Exploration of Generalized Linear Models

Luke Andrade, Soana Ballolli, Dana Gestosani, Himani Patel
Submitted to: Jack Mardekian, PhD
Department of Statistics, Rutgers University
December 6, 2022

Abstract

Our project investigates three different generalized linear models by applying them to the Palmer penguin dataset. This set is a collection of data about 344 observations studying 8 variables: species, island, bill length, bill depth, flipper length, body mass, sex, and year. Our report applies this data to the Gaussian, multinomial, and ANCOVA models. Through our study, we found ... (include results here)

Introduction

Our report is an investigation into generalized linear models (GLMs). These models unify linear and nonlinear regression models to develop models for response variables whose distributions are nonnormal and are part of the exponential family (ex: normal, Poisson, binomial, exponential, and gamma distributions). The fundamental idea of GLMs is its two components: the response distribution and the link function. A link function is a function that relates the mean of the response distribution to a linear predictor. This function allows statisticians to map a non-linear relationship to a linear one.

There are multiple advantages to using GLMs over simple linear regression. The main benefit is that the response variable can have any form of the exponential distribution; it does not need to be transformed to the normal distribution. GLMs are also more flexible and less susceptible to overfitting. There are many different types of GLMs. In our project, we specifically focus on the Gaussian, multinomial, and ANCOVA models.

To apply these models, we will utilize the Palmer penguin dataset. This set was collected from the Palmer Archipelago in Antarctica. It consists of data from 344 penguins across three species (chinstrap, gentoo, adelie) collected from three different islands in the archipelago. Its variables include species, island, bill length (mm), bill depth (mm), flipper length (mm), body mass (g), sex, and year. The application of this data will allow us to learn more about these generalized linear models and notice the differences between the models we have chosen to study.

Programs Used and Packages Required

For this project, the use of R and RStudio were utilized in order to fit the data across all three GLMs. Certain packages were required in order to complete the functions for each model. These included the packages: palmerpenguins, tidyverse, caret, VGAM, nnet, rstatix, car, and multcomp.

The Data

```
head(palmerpenguins::penguins)
```

```
## # A tibble: 6 x 8
##   species island    bill_length_mm bill_depth_mm flipper_l~1 body_~2 sex    year
##   <fct>   <fct>              <dbl>         <dbl>       <int>   <int> <fct> <int>
## 1 Adelie  Torgersen           39.1          18.7         181    3750 male   2007
## 2 Adelie  Torgersen           39.5          17.4         186    3800 fema~  2007
## 3 Adelie  Torgersen           40.3          18           195    3250 fema~  2007
## 4 Adelie  Torgersen           NA            NA            NA      NA <NA>   2007
## 5 Adelie  Torgersen           36.7          19.3         193    3450 fema~  2007
## 6 Adelie  Torgersen           39.3          20.6         190    3650 male   2007
## # ... with abbreviated variable names 1: flipper_length_mm, 2: body_mass_g
```

```
summary(palmerpenguins::penguins)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

Gaussian Model

What is a Gaussian Model

A Gaussian or normal distribution model is used to model functions with a finite number of points. This model is part of the exponential family of distributions. When performing a linear regression using a Gaussian model, the distribution of y given x is a Gaussian distribution with some mean mu and variance $sigma^2$. A linear relationship between the data and the parameters of the distribution is expected. The link function for the gaussian model is the identity function.

Assumption for a Gaussian Model

1. Cases are independent
2. The response fits a distribution in the exponential family
3. Linearity between the transformed expected response in terms of the link function and the explanatory variables

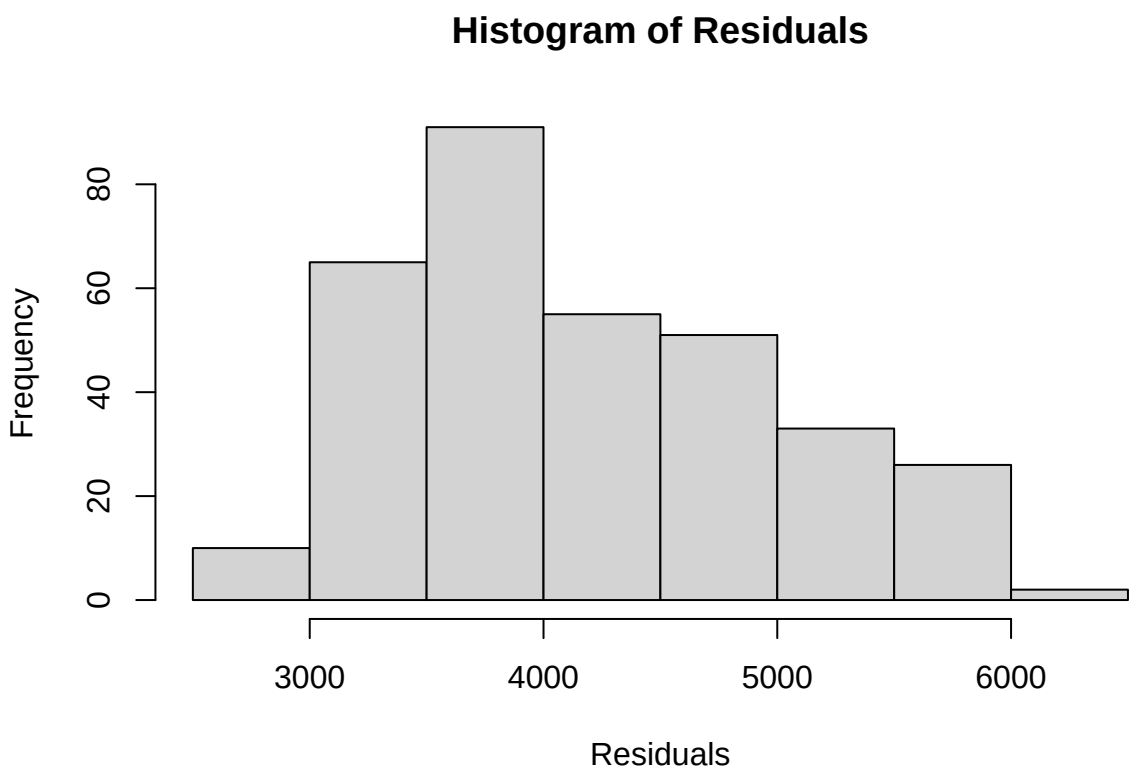
Setup

```
library(palmerpenguins)
library(tidyverse)
library(caret)
penguins_df <- na.omit(penguins)
continuous <- select_if(penguins_df, is.numeric)
```

Checking Assumption 1 and 2

Each case from the data set is independent.

```
hist(penguins_df$body_mass_g, xlab = 'Residuals', main = 'Histogram of Residuals')
```



This doesn't exactly appear to be normally distributed however we will continue with the modeling.

Splitting the Data into Test and Training Sets

```
trainindex <- createDataPartition(penguins_df$species, p = 0.85, list = FALSE)
training_set <- penguins_df[trainindex,]
testing_set <- penguins_df[-trainindex,]
```

```
model <- glm(body_mass_g ~ bill_length_mm + bill_depth_mm + flipper_length_mm, data = training_set)
summary(model)
```

```
##
## Call:
## glm(formula = body_mass_g ~ bill_length_mm + bill_depth_mm +
##     flipper_length_mm, data = training_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -874.22  -299.96   -27.24   234.08  1273.27
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6156.095    608.447 -10.118   <2e-16 ***
## bill_length_mm        6.577      5.704   1.153    0.250
## bill_depth_mm        11.783     14.670   0.803    0.423
## flipper_length_mm    49.163      2.692  18.263   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 152386.7)
##
##     Null deviance: 177681965  on 284  degrees of freedom
## Residual deviance:  42820675  on 281  degrees of freedom
## AIC: 4216
##
## Number of Fisher Scoring iterations: 2
```

As we can see the only regression coefficient that has a low enough p value to reject the null hypothesis is for the flipper length variable so we will remake the model to only include that variable.

```
model <- glm(body_mass_g ~ flipper_length_mm, data = training_set)
summary(model)
```

```
##
## Call:
## glm(formula = body_mass_g ~ flipper_length_mm, data = training_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -866.88  -267.75   -18.99   235.36  1280.88
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       -5807.981    338.382  -17.16   <2e-16 ***
## flipper_length_mm    49.876      1.681   29.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 152691.3)
##
##     Null deviance: 177681965  on 284  degrees of freedom
## Residual deviance:  43211627  on 283  degrees of freedom
## AIC: 4214.6
##
## Number of Fisher Scoring iterations: 2
```

Making Predictions with the Model

```
predictions <- predict.glm(model, testing_set, type = "link")
head(predictions)
```

```
##        1        2        3        4        5        6
## 3917.748 3718.246 3917.748 4017.499 3518.743 3319.241
```

Calculating RMSE of the Model

We will calculate the root mean squared error to evaluate the error of the model

```
RMSE <- sqrt(sum((predictions - testing_set$body_mass_g)^2) / length(predictions))
RMSE
```

```
## [1] 408.9694
```

This number by itself doesn't give us too much information. However, we could compare it to the RMSE of other models to compare the effectiveness of different models.

Comparing RMSE

We will compare it to the RMSE of the original model we created.

```
model2 <- glm(body_mass_g ~ bill_length_mm + bill_depth_mm + flipper_length_mm, data = training_set)
predictions2 <- predict.glm(model2, testing_set)
RMSE2 <- sqrt(sum((predictions2 - testing_set$body_mass_g)^2) / length(predictions))
RMSE2
```

```
## [1] 410.6734
```

We can see that the previous model had a slightly lower RMSE indicating better predictive power.

Multinomial Logistic Model

What is a Multinomial Logistic Model

Multinomial Logistic regression is used to predict a single categorical variable using one or more other variables. It extends the approach for situations where the independent variable has more than two categories. This model can be used for classification. If our dependent variable is a categorical variable, we would be able to predict the factor level based on other variables that could be continuous. The link function for this model is the generalized logit.

Assumptions of Multinomial Logistic Models

1. Linearity
2. No Outliers
3. Independence
4. No Multicollinearity

Setup

```
library(caret)
library(tidyverse)
require(nnet)
library(VGAM)
library(car)
penguins_df <- palmerpenguins::penguins
```

```
penguins_df %>%
  group_by(species, island) %>%
  summarise(n_records = n())
```

```
## # A tibble: 5 x 3
## # Groups:   species [3]
##   species   island    n_records
##   <fct>     <fct>         <int>
## 1 Adelie    Biscoe           44
## 2 Adelie    Dream            56
## 3 Adelie    Torgersen        52
## 4 Chinstrap Dream            68
## 5 Gentoo    Biscoe          124
```

We can see that the Chinstrap and Gentoo species only appear to inhabit one island while the Adelie species inhabits three different islands islands.

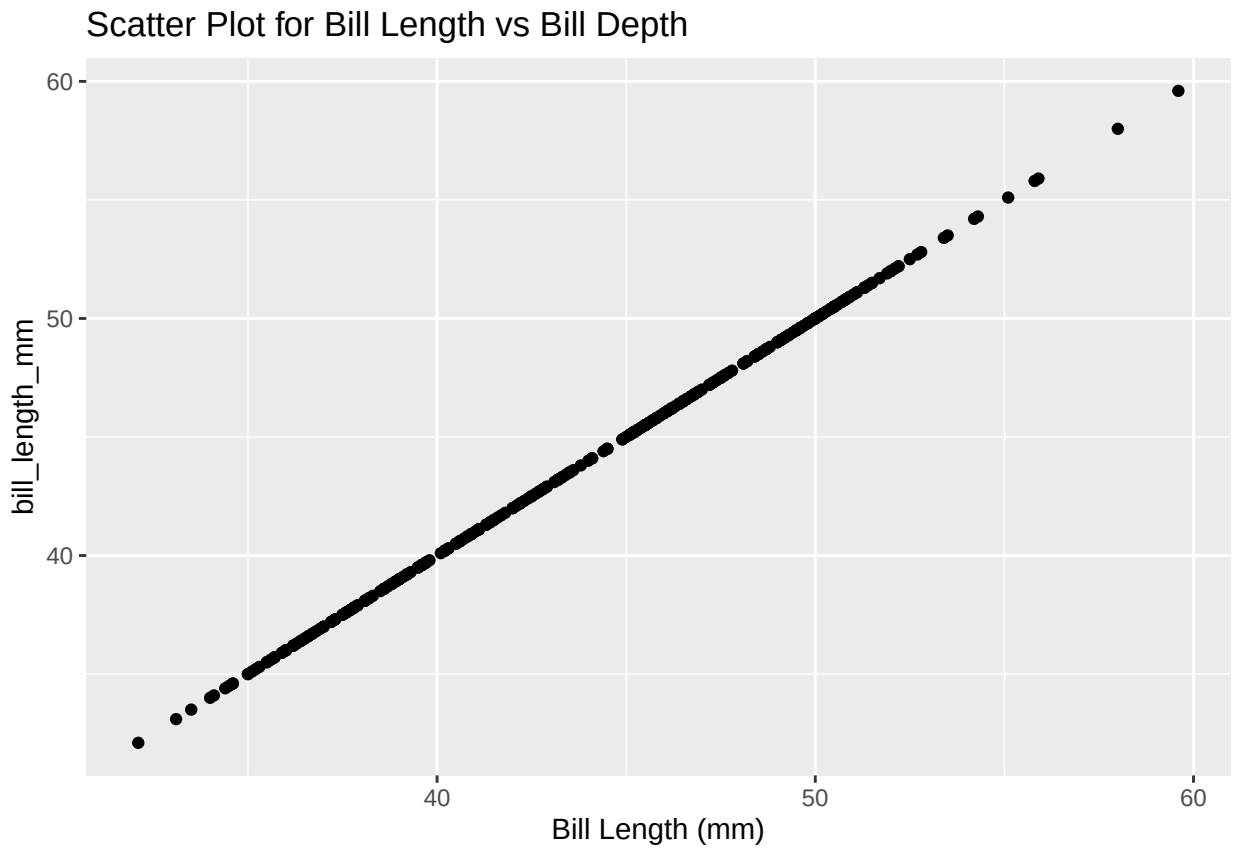
```
penguins_df <- penguins_df %>%
  mutate(species_binary = ifelse(species == 'Adelie', 'Adelie', 'Other'))

penguins_df$species_binary <- factor(penguins_df$species_binary, levels = c("Other", "Adelie"))
```

Checking Assumption 1

Linearity between the response and predictors.

```
ggplot(penguins_df, aes(x = bill_length_mm, y = bill_length_mm)) +
  geom_point() +
  ggtitle("Scatter Plot for Bill Length vs Bill Depth") +
  xlab("Bill Length (mm)")
```




```
  ylab("Bill Depth (mm")
```

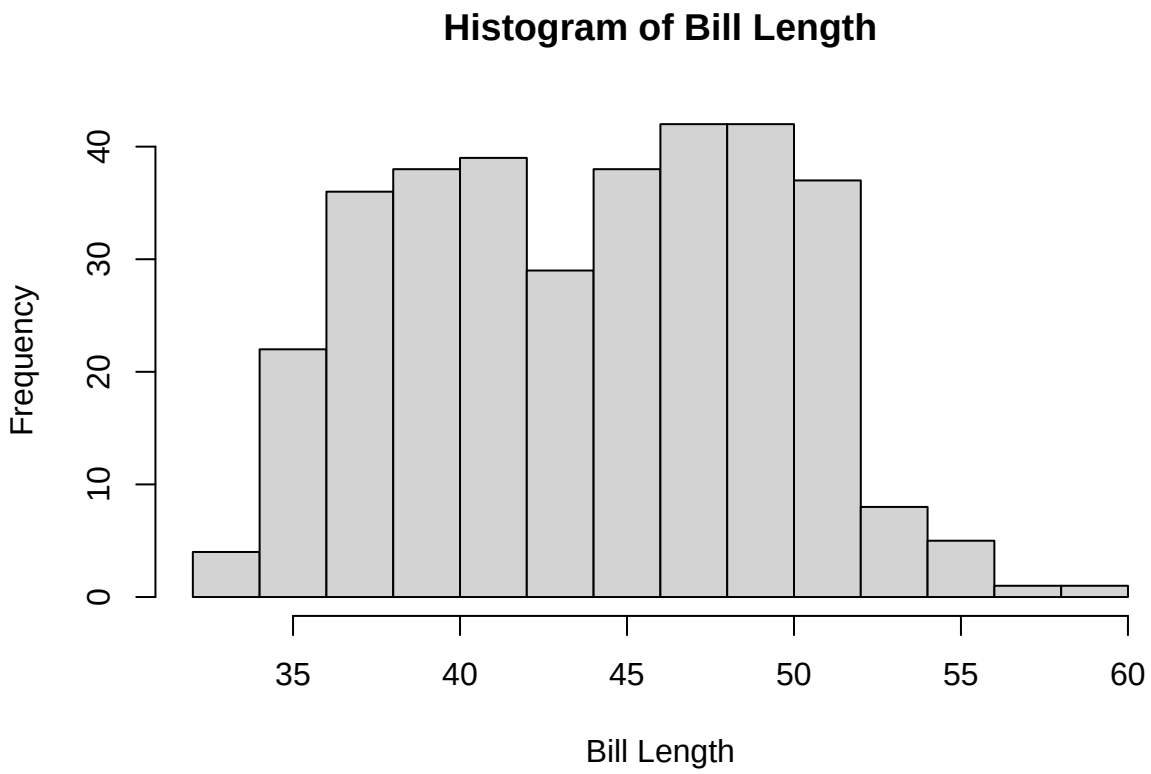
```
## $y
## [1] "Bill Depth (mm"
##
## attr(,"class")
## [1] "labels"
```

We can see that the relationship seems to be linear.

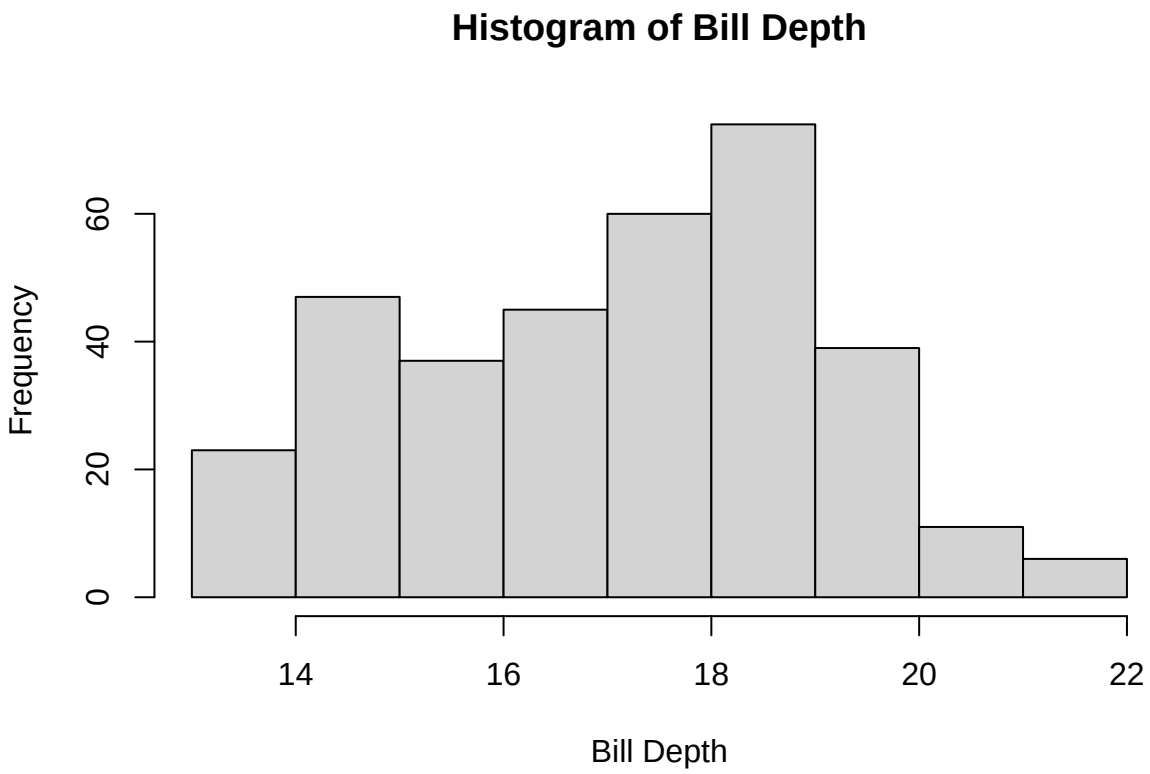
Checking Assumption 2

No significant outliers

```r
hist(penguins_df$bill_length_mm, xlab = "Bill Length", main = "Histogram of Bill Length")
```




```r
hist(penguins_df$bill_depth_mm, xlab = "Bill Depth", main = "Histogram of Bill Depth")
```



We can see that there are no significant outliers in the data.

Checking Assumption 3

Independence

```
fit <- vglm(species~island, multinomial, data = penguins_df)
anova(fit)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Model: 'multinomial', 'VGAMcategorical'
##
## Link: 'multilogitlink'
##
## Response: species
##
##        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## island  4   357.87       686     721.82 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With the low p-value we would reject the null hypothesis and conclude that there is independence.

Checking Assumption 4

Little or no multicollinearity between the predictors

```
penguins_df2 <- penguins %>%
  mutate(species = as.numeric(factor(species)),
         island = as.numeric(factor(island)),
         sex = as.numeric(factor(sex)))
model_of_interest <- lm(species ~ bill_length_mm + bill_depth_mm, data = penguins_df2)
vif(model_of_interest)
```

```
## bill_length_mm  bill_depth_mm
##       1.058481       1.058481
```

Since each of the variables we are using have a VIF < 5, multicollinearity is not an issue for our model.

Training Set, Testing Set, and Setting a Reference Level

The first step is the split the data into a training and testing set if both sets do not already exist for the desired data set.

```
index <- createDataPartition(penguins_df$species, p = 0.70, list = FALSE)
train <- penguins_df[index,]
test <- penguins_df[-index,]
```

Next we will set the reference level to the species Adelie since it is the only species that inhabits all the islands in this dataset.

```
train$species <- relevel(train$species, ref = "Adelie")
test$species <- relevel(test$species, ref = "Adelie")
```

Training the Model

```
multinom_model <- multinom(species ~ bill_length_mm + bill_depth_mm, data = train)
```

```
## # weights:  12 (6 variable)
## initial  value 263.666949
## iter  10 value 21.332388
## iter  20 value 13.452623
## iter  30 value 13.055288
## iter  40 value 12.481676
## iter  50 value 12.378270
## iter  60 value 12.270758
## iter  70 value 12.198297
## iter  80 value 12.140822
## iter  90 value 12.099324
## iter 100 value 12.091189
## final  value 12.091189
## stopped after 100 iterations
```

```
summary(multinom_model)
```

```
## Call:
## multinom(formula = species ~ bill_length_mm + bill_depth_mm,
##     data = train)
##
## Coefficients:
##           (Intercept) bill_length_mm bill_depth_mm
## Chinstrap   -21.74697       1.931521     -3.472306
## Gentoo       74.58060       2.747014    -11.526408
##
## Std. Errors:
##           (Intercept) bill_length_mm bill_depth_mm
## Chinstrap    13.97655      0.6683862      1.492480
## Gentoo       49.43504      0.7632730      3.991214
##
## Residual Deviance: 24.18238
## AIC: 36.18238
```

Computing p-values For the Regression Coefficients

```
(z <- summary(multinom_model)$coefficients / summary(multinom_model)$standard.errors )
```

```
##           (Intercept) bill_length_mm bill_depth_mm
## Chinstrap   -1.555961       2.889828     -2.326534
## Gentoo       1.508659       3.598993     -2.887945
```

```
(p <- (1 - pnorm(abs(z), 0, 1 )) *2 )
```

```
##           (Intercept) bill_length_mm bill_depth_mm
## Chinstrap   0.1197174   0.0038545251   0.019990091
## Gentoo      0.1313861   0.0003194518   0.003877675
```

Since all of the p-values are small we will reject the null hypothesis that the regression coefficients are equal to 0. In general, it is good practice to set an alpha level before starting any tests and using a Bonferroni correction when testing.

Converting the Coefficients to Odds by Taking the Exponentional of the Coefficients

```
exp(coef(multinom_model))
```

```
##            (Intercept) bill_length_mm bill_depth_mm
## Chinstrap 3.592616e-10       6.899998  3.104535e-02
## Gentoo    2.454382e+32      15.595994  9.866076e-06
```

By taking the exponential of the coefficients, we are able to see the change in the odds ratio with a 1 unit increase.

Model Prediction and Validation

```
multinom_preds <- predict(multinom_model, test, type = "class")
head(multinom_preds)
```

```
## [1] Adelie Adelie Adelie Adelie Adelie Adelie
## Levels: Adelie Chinstrap Gentoo
```

Viewing the First Few Predictions

```
head(round(fitted(multinom_model), 2))
```

```
##   Adelie Chinstrap Gentoo
## 1      1         0      0
## 3      1         0      0
## 4      1         0      0
## 5      1         0      0
## 6      1         0      0
## 7      1         0      0
```

Multinomial Regression predicts the probability of a particular observation.

Building a Classification Table

```
multinom_cm <- table(test$species, multinom_preds)
multinom_cm
```

```
##            multinom_preds
##             Adelie Chinstrap Gentoo
##   Adelie        45         0      0
##   Chinstrap      1        16      3
##   Gentoo         0         1     36
```

Calculating Accuracy

```
round((sum(diag(multinom_cm))/sum(multinom_cm))*100,2)
```

```
## [1] 95.1
```

ANCOVA

What is ANCOVA

ANCOVA stands for analysis of covariance. Simply put, ANCOVA is a combination of ANOVA and linear regression as it deals with categorical and continuous variables. It is similar to ANOVA, analysis of variance, which tests for differences in mean responses to a categorical factor level. ANCOVA differs from ANOVA because it includes a continuous covariate in the model. The job of the covariate is to remove the unnecessary variation from the response variable. ANCOVA is useful when the covariate has a linear relationship with the dependent variable and does not have relationship with the categorical variable.

Assumptions of ANCOVA

1. Linearity between covariate and response at each level of the grouping variable
2. Homogeneity of regression slopes
3. Outcome variable is normally distributed
4. Homoscedasticity for all groups
5. No significant outliers

Setup

```
library(palmerpenguins)
library(tidyverse)
library(rstatix)
library(car)
library(multcomp)
df <- penguins
```

palmerspenguins contains the dataset penguins that we will be working with

tidyverse contains packages such as dplyr that is used for data manipulation and ggplot2 which is used for graphing

rstatix contains the function anova_test

car contains the function Anova

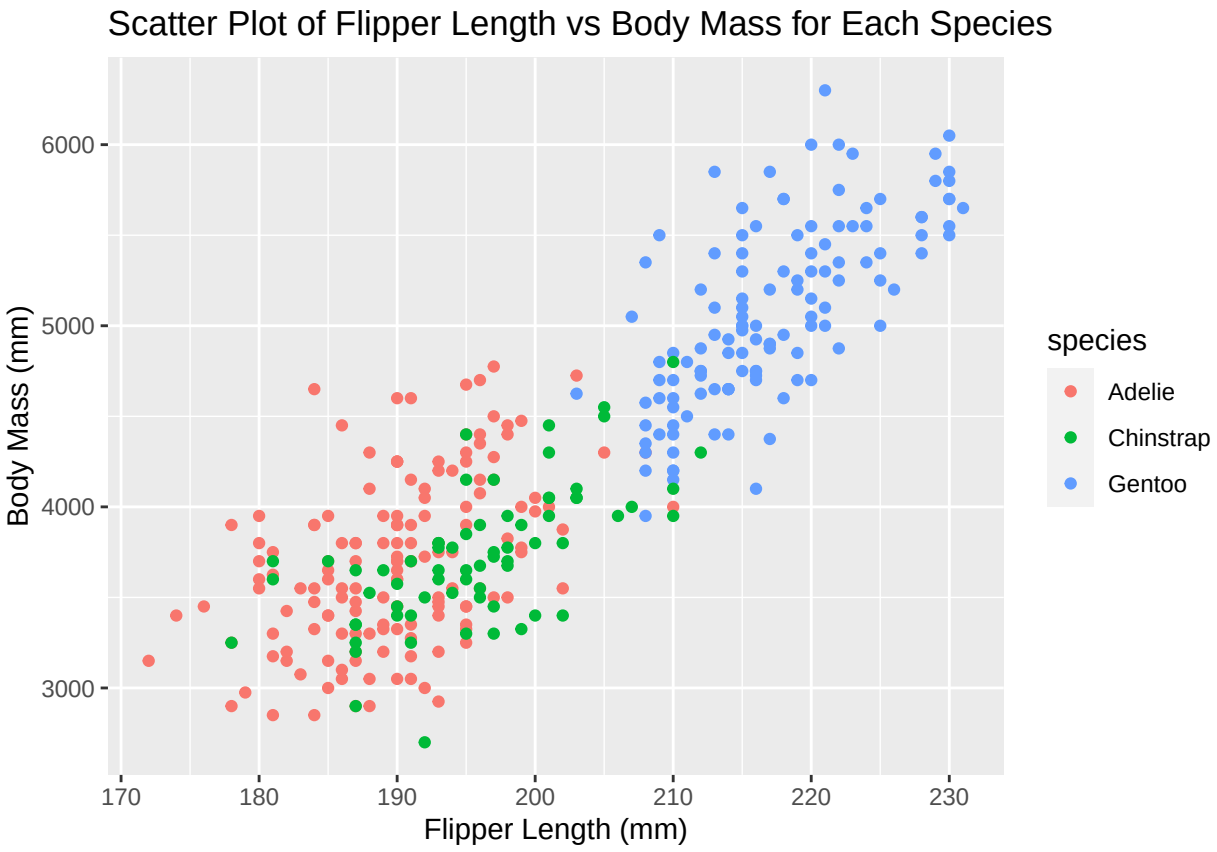
multcomp contains the function glht

Checking assumption 1

Linearity between covariate and response at each level of the grouping variable

This can be done by graphing a scatter plot of the predictor vs the covariate grouped by the categorical variable. In this case, a scatter plot of flipper length vs body mass separated by species.

```
ggplot(df, aes(x = flipper_length_mm, y = body_mass_g, color = species)) +
  geom_point() +
  ggtitle('Scatter Plot of Flipper Length vs Body Mass for Each Species') +
  xlab('Flipper Length (mm)') +
  ylab('Body Mass (mm)')
```



The predictor appears to be linear at every level of species.

Checking Assumption 2

Homogeneity of regression slopes

This can be checked by seeing if the interaction is significant between the group variable and the predictor.

```
anova_test(df, body_mass_g ~ flipper_length_mm + species + flipper_length_mm * species)
```

```
## ANOVA Table (type II tests)
##
##                       Effect DFn DFd       F        p p<.05   ges
## 1         flipper_length_mm   1 336 180.398 3.22e-33     * 0.349
## 2                   species   2 336  18.886 1.69e-08     * 0.101
## 3 flipper_length_mm:species   2 336   5.532 4.00e-03     * 0.032
```

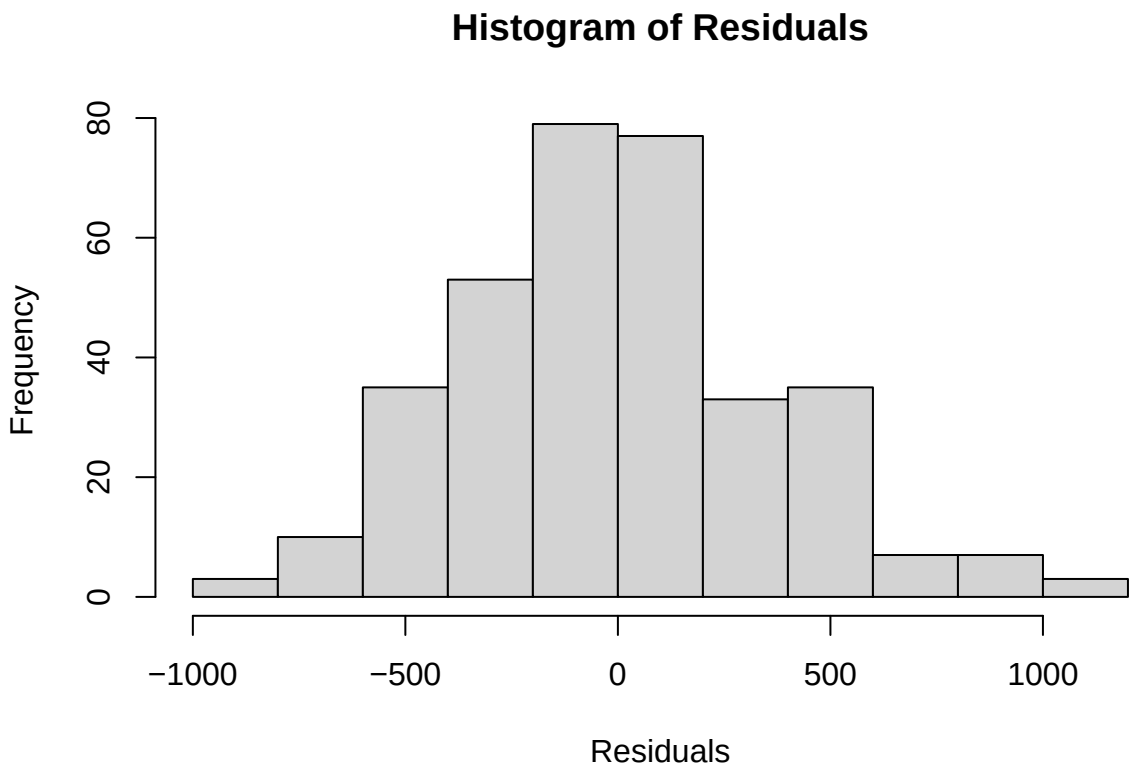
The interaction between flipper length and species is insignificant therefore not homogeneous. We will proceed for the sake of the exploration of ANCOVA but we will conclude that the results are invalid.

Checking Assumption 3

Outcome variable is normally distributed

This can be checked by viewing a histogram of the residuals

```r
model <- lm(body_mass_g ~ flipper_length_mm * species, data = df)
hist(model$residuals, xlab = 'Residuals', main = 'Histogram of Residuals')
```



The histogram of the residuals appears to be approximately normally distributed.

Checking Assumption 4

Homoscedasticity for all groups

This can be checked using Bartlett's Test

```r
newdf <- as.data.frame(cbind(df$species, model$residuals))
names(newdf) <- c('species', 'residuals')
bartlett.test(residuals ~ species, data = newdf)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  residuals by species
## Bartlett's K-squared = 8.2283, df = 2, p-value = 0.01634
```

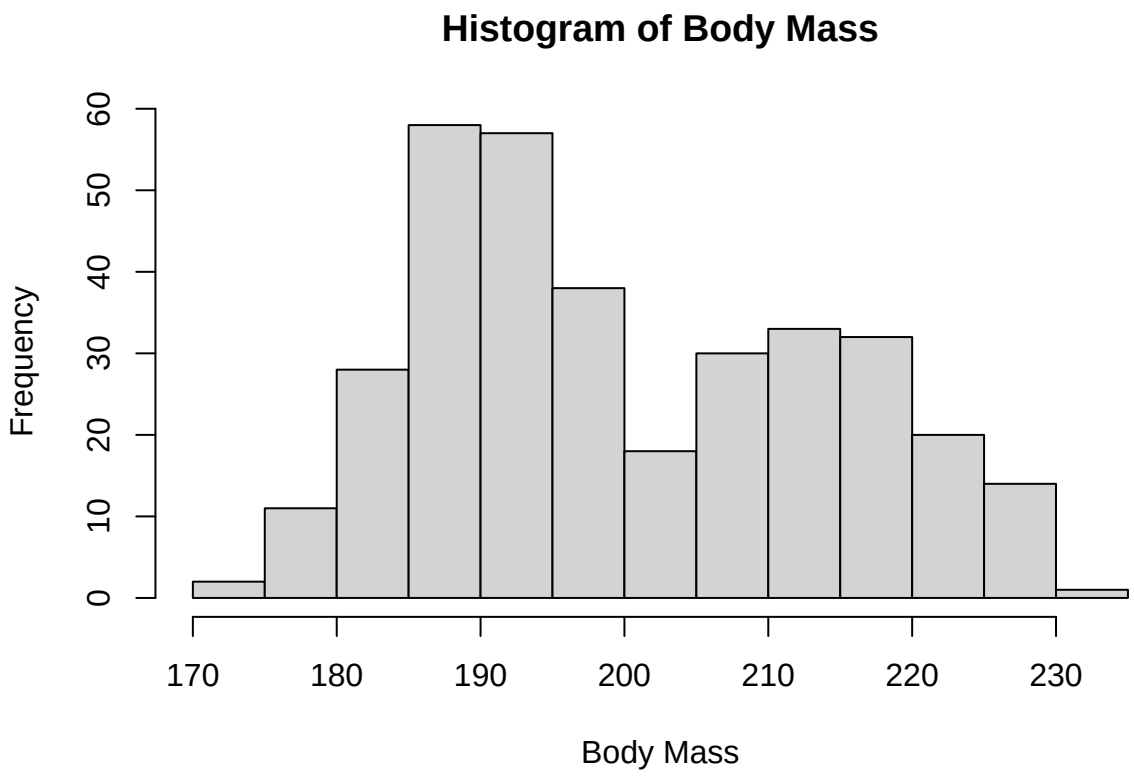
With a p-value of 0.3182 we would fail to reject the null hypothesis that the variances are the same across the three species if we were to use 0.01 as the alpha level. We can't use this assumption as it wouldn't result in failing to reject the null hypothesis.

Checking Assumption 5

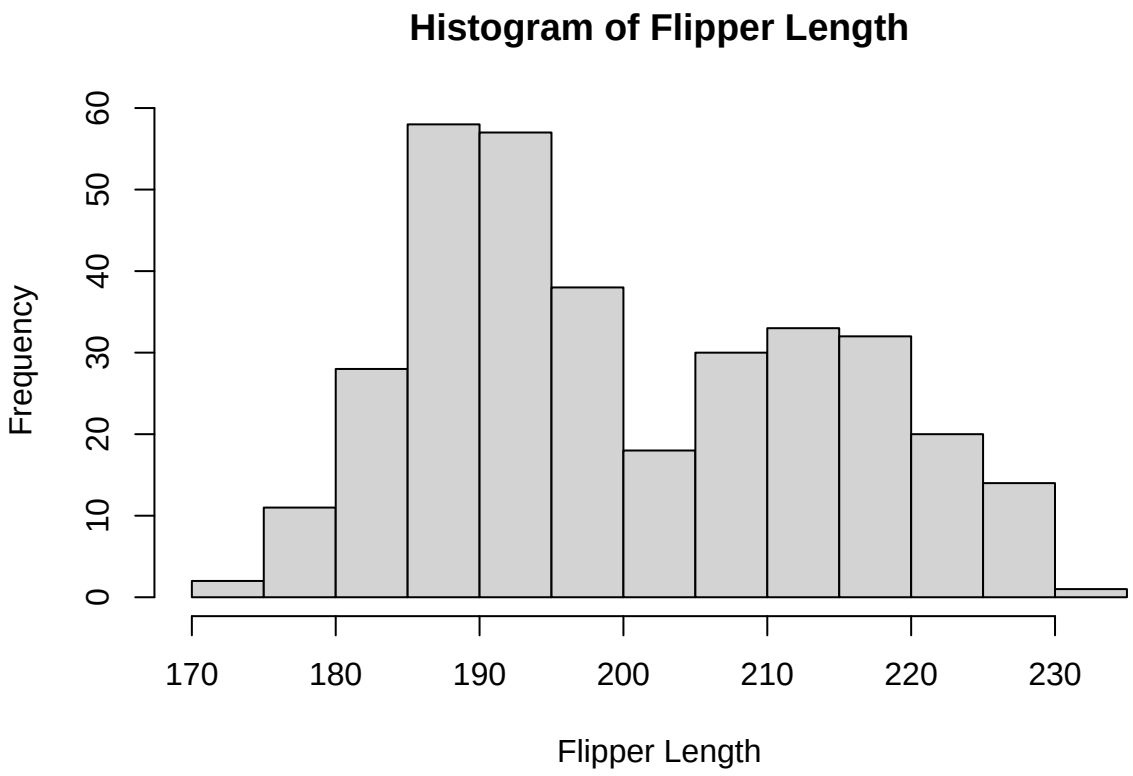
No significant outliers

This can be checked by viewing a histogram

```
hist(df$flipper_length_mm, xlab = 'Body Mass', main = 'Histogram of Body Mass')
```




```
hist(df$flipper_length_mm, xlab = 'Flipper Length', main = 'Histogram of Flipper Length')
```



There does not appear to be any significant outliers in the data.

With only 3 of our 5 assumptions holding, this exact ANCOVA would not yield valid results.

Running ANCOVA

We will perform ANCOVA with body mass as the response, flipper length as the covariate, and species as the factor variable.

```
fit <- aov(body_mass_g ~ flipper_length_mm + species, data = df)
Anova(fit, type = 'III')
```

```
## Anova Table (Type III tests)
##
## Response: body_mass_g
##                     Sum Sq  Df F value    Pr(>F)
## (Intercept)        6717050   1  47.630 2.551e-11 ***
## flipper_length_mm 24776495   1 175.687 < 2.2e-16 ***
## species            5187807   2  18.393 2.615e-08 ***
## Residuals         47666988 338
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When controlling for our covariate, flipper length, we can see that species has a significant impact on the body mass. We reject the null hypothesis and conclude that the mean body mass is not the same for each species.

The covariate, flipper length has a significant relationship with the body mass of penguins. There was also a significant effect of the species on the body mass after controlling for the effect of the flipper length.

Post Hoc Test

Next, we must run post hoc tests for comparing multiple means. We will perform this via Tukey contrasts.

```
posthoc <- glht(fit, linfct = mcp(species = 'Tukey'))
summary(posthoc)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = body_mass_g ~ flipper_length_mm + species, data = df)
##
## Linear Hypotheses:
##                          Estimate Std. Error t value Pr(>|t|)
## Chinstrap - Adelie == 0  -206.51      57.73  -3.577  0.00115 **
## Gentoo - Adelie == 0      266.81      95.26   2.801  0.01398 *
## Gentoo - Chinstrap == 0   473.32      86.75   5.456  < 0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

For the difference in means for each species, we reject the null hypothesis that they are equal and conclude that they are indeed not equal.

```
confint(posthoc)
```

```
##
##   Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: aov(formula = body_mass_g ~ flipper_length_mm + species, data = df)
##
## Quantile = 2.3355
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                          Estimate  lwr       upr
## Chinstrap - Adelie == 0 -206.5101 -341.3391  -71.6812
## Gentoo - Adelie == 0     266.8096   44.3228  489.2964
## Gentoo - Chinstrap == 0  473.3197  270.7264  675.9131
```

When we run a confidence interval for the difference in means, we can see that 0 is not in any of the intervals. In fact, none of the intervals are even close to 0.

Results

The Gaussian model yielded a model with only flipper length as a significant regressor to predict body mass with a lower RMSE than the original model (predicting body mass using bill length, bill depth, and flipper length).

The multinomial model showed that bill length and bill depth are both significant in predicting the species of penguin with a 94.06 accuracy.

The ANCOVA model showed that flipper length, the covariate, has a significant relationship with the body mass of penguins and when controlling for this covariate, species also had a significant impact on the body mass. When testing for a difference of means, it was concluded that they are not equal and a difference of means indeed exists.

Discussion

Literature Cited

Ancova Part I: Stat 502. PennState: Statistics Online Courses. (n.d.). Retrieved December 6, 2022, from https://online.stat.psu.edu/stat502_fa21/lesson/9

Cheng Hua, Y.-J. C. (2021, April 29). Chapter 11 Multinomial Logistic Regression. Companion to BER 642: Advanced regression methods. Retrieved December 6, 2022, from https://bookdown.org/chua/ber642_advanced_regression/multinomial-logistic-regression.html

Datasciencebeginners. (2020, May 27). Multinomial logistic regression with R: R-bloggers. R. Retrieved December 6, 2022, from https://www.r-bloggers.com/2020/05/multinomial-logistic-regression-with-r/

GLM in R: Learn how to construct generalized linear model in R. EDUCBA. (2021, October 22). Retrieved December 6, 2022, from https://www.educba.com/glm-in-r/

GLM: Fitting generalized linear models. RDocumentation. (n.d.). Retrieved December 6, 2022, from https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm

Gorman KB, Williams TD, Fraser WR (2014). Ecological sexual dimorphism and environmental variability within a community of Antarctic penguins (genus Pygoscelis). PLoS ONE 9(3):e90081. https://doi.org/10.1371/journal.pone.0090081

Huang, W.-M. (n.d.). ANCOVA (Analysis of Covariance). Retrieved December 6, 2022, from https://www.lehigh.edu/~wh02/ancova.html#:~:text=ANCOVA&text=Analysis%20of%20covariance%20is%20used,co%2Dvary%20with%

Introduction to Glms: Stat 504. PennState: Statistics Online Courses. (n.d.). Retrieved December 6, 2022, from https://online.stat.psu.edu/stat504/lesson/6/6.1

Introduction to palmerpenguins. • palmerpenguins. (n.d.). Retrieved December 6, 2022, from https://allisonhorst.github.io/palmerpenguins/articles/intro.html#highlights

Lani, J. (2021, August 11). Analysis of covariance (ANCOVA). Statistics Solutions. Retrieved December 6, 2022, from https://www.statisticssolutions.com/analysis-of-covariance-ancova/

MULTINOMIAL LOGISTIC REGRESSION | STATA DATA ANALYSIS EXAMPLES. UCLA Statistical Methods and Data Analytics. (n.d.). Retrieved December 6, 2022, from https://stats.oarc.ucla.edu/stata/dae/multinomiallogistic-regression/

Renard, M. (2021, April 13). Doing and reporting your first ANOVA and Ancova in R. Medium. Retrieved December 6, 2022, from https://towardsdatascience.com/doing-and-reporting-your-first-anova-and-ancova-in-r-1d820940f2ef