Categorical Predictors & Dummy Variables

Arjun Mannem

2023-12-05

# Load necessary packages  
library(tidyverse)

## Warning: package 'ggplot2' was built under R version 4.3.2

## Warning: package 'dplyr' was built under R version 4.3.2

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggthemes)  
library(flextable)

## Warning: package 'flextable' was built under R version 4.3.2

##   
## Attaching package: 'flextable'  
##   
## The following object is masked from 'package:purrr':  
##   
## compose

library(corrr)

## Warning: package 'corrr' was built under R version 4.3.2

library(ggfortify)  
library(broom)  
library(car)

## Warning: package 'car' was built under R version 4.3.2

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.3.2

##   
## Attaching package: 'car'  
##   
## The following object is masked from 'package:dplyr':  
##   
## recode  
##   
## The following object is masked from 'package:purrr':  
##   
## some

# Set ggplot theme for visualizations  
theme\_set(ggthemes::theme\_few())  
  
# Set default ggplot2 colors  
my\_palette <- ggthemes::colorblind\_pal()(8)[c(4, 6:8, 2:3, 5, 1)]  
  
# Set options for flextables  
set\_flextable\_defaults(na\_str = "NA")  
  
# Load function for printing tables nicely  
source("https://raw.githubusercontent.com/dilernia/STA323/main/Functions/make\_flex.R")  
  
# Load function for visualizing VIF & GVIF values  
source("https://raw.githubusercontent.com/dilernia/STA323/main/Functions/vif\_plot.R")

# Loading penguins data  
penguins <- readr::read\_csv("https://raw.githubusercontent.com/dilernia/STA323/main/Data/penguins.csv")

## Rows: 344 Columns: 8  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): species, island, sex  
## dbl (5): bill\_length\_mm, bill\_depth\_mm, flipper\_length\_mm, body\_mass\_g, year  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

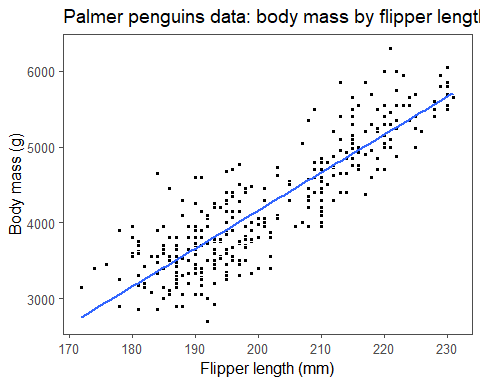
# Dropping missing values  
penguinsComplete <- penguins %>%   
 dplyr::filter(!is.na(sex)) %>%   
 dplyr::mutate(sex = fct\_relevel(sex, "male", "female"))  
  
# Printing sample of rows from data set  
penguinsComplete %>%   
 dplyr::select(species, sex, body\_mass\_g, flipper\_length\_mm) %>%   
 dplyr::sample\_n(size = 5) %>%   
 make\_flex(caption = "Five randomly selected rows and four variables from the Palmer Penguins data.")

Table 1: Five randomly selected rows and four variables from the Palmer Penguins data.

| species | sex | body\_mass\_g | flipper\_length\_mm |
| --- | --- | --- | --- |
| Adelie | male | 4,250.00 | 190.00 |
| Adelie | female | 3,325.00 | 195.00 |
| Gentoo | male | 5,750.00 | 222.00 |
| Chinstrap | male | 3,600.00 | 195.00 |
| Adelie | male | 3,550.00 | 194.00 |

# Creating a scatter plot of body mass by flipper length  
penguinsComplete %>%   
 ggplot(aes(x = flipper\_length\_mm, y = body\_mass\_g)) +   
 geom\_point(pch = 21, fill = "black",  
 color = "white") +   
 geom\_smooth(aes(x = flipper\_length\_mm, y = body\_mass\_g), method = "lm", se = FALSE) +  
 labs(y = "Body mass (g)",   
 x = "Flipper length (mm)",   
 title = "Palmer penguins data: body mass by flipper length")

## `geom\_smooth()` using formula = 'y ~ x'



# Fitting SLR model  
slrModel <- lm(body\_mass\_g ~ flipper\_length\_mm, data = penguinsComplete)  
  
# Printing SLR model output  
slrModel %>%   
 broom::tidy() %>%   
 make\_flex(ndigits = 1,  
 caption = "SLR model output.")

Table 2: SLR model output.

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -5,872.1 | 310.3 | -18.9 | <2e-16 |
| flipper\_length\_mm | 50.2 | 1.5 | 32.6 | <2e-16 |

**How do we interpret the estimated slope?**

For every 1mm increase in a penguin’s flipper length, we expect its body mass to increase by 50.2 grams.

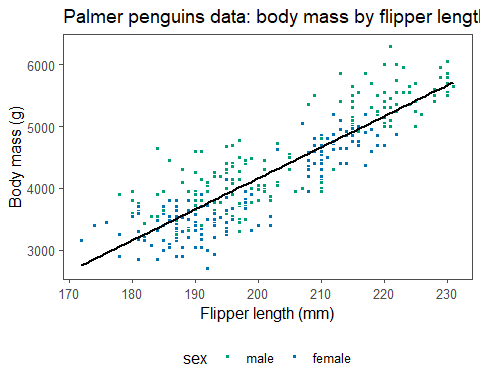
**How do we interpret the estimated intercept?**

Since 0 is outside the scope of the data set

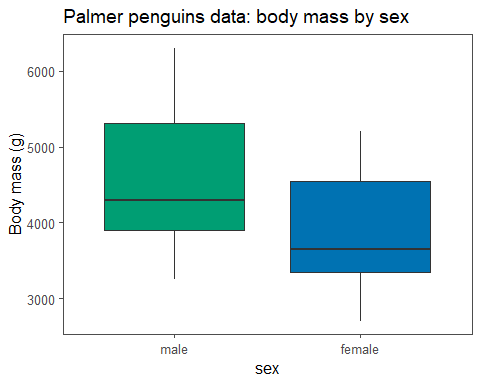
#### What if we wanted to incorporate categorical variables in our model?

# Creating scatter plot of body mass by flipper length coloring points by sex  
penguinsComplete %>%   
 ggplot(aes(x = flipper\_length\_mm, y = body\_mass\_g)) +   
 geom\_point(aes(x = flipper\_length\_mm, y = body\_mass\_g, fill = sex), pch = 21, color = "white") +   
 geom\_smooth(aes(x = flipper\_length\_mm, y = body\_mass\_g),   
 method = "lm", se = FALSE, color = "black") +  
 labs(y = "Body mass (g)", x = "Flipper length (mm)", title = "Palmer penguins data: body mass by flipper length") +  
 scale\_fill\_manual(values = my\_palette) +  
 theme(legend.position = "bottom")

## `geom\_smooth()` using formula = 'y ~ x'



# Creating side-by-side boxplots for body mass by sex  
penguinsComplete %>%   
 ggplot(aes(x = sex, y = body\_mass\_g, fill = sex)) +   
 geom\_boxplot() +   
 labs(y = "Body mass (g)", x = "sex", title = "Palmer penguins data: body mass by sex") +  
 scale\_fill\_manual(values = my\_palette) +   
 theme(legend.position = "none")



**Does a penguin’s body mass (g) appear to vary based on its sex? Do male or female penguins have larger average body masses?**

### Categorical Predictors

Example 1: Predicting how many points a basketball team will score using playing location (home, away, or a neutral court)

Example 2: Estimating the effect of different treatments on blood pressure (Drug A, Drug B, and a placebo)

**How many dummy variables would be needed to incorporate the categorical predictors in Examples 1 and 2?**

Example 1: 2 dummy variables

Example 2: 2 dummy variables

### Palmer Penguins Data

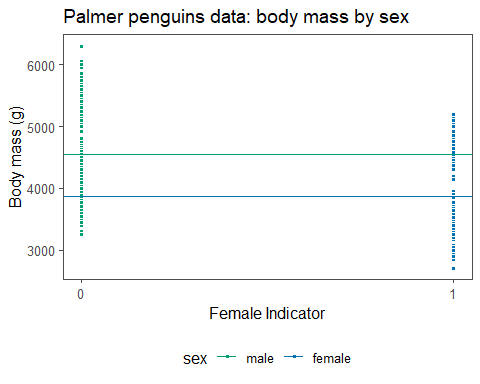
# Set seed value for reproducible random number generation  
set.seed(1994)  
  
# Creating explicit dummy variable for sex of the penguin  
penguinsDummy <- penguinsComplete %>%   
 dplyr::mutate(sexfemale = ifelse(sex == "female", 1, ifelse(sex == "male", 0, NA)))  
  
# Displaying data with dummy variable  
penguinsDummy %>%   
 dplyr::select(species, sex, sexfemale, body\_mass\_g, flipper\_length\_mm) %>%   
 sample\_n(size = 5) %>%  
 make\_flex(caption = "Five randomly selected rows from the Palmer Penguins data displaying a dummy variable.")

Table 3: Five randomly selected rows from the Palmer Penguins data displaying a dummy variable.

| species | sex | sexfemale | body\_mass\_g | flipper\_length\_mm |
| --- | --- | --- | --- | --- |
| Adelie | male | 0.00 | 4,600.00 | 191.00 |
| Chinstrap | female | 1.00 | 3,575.00 | 190.00 |
| Chinstrap | male | 0.00 | 3,900.00 | 196.00 |
| Gentoo | male | 0.00 | 5,200.00 | 219.00 |
| Chinstrap | male | 0.00 | 3,950.00 | 206.00 |

#### Modeling Penguin Body Mass: Theoretical Model

# Group means for plotting line of best fit  
bodyMassMeans <- penguinsDummy %>%   
 dplyr::group\_by(sex) %>%   
 dplyr::summarize(Mean = mean(body\_mass\_g),   
 sexfemale = mean(sexfemale))  
  
# Visualizing group means  
penguinsDummy %>%   
 ggplot(aes(x = sexfemale, y = body\_mass\_g,   
 fill = sex)) +   
 geom\_point(pch = 21, color = "white") +  
 geom\_hline(data = bodyMassMeans,   
 aes(yintercept = Mean, color = sex)) +  
 labs(y = "Body mass (g)", x = "Female Indicator", title = "Palmer penguins data: body mass by sex") +  
 scale\_fill\_manual(values = my\_palette) +   
 scale\_color\_manual(values = my\_palette) +   
 scale\_x\_continuous(breaks = c(0, 1)) +  
 theme(legend.position = "bottom")



# Fitting SLR model with sex as predictor  
slrModel2 <- lm(body\_mass\_g ~ sex,   
 data = penguinsComplete)  
  
# Printing SLR model output  
slrModel2 %>%   
 broom::tidy() %>%   
 make\_flex(ndigits = 1,  
 caption = "SLR model output.")

Table 4: SLR model output.

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 4,545.7 | 56.3 | 80.7 | <2e-16 |
| sexfemale | -683.4 | 80.0 | -8.5 | 5e-16 |

**Which group is the baseline / reference category?**

The male category is the baseline / reference category since it does not have a dummy variable as a predictor in the model.

**How do we interpret the estimated slope?**

Since βHat1 = -683.4, we expect that female penguins have body masses that are 683.4 grams less than that of male penguins, on average.

**How do we interpret the estimated intercept?**

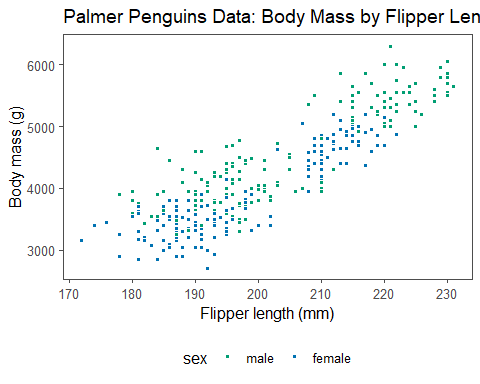
Since βHat0 = 4,545.7, this means that the expected body mass of the male penguins is 4,545.7 grams.

**What is the expected body mass (g) of a female penguin?**

yHat = 4545.7 − 683.4 \* 1 = 3862.3 grams

#### Modeling Penguin Body Mass

# Creating scatter plot of body mass by flipper length colored by sex  
penguinsComplete %>%   
 ggplot(aes(x = flipper\_length\_mm, y = body\_mass\_g,  
 fill = sex)) +   
 geom\_point(pch = 21, color = "white") +   
 labs(y = "Body mass (g)", x = "Flipper length (mm)", title = "Palmer Penguins Data: Body Mass by Flipper Length") +  
 scale\_fill\_manual(values = my\_palette) +   
 theme(legend.position = "bottom")



# Fitting MLR model  
mlrModel <- lm(body\_mass\_g ~ sex + flipper\_length\_mm,  
 data = penguinsComplete)  
  
# Extracting MLR model coefficients  
betaHats <- coefficients(mlrModel)  
  
# Printing MLR model output  
mlrModel %>%   
 broom::tidy() %>%   
 make\_flex(ndigits = 1,  
 caption = "MLR model output.")

Table 5: MLR model output.

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -5,062.4 | 296.0 | -17.1 | <2e-16 |
| sexfemale | -347.9 | 40.3 | -8.6 | 3e-16 |
| flipper\_length\_mm | 47.0 | 1.4 | 32.6 | <2e-16 |

**How do we interpret the estimated slope for the sex of the penguin?**

Since βHat1 = -347.9, we expect female penguins have body masses that are 347.9 grams less than that of male penguins with same flipper length.

**How do we interpret the estimated intercept?**

Since flipper length = 0 is outside the scope of the data set, interpreting the intercept would be extrapolation, and thus is inappropriate.

**How do we interpret the estimated slope for the flipper lengths of the penguins?**

Since βHat2 = 47.0, for every 1mm increase in a penguin’s flipper length, we expect its body mass to increase by 47.0 grams.

#### Modeling Penguin Body Mass using Species

**How many dummy variables are used to represent the species variable?**

2 dummy variables since there are 3 species.

**Based on the model above, which species is the baseline / reference category? Why?**

Adelie species is the baseline / reference category,since it does not have a dummy variable.( X1=1 for an chinstrap penguin and 0 otherwise, X2=1 for a gentoo penguin and 0 otherwise.)

# Set seed value for reproducible random number generation  
set.seed(1994)  
  
# Creating explicit dummy variables for species  
penguinsDummy2 <- penguinsComplete %>%   
 dplyr::mutate(speciesChinstrap = ifelse(species == "Chinstrap", 1, 0),  
 speciesGentoo = ifelse(species == "Gentoo", 1, 0))  
  
# Displaying data with dummy variabels for species  
penguinsDummy2 %>%   
 dplyr::select(species, speciesChinstrap, speciesGentoo, body\_mass\_g, flipper\_length\_mm) %>%   
 dplyr::sample\_n(size = 5) %>%  
 make\_flex(caption = "Five randomly selected rows from the Palmer Penguins data displaying dummy variables.")

Table 6: Five randomly selected rows from the Palmer Penguins data displaying dummy variables.

| species | speciesChinstrap | speciesGentoo | body\_mass\_g | flipper\_length\_mm |
| --- | --- | --- | --- | --- |
| Adelie | 0.00 | 0.00 | 4,600.00 | 191.00 |
| Chinstrap | 1.00 | 0.00 | 3,575.00 | 190.00 |
| Chinstrap | 1.00 | 0.00 | 3,900.00 | 196.00 |
| Gentoo | 0.00 | 1.00 | 5,200.00 | 219.00 |
| Chinstrap | 1.00 | 0.00 | 3,950.00 | 206.00 |

# Fitting linear model with species as predictor  
speciesModel <- lm(body\_mass\_g ~ species,   
 data = penguinsComplete)  
  
betaHats <- coefficients(speciesModel)  
# Printing MLR model output  
speciesModel %>%   
 broom::tidy() %>%   
 make\_flex(ndigits = 1,  
 caption = "MLR model output.")

Table 7: MLR model output.

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 3,706.2 | 38.1 | 97.2 | <2e-16 |
| speciesChinstrap | 26.9 | 67.7 | 0.4 | 0.7 |
| speciesGentoo | 1,386.3 | 56.9 | 24.4 | <2e-16 |

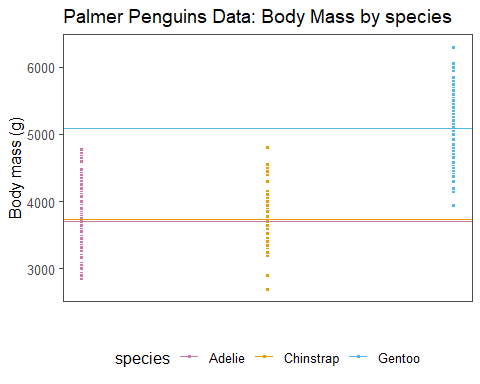
yHat = 3706.2 + 26.9X1 + 1386.3X2

**Interpretation of estimated intercept**: Since βHat0 = 3706.2, the expected bodymass of an Adelie penguin is 3706.2 grams.

**Interpretation of estimated slope for Chinstrap penguins**: Since βHat1 = 26.9, the expected bodymass of an Chinstrap penguin is 26.9 grams greater than that of Adelie species. Here, 3706.2 + 26.9 = 3733.1 grams.

**Interpretation of estimated slope for Chinstrap penguins**: Since βHat2 = 1386.3, the expected bodymass of an Chinstrap penguin is 1386.3 grams greater than that of Adelie species. Here, 3706.2 + 1386.3 = 5092.5 grams.

# Group means for plotting line of best fit  
bodyMassMeans <- penguinsDummy %>%   
 dplyr::group\_by(species) %>%   
 dplyr::summarize(Mean = mean(body\_mass\_g))  
  
# Plotting group data  
penguinsDummy %>%   
 ggplot(aes(x = as.integer(as.factor(species)), y = body\_mass\_g,   
 fill = species)) +   
 geom\_point(pch = 21, color = "white") +  
 labs(y = "Body mass (g)", x = "", title = "Palmer Penguins Data: Body Mass by species") +  
 geom\_hline(data = bodyMassMeans,   
 aes(yintercept = Mean, color = species)) +  
 scale\_fill\_manual(values = my\_palette[4:6]) +  
 scale\_color\_manual(values = my\_palette[4:6]) +  
 scale\_x\_continuous(breaks = 0) +  
 theme(legend.position = "bottom")



# Fitting linear models  
speciesFlipModel <- lm(body\_mass\_g ~ species + flipper\_length\_mm, data = penguinsComplete)  
  
betaHats <- coefficients(speciesFlipModel)  
  
# Printing MLR model output  
speciesFlipModel %>%   
 broom::tidy() %>%   
 make\_flex(ndigits = 1,  
 caption = "MLR model output.")

Table 8: MLR model output.

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -4,013.2 | 586.2 | -6.8 | 4e-11 |
| speciesChinstrap | -205.4 | 57.6 | -3.6 | 4e-04 |
| speciesGentoo | 284.5 | 95.4 | 3.0 | 0.003 |
| flipper\_length\_mm | 40.6 | 3.1 | 13.2 | <2e-16 |

**How do we interpret the estimated intercept?**

βHat0 = -4013.2

Since a flipper length of 0mm is outside the range of our observed data set, it would be **extrapolation** to interpret this value, and thus this is inappropriate.

**How do we interpret the estimated slopes?**

1. Since βHat\_Chinstrap = -205.4, the expected body mass in grams of Chinstrap penguins is 205.4 grams less than that of the Adelie penguins with the same flipper lengths.
2. Since βHat\_Gentoo = 284.5, the expected body mass in grams of Gentoo penguins is 284.5 grams more than that of the Adelie penguins with the same flipper lengths.
3. Since βHat\_Flipperlength = 40.6, for every 1mm increase in penguin’s flipper length, we expect its body mass in grams to increase by 40.6 grams, holding its species constant.

**What is the expected body mass (g) of a Chinstrap penguin with a flipper length of 195mm? Is this extrapolation?**

yHat = -4013.2 - 205.4 \* (x1) +284.5 \* (x2) + 40.6 \* (x3)

yHat = -4013.2 - 205.4 \* (1) +284.5 \* (0) + 40.6 \* 195

yHat = 3698.4 grams

No, this is not extrapolation since 195mm is within the range of the observed flipper lengths for the Chinstrap penguins.

# Obtaining predicted value  
predict(speciesFlipModel,   
 newdata = tibble(species = "Chinstrap",  
 flipper\_length\_mm = 195))

## 1   
## 3699.648

## 1   
## 3699.648

**If a Chinstrap penguin with a flipper length of 195mm had a body mass of 3900 grams, what is its corresponding residual? What does this mean?**

residual = y - yHat

residual = 3900 - 3698.4 = 201.6 grams

This penguin weighed about slightly more(200 grams) than our model expected, given its species and flipper lengths.

**What is the expected body mass (g) of a Chinstrap penguin with a flipper length of 225mm? Is this extrapolation?**

Since 225mm is outside the range of the observed flipper lengths for the Chinstrap penguins, such a predicted value would be extrapolation, and thus is inappropriate to calculate.

**Does it seem reasonable to constrain the regression lines to be parallel for each group?**

Since parallel lines fit the data reasonably well, it does seem reasonable to constrain the regression lines to be parallel for each group.