven observations with missing values are dropped, leaving 333 observations in the data set.

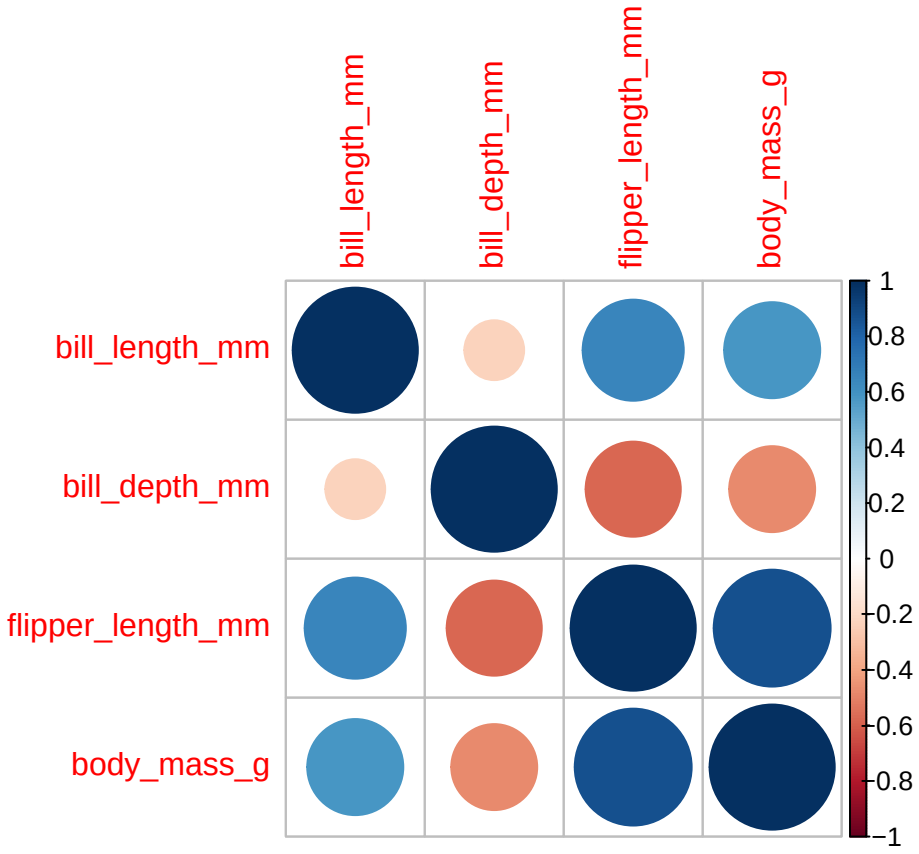
```
[1] 333
```

There may be a need to transform existing features or create new ones in preparation for modeling. This process begins by assessing the numeric body characteristic features for possible power transformation. Below are skewness statistics for each feature, with negative values reflecting left skewness and positive values reflecting right skewness. Larger values are associated with greater levels of skewness. None of the predictors show large skew. For ease of interpretability of model coefficients, they are left not transformed.

	Skewness
bill_length_mm	0.0449328
bill_depth_mm	-0.1483741
flipper_length_mm	0.3569099

	Skewness
body_mass_g	0.4680001

Regarding multicollinearity, Pearson's correlation coefficient is used to check correlations between the body characteristic features. A plot of the correlations confirms the relative relationships identified earlier, though none of them feature correlations that exceed 0.90–a default used by the `caret` package. At approximately 0.87, the pair of `flipper_length_mm` and `body_mass_g` are close, but they won't be removed from consideration.




```
[1] 0
```

Lastly, the dataset is split 80/20 into a training set (n = 267) and a test set (n = 66). The latter will be held out for validation of the final model.

Model 1

The first model is simple and regresses response `adelie` on all features in the training set. Factors `island`, `sex`, and `year` are converted to sets of dummies for each level of the original factor. Then, the first level dummy of each set is left out to avoid perfect collinearity between levels of the set.

```
Warning: glm.fit: algorithm did not converge
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred


Call:
glm(formula = adelie ~ ., family = binomial, data = train)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-9.737e-05  -2.100e-08  -2.100e-08   2.100e-08   8.505e-05

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        4.011e+02  3.369e+05   0.001    0.999
islandDream       -4.308e+01  5.982e+04  -0.001    0.999
islandBiscoe      -3.097e+01  1.087e+05   0.000    1.000
bill_length_mm    -2.052e+01  6.822e+03  -0.003    0.998
bill_depth_mm      2.893e+01  1.019e+04   0.003    0.998
flipper_length_mm -2.101e-01  1.841e+03   0.000    1.000
body_mass_g        8.421e-03  7.533e+01   0.000    1.000
sexmale            3.669e+01  8.331e+04   0.000    1.000
year2008          -1.586e+01  9.113e+04   0.000    1.000
year2009           3.999e+00  3.066e+04   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.6605e+02  on 266  degrees of freedom
Residual deviance: 2.9749e-08  on 257  degrees of freedom
AIC: 20

Number of Fisher Scoring iterations: 25
```

The estimated model coefficients are all relatively high, and their respective estimated standard errors are even higher. The associated z-scores are all approximately zero, leading to very high p-values. The median residual error is also approximately zero.

It is immediately clear that there is an issue with this model. The warning notes that the underlying model algorithm failed to converge on parameter estimates and that at least some of the fitted probabilities were zero or one. The former suggests an instance of the separation and maximum-likelihood estimation issue arising earlier with `adelie` (and `species`) and `islands`. Holding out the `islandTorgersen` dummy, which perfectly predicts `adelie`, made no difference, however.

Model 2

The process for the second model starts over, focusing on features based upon the EDA. Considering the feature boxplots for levels of `adelie`, body characteristics `bill_length_mm` and `flipper_length_mm` appear to show clear differences between "Adelie" and "Not Adelie". Notably, for both features, their distributions across the levels of `adelie` overlap minimally, which suggests they could be strong predictors. By comparison, the distributions for the other two body characteristic features show greater overlap. Additionally, excluding `island`, `sex` and `year` are both balanced across "Adelie" and "Not Adelie", so they may not be strong predictors.

The second model regresses `adelie` on `bill_length_mm` and `flipper_length_mm`.

```
Call:
glm(formula = adelie ~ bill_length_mm + flipper_length_mm, family = "binomial",
    data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.50847  -0.08729  -0.00840   0.09018   2.38805

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       62.50863    9.89586   6.317 2.67e-10 ***
bill_length_mm    -1.00166    0.17221  -5.816 6.01e-09 ***
flipper_length_mm -0.09860    0.03289  -2.998  0.00272 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 366.05  on 266  degrees of freedom
Residual deviance:  67.34  on 264  degrees of freedom
AIC: 73.34

Number of Fisher Scoring iterations: 8
```

Unlike the first model, the second one converges upon a maximum-likelihood estimate. Its AIC of approximately 73.34 seems okay. All of the model coefficients are statistically significant ($\alpha = 0.05$) with p-values of approximately zero, and the negative coefficients on `bill_length_mm` and `flipper_length_mm` suggest that those features share negative relationships with `adelie`. The intercept's coefficient and standard error are quite high, but they are not a concern given that encountering a penguin whose `bill_length_mm` and `flipper_length_mm` are zero is impossible.

```
                                    2.5 %       97.5 %
(Intercept)       1.403308e+27 7.697726e+19 9.967371e+36
bill_length_mm    3.672690e-01 2.474669e-01 4.917264e-01
flipper_length_mm 9.061038e-01 8.452848e-01 9.635227e-01
```

Looking at the coefficients as odds-ratios eases their interpretation. Per this model, an increase of one mm in `bill_length_mm` is associated with a decrease of roughly 63 percent in the likelihood of being an "Adelie" penguin (`adelie` == "Adelie"). The 95% confidence interval, based on log likelihood, about that odds-ratio estimate ranges from a rough 75 percent decrease in likelihood to a rough 51 percent decrease in likelihood. Similarly, an increase of one mm in `flipper_length_mm` is associated with a decrease of roughly nine percent in the likelihood of being an "Adelie" penguin; its 95% confidence interval ranges from a rough fifteen percent decrease to a rough four percent decrease.

```
[1] 298.711
```

```
[1] 2
```

```
[1] 1.366874e-65
```

The chi-square test statistic of approximately 298.71 with 2 degrees of freedom–the number of model predictors–has a p-value of approximately 0, which is statistically significant at $\alpha = 0.05$. So, the model fits the data better than a null model consisting solely of an intercept.

```
# A tibble: 4 x 6
  .metric  .estimator  mean     n std_err .config
  <chr>    <chr>      <dbl> <int>   <dbl> <fct>
1 accuracy binary     0.951     5 0.0152  Preprocessor1_Model1
2 pr_auc   binary     0.992     5 0.00432 Preprocessor1_Model1
3 sens     binary     0.953     5 0.0121  Preprocessor1_Model1
4 spec     binary     0.956     5 0.0236  Preprocessor1_Model1
```

Incorporating five-fold cross-validation of the model using the training set provides additional insights. The model performs well across metrics, with mean accuracy across the five folds of approximately 0.95, mean sensitivity of approximately 0.95, mean specificity of approximately 0.96, and mean area under the ROC curve of approximately 0.99. However, better measures of its actual performance come when predicting on the test set.

```
         Actual
Predicted Not Adelie Adelie
        0         37      1
        1          0     28
```

A confusion matrix of the model's predictions on the test set suggests good performance. There are 28 true positives (TP), 37 true negatives (TN), and 1 false negative (FN). There are zero false positives (FP).

2. For your model from #1, please provide: AUC, Accuracy, TPR, FPR, TNR, FNR (20)

Accuracy

Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

```
[1] 0.9848485
```

The model's accuracy on the test set is approximately 0.98.

Sensitivity (True Positive Rate)

Sensitivity = $\frac{TP}{TP+FN}$

```
[1] 0.9655172
```

The model's sensitivity on the test set is approximately 0.97.

Specificity (True Negative Rate)

Specificity = $\frac{TN}{TN+FP}$

```
[1] 1
```

The model's specificity on the test set is approximately one.

False Positive Rate

False Positive Rate = $\frac{FP}{TN+FP}$

```
[1] 0
```

The model's false positive rate–the complement of its specificity–is approximately zero.

False Negative Rate

False Negative Rate = $\frac{FN}{TP+FN}$

```
[1] 0.03448276
```

The model's false negative rate–the complement of its sensitivity–is approximately 0.03.

Area under curve (AUC)

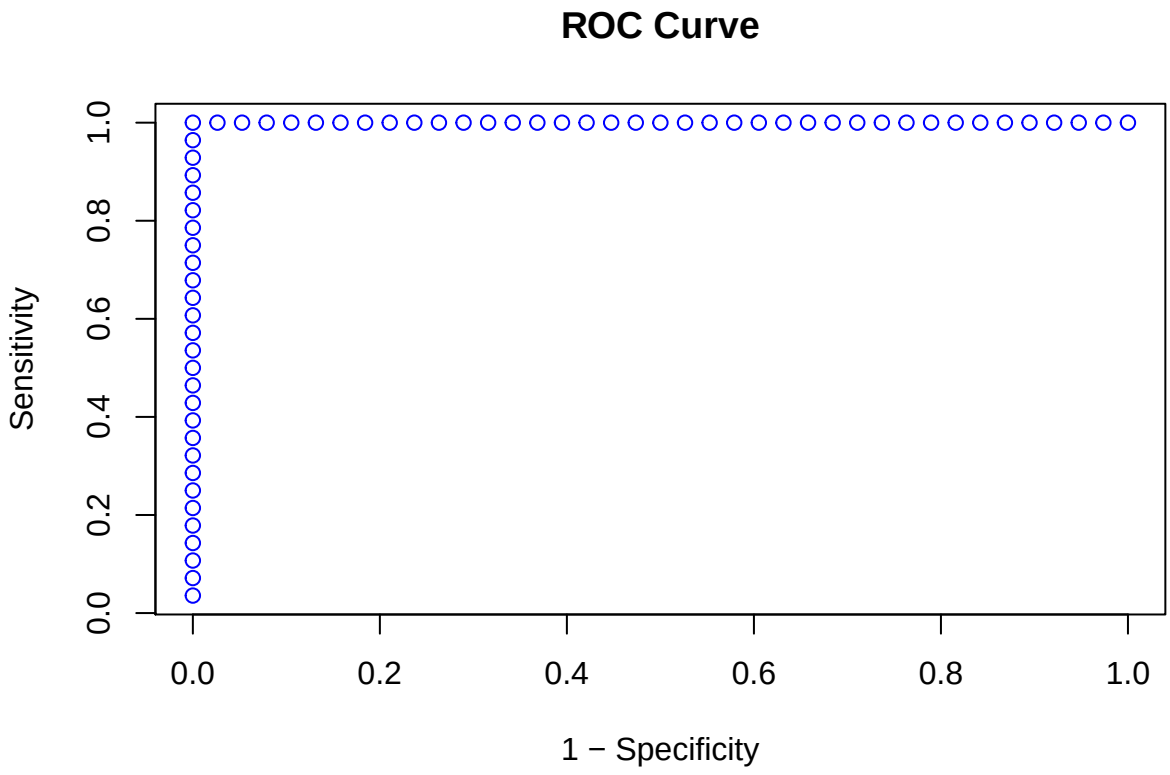


Figure 1: Model ROC curve

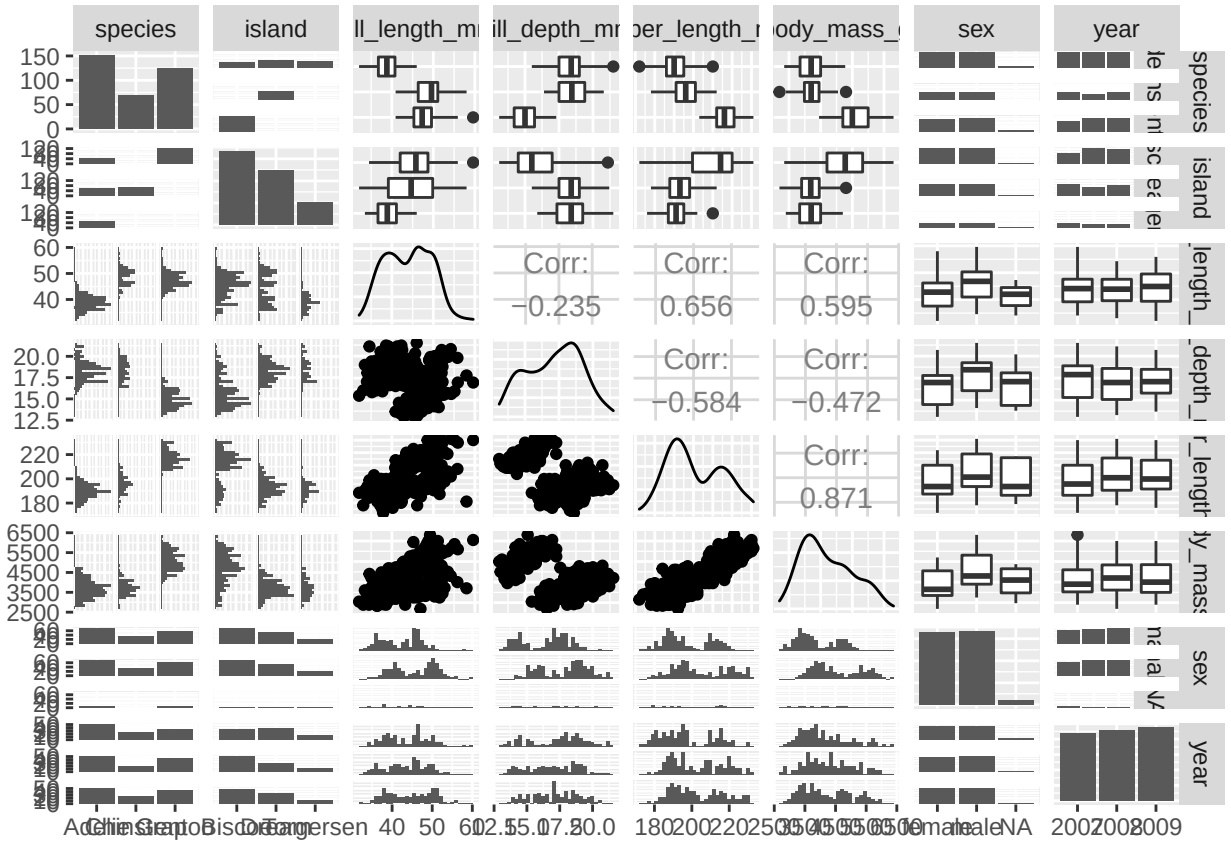
```
[1] 1
```

The AUC for the model's ROC curve is approximately one.

In sum, the model predicts the test set near perfectly. This performance seems more than sufficient for a model including only two predictors: features `bill_length_mm` and `flipper_length_mm`. Presumably, incorporating additional features could result in perfect predictive performance, but doing so would come at the cost of some simplicity.

3. Multinomial Logistic Regression (40)

Construction of a multinomial logistic regression model to predict `species` mirrors the EDA and data preprocessing performed for the binomial logistic regression model.



Here again, visualizing the distributions of the response and possible features reveals information to inform modeling. First, there are clear differences in body characteristics between values of `species`. Specifically, relative to the other two species, "Gentoo" penguins tend to have lower values for `bill_depth_mm` and higher values for `flipper_length_mm` and `body_mass_g`. Second, "Gentoo" penguins are found solely on "Biscoe" island, while "Chinstrap" penguins are found solely on "Dream" island. The feature `island` could be *too* predictive of `species` to ensure a model's maximum-likelihood estimate will converge. And third, again, each of the species appears well-balanced across the levels of factors `sex` and `year`.

```
[1] 333
```

After dropping the eleven incomplete observations, and considering the limited skewness and multicollinearity identified earlier, 333 observations are left in the data set. This full set is split 80/20 into training (n = 268) and test sets (n = 65); the latter will be held out.

Model 1

The first multinomial model regresses response `species` on all available features. Factors `island`, `sex`, and `year` are converted to sets of dummies for each level of the original factor, then the first level dummy of each set is left out.

```
# weights:  33 (20 variable)
initial  value 294.428093
iter  10 value 11.935995
iter  20 value 1.898077
iter  30 value 0.438159
iter  40 value 0.002732
iter  50 value 0.000495
iter  60 value 0.000263
iter  70 value 0.000252
final  value 0.000083
converged

Warning in sqrt(diag(vc)): NaNs produced

Call:
multinom(formula = species ~ ., data = multi_train)

Coefficients:
          (Intercept) islandDream islandTorgersen bill_length_mm
Chinstrap  -167.05174    61.41441       0.9096979      18.538956
Gentoo      -16.94195   -35.42267    -107.1561319       9.200488
          bill_depth_mm flipper_length_mm body_mass_g   sexmale year2008
Chinstrap     -25.18528        -1.4763245  0.01261848 -60.91981 25.70718
Gentoo        -25.03192        -0.7964594  0.05165874 -29.11618 18.10889
          year2009
Chinstrap 21.59569
Gentoo    21.60720

Std. Errors:
          (Intercept)  islandDream islandTorgersen bill_length_mm
Chinstrap  10.4761448 1.027534e+01             NaN       83.17652
Gentoo      0.1427074 1.375313e-18    9.727807e-36       13.05603
          bill_depth_mm flipper_length_mm body_mass_g   sexmale
Chinstrap    100.785671          11.89442   0.5663817 66.716225
Gentoo         5.250032          46.08011   2.0516490  1.504335
              year2008  year2009
Chinstrap 0.3103352780 47.010136
Gentoo    0.0001794778  1.367312

Residual Deviance: 0.0001662986
AIC: 40.00017
```

Unlike its binomial regression model that included all available features, this multinomial model converged on a minimal negative-log likelihood estimate. Its coefficients, representing "Chinstrap" and "Gentoo", reference "Adelie". In effect, it combines two binomial comparisons: "Chinstrap" and "Adelie", and "Gentoo" and "Adelie".

However, the coefficients are very high and have very high standard errors, which prompts skepticism. Further, the residual deviance is very low, perhaps too low. The model's AIC is approximately forty.

```
           (Intercept)  islandDream islandTorgersen bill_length_mm
Chinstrap 2.820666e-73 4.698299e+26    2.483572e+00   1.125554e+08
Gentoo    4.387361e-08 4.131713e-16    2.901905e-47   9.901956e+03
          bill_depth_mm flipper_length_mm body_mass_g      sexmale
Chinstrap  1.153906e-11         0.2284759    1.012698 3.490296e-27
Gentoo     1.345163e-11         0.4509227    1.053016 2.264669e-13
             year2008   year2009
Chinstrap 146044649620 2392698356
Gentoo        73213631 2420412658
```

Exponentiating the logit coefficients reveals a set of unrealistic odds relative to the base level "Adelie". For example, a one mm increase in `bill_length_mm` is associated with a likelihood for "Chinstrap" that is hundreds of millions of times higher than that of "Adelie", while a one mm increase in `bill_depth_mm` raises the likelihood of "Chinstrap" by essentially zero. Unfortunately, this model is scrapped.

Model 2

The second multinomial model echoes its binomial counterpart, regressing `species` on `bill_length_mm` and `flipper_length_mm`. These features seems like a reasonable starting point given their previous predictive performance.

```
# weights:  12 (6 variable)
initial  value 294.428093
iter  10 value 43.574763
iter  20 value 36.706114
iter  30 value 32.748846
iter  40 value 31.964391
iter  50 value 31.787352
iter  60 value 31.783526
final  value 31.757087
converged

Call:
multinom(formula = species ~ bill_length_mm + flipper_length_mm,
    data = multi_train)

Coefficients:
          (Intercept) bill_length_mm flipper_length_mm
Chinstrap   -28.30133      1.2104751        -0.1299593
Gentoo     -125.83465      0.3516171         0.5408406

Std. Errors:
          (Intercept) bill_length_mm flipper_length_mm
Chinstrap   11.128707      0.2293308        0.06762680
Gentoo       1.544672      0.2496595        0.05797607

Residual Deviance: 63.51417
AIC: 75.51417
```

The model converges to an optimal log-likelihood of approximately 31.76. Its AIC is approximately 75.51, and its residual deviance, across all observations, is approximately 63.51. Unlike the initial multinomial model, this one has coefficient estimates that, at first glance, appear reasonable.

```
           (Intercept) bill_length_mm flipper_length_mm
Chinstrap 5.115526e-13       3.355078         0.8781312
Gentoo    2.242352e-55       1.421364         1.7174499
```

The exponentiated coefficients reveal information about the relationships between species based upon `bill_length_mm` and `flipper_length_mm`. Considering `bill_length_mm`, an increase in bill length of one mm is associated with an approximate 235 percent increase in the likelihood of predicting "Chinstrap" relative to "Adelie", and is associated with an approximate 42 percent increase in the likelihood of predicting "Gentoo" relative to "Adelie". Considering `flipper_length_mm`, an increase in flipper length of one mm is associated with an approximate 12 percent decrease in the likelihood of predicting "Chinstrap" relative to "Adelie", and is associated with an approximate 72 percent increase in the likelihood of predicting "Gentoo" relative to "Adelie".

Model 3

A third model improves slightly upon the second model. It regresses `species` on `bill_length_mm`, `flipper_length_mm`, and `year`. Per EDA, the last feature appears to show slight differences between `species` levels regarding the balance of observations across years.

```
# weights:  18 (10 variable)
initial  value 294.428093
iter  10 value 47.627902
iter  20 value 35.177922
iter  30 value 32.071987
iter  40 value 31.220926
final  value 31.218813
converged

Call:
multinom(formula = species ~ bill_length_mm + flipper_length_mm +
    year, data = multi_train)

Coefficients:
          (Intercept) bill_length_mm flipper_length_mm   year2008
Chinstrap   -30.01243      1.2316303        -0.1263817  0.4538026
Gentoo     -135.29980      0.3746701         0.5886083 -1.4149458
             year2009
Chinstrap -0.09026852
Gentoo    -1.79977236

Std. Errors:
          (Intercept) bill_length_mm flipper_length_mm year2008 year2009
Chinstrap   12.462517      0.2439740        0.07746277 1.116785 1.163715
Gentoo       1.660759      0.2764479        0.06513135 2.136743 2.172802

Residual Deviance: 62.43763
AIC: 82.43763
```

The model converges to an optimal log-likelihood of approximately 31.22–slightly lower than the second model's–and its AIC is approximately 82.44–slightly higher than the second's.

```
           (Intercept) bill_length_mm flipper_length_mm  year2008
Chinstrap 9.242011e-14       3.426812         0.8812784 1.5742873
```

```
Gentoo    1.737967e-59       1.454512         1.8014796 0.2429388
           year2009
Chinstrap 0.9136858
Gentoo    0.1653365
```

The exponentiated coefficients for `bill_length_mm`, `flipper_length_mm`, and `year` Considering `bill_length_mm`, an increase in bill length of one mm is associated with an approximate 243 percent increase in the likelihood of predicting "Chinstrap" relative to "Adelie". This estimate, after accounting for `year`, represents an increase in absolute value over its second model counterpart. The coefficient on "Gentoo", which indicates that a one mm increase in bill length is associated with an approximate 45 percent increase in likelihood of prediction, exhibits similar behavior compared to the same coefficient in the second model.

Adding `year` also amplifies the coefficients on `flipper_length_mm`. Here, a one mm increase in flipper length is associated with an approximate 11 percent decrease in the likelihood of predicting "Chinstrap" relative to "Adelie", and is associated with an approximate 80 percent increase in the likelihood of predicting "Gentoo" relative to "Adelie".

Regarding `year` itself, moving from "2007" to "2008" is associated with an approximate 57 percent increase in the likelihood of predicting "Chinstrap" relative to "Adelie" and an approximate 76 percent decrease in the likelihood of predicting "Gentoo" relative to "Adelie". By comparison, moving from "2007" to "2009" is associated with an approximate 9 percent decrease in the likelihood of predicting "Chinstrap" and an approximate 84 percent decrease in the likelihood of predicting "Gentoo".

```
          (Intercept) bill_length_mm flipper_length_mm   year2008
Chinstrap   -2.408216       5.048204         -1.631515  0.4063473
Gentoo     -81.468648       1.355301          9.037251 -0.6621974
             year2009
Chinstrap -0.07756924
Gentoo    -0.82831847
```

```
          (Intercept) bill_length_mm flipper_length_mm  year2008  year2009
Chinstrap   0.0160307   4.459837e-07         0.1027817 0.6844874 0.9381707
Gentoo      0.0000000   1.753216e-01         0.0000000 0.5078447 0.4074902
```

However, calculating z-score test statistics reveals that the coefficients on the `year` dummies are not statistically significant ($\alpha = 0.05$). Nor are all of the z-scores on `bill_length_mm` and `flipper_length_mm` significant at the same α. Statistical significance is not necessarily reliable for decision making, but here it suggests that `year` adds little to the model.

In response, the second model, which performs similarly with one less model parameter, is used for evaluation on the test set.

Test Set Prediction

```
[1] 0.9538462
```

There are limited options for assessing the multinomial model's predictive performance on the test set. One basic way is to calculate the proportion of observation predictions that match that observation's `species` value. Using the second multinomial model that regresses `species` on `bill_length_mm` and `flipper_length_mm`, approximately 95 percent of predictions matched the associated `species` values. This performance seems okay for the purposes of this exercise.

Sources

Cross Validated (2020).*Help testing the predictive quality of a binomial GLM (currently attempting using the "caret" package)*. Stack Exchange. Accessed February 15, 2021 from https://stats.stackexchange.com/questions/459724/help-testing-the-predictive-quality-of-a-binomial-glm-currently-attempting-usin

Institute for Digital Research & Education Statistical Consulting (2021). *Logit regression | R data analysis examples*. University of California, Los Angeles. Accessed February 14, 2021 from https://stats.idre.ucla.edu/r/dae/logit-regression/

Institute for Digital Research & Education Statistical Consulting (2021). *Multinomial logistic regression | R data analysis examples*. University of California, Los Angeles. Accessed February 14, 2021 from https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/