# Analysis of Environmental Data

## Deck 7 Regression Modeling

Michael France Nelson

Eco 602 – University of Massachusetts, Amherst – Fall 2021

Michael France Nelson

# Announcements: Oct 26th

- In-Class Likelihood group self-select *should be fixed.
  - I had it associated with the wrong grouping in Moodle
- For this week
  - Tuesday
    - Finish in-class likelihood
    - Start Deck 7
  - Thursday
    - Continue Deck 7
    - In-class confidence intervals

# Announcements: Oct 28th

- Today: In-class confidence intervals
  - Critical values
  - CI calculations
- Today: Special lab 7 office hours: noon – 3PM
  - In-person (in my office) or virtual (via course Zoom channel)
- Next week:
  - Review of in-class likelihood
  - Finish Deck 7, start Deck 8
  - Ginkgo data collection

# Announcements: Nov 2

- Week 10 reading questions
  - I want to push these back by 1 week, they'll be converted to "week 11 questions" and due date will be adjusted.
  - Final reading list will be updated.
  - Subsequent weeks' questions may also need to be adjusted – stay tuned for more info on Thursday.
- Ginkgoes aren't quite ready – stay tuned.

# Announcements: Nov 2

- Today:
  - Finish Deck 7
  - Review in-class likelihood (if there's time)
  - In-class t-tests

- Thursday
  - Start Deck 8
  - In-class regression

# Announcements: Nov 4

- Today:
  - Finish Deck 7
  - Review in-class likelihood
  - Review in-class t-tests
- Using Models 1 is due Sunday
- Using Models 2 will be available later today - Due

# Model Coefficients and the ANOVA Table

# What's in This Section?

Take-Home Concepts

- Interpreting model coefficient tables for categorical variables
- Interpreting model coefficient tables for continuous variables
- Interpreting the ANOVA table
- Intro to dummy variables

# Group 1 model interpretation

**Group 1 models are *linear in the parameters***

This makes the interpretation of model terms *relatively* easy.

- But note, there is still lots of complexity especially when we mix continuous and categorical terms and interaction terms.

Recall the basic equation:

$$y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \epsilon$$

- When all of the predictor variables have a value of zero, we expect $y$ to have a value of $\alpha$, on average.
- For every 1-unit change in $x_1$ we expect a $\beta_1$-unit change in $y$, on average.

# Group 1 model summary presentations

Table of model coefficients model summary.

- This table tells us the strength of effects of predictors, overall model significance test

ANOVA table.

- This table shows the model variability attributed to each factor, factor-specific significance tests

# Group 1 model interpretation

| Model Coefficients | ANoVA Table |
| --- | --- |
| Intercept: What is the value of the response when the predictor has value zero?<br>Slope: What is the change in the response with each unit change in the predictor?<br>Standard Errors: shape of sampling distribution<br>F-test: overall model significance test | Degrees of freedom: Reflects the number of samples, number of factor levels, number of individuals per factor level etc.<br>Sum of squares: Reflects the total squared deviation from the mean explained by a source.<br>Mean squares: Mean SS due to a source (per DF)<br>F tests: Test for ratio of variability explained by a particular predictor variable |

# ANOVA table vs. model coefficient table

| Model coefficient table tells you | ANOVA table tells you |
| --- | --- |
| 1. Intercept and slope coefficients<br>2. Overall model significance test, correlation test | 1. Variability explained by each factor in the model<br>2. Significance tests for each factor separately |

# 1-way ANOVA

When we have a continuous response and a single categorical predictor with 2 levels we can use a t-test.

What if there are 3 or more levels?

- The t-test is not enough.
- Analysis of Variance is a generalization of the t-test for 3 or more groups.

# Model Coefficient Tables: Dummy Variables

When you fit a model using a categorical predictor with n levels, the algorithm first detects all of the factor levels present in the data, then creates a set of n - 1 *dummy variables*.

- The dummy variables allow the model-building process to treat each factor level as if it were a separate, numerical predictor that can take on only values of zero or one.

| species | speciesGentoo | speciesChinstrap |
|---|---|---|
| Adelie | 0 | 0 |
| Gentoo | 1 | 0 |
| Chinstrap | 0 | 1 |

# Model Coefficient Tables: Interpretation for Categorical Predictors

Since each factor level is treated as a predictor variable, there will be slope parameters for each.

When R builds a model, it selects one of the factor levels to serve as the *base case*.

- When the model contains only categorical variables, the base case is analogous to the *intercept* term in a model, i.e. the $\alpha$.

It'll be easier to understand with an example.

# 1-way ANOVA: Palmer Penguins

The procedure for conducting an ANOVA in R is:

- Create a linear model fit with lm().
- Use anova() to perform the Analysis of Variance and print the ANOVA table.

Recall that ANOVA is really a just a different way of looking at a linear model.

- To better understand the relationship, we'll focus on the model coefficient table first:

```
lm(
  formula = body_mass_g ~ species,
  data = penguins)
```

```
Call:
lm(formula = body_mass_g ~ species,
 data = penguins)

Coefficients:
      (Intercept)
          3700.66
speciesChinstrap
            32.43
   speciesGentoo
          1375.35
```

# Factor Base Cases

There are slopes for Chinstrap and Gentoo, but where is the Adelie coefficient?

- Recall: the *base case* is the intercept in a 1-way ANOVA.

R assigned “Adelie” to be the base case.

- Notice how R formats the factor-level coefficient names:
  - the variable name prepended to the factor level.

# Interpreting the Coefficient Table

```
Call:
lm(formula = body_mass_g ~ species,
 data = penguins)

Coefficients:
     (Intercept)
         3700.66
speciesChinstrap
           32.43
   speciesGentoo
         1375.35
```

- Mean Adelie penguin mass is 3700 grams
- Mean Chinstrap penguin mass is 3700 + 32 grams
- Mean Gentoo penguin mass is 3700 + 1375 grams

Everything is relative to the base case!

# Interpreting the Coefficient Table

```
Call:
lm(formula = body_mass_g ~ species,
 data = penguins)

Coefficients:
      (Intercept)
          3700.66
speciesChinstrap
            32.43
   speciesGentoo
          1375.35
```

- The intercept is 3700 grams: Adelie penguins weigh 3700g, on average
- The regression slope for Chinstrap is 32 grams per unit.
  - Adding one ‘penguin unit’ increases the penguin mass by 32 grams, on average.
- The regression slope for Gentoo slope 1375 grams
  - Adding one ‘penguin unit’ increases the penguin mass by 1375 grams, on average.

Everything is relative to the base case!

# Interpreting the Coefficient Table

```
Call:
lm(formula = body_mass_g ~ species,
 data = penguins)

Coefficients:
     (Intercept)
         3700.66
speciesChinstrap
           32.43
   speciesGentoo
         1375.35
```

We can obtain the mean masses of each species from the model coefficient table.

- Mean Chinstrap penguin mass
  - $3733 = 3701 + 1 \times 32 + 0 \times 1375$
- Mean Gentoo penguin mass:
  - $5076 = 3701 + 0 \times 32 + 1 \times 1375$

# Dummy Variables

If we consider $x_{chin}$ a dummy variable which is equal to 1 if the observation is a Chinstrap penguin and 0 otherwise, and likewise for $x_{gentoo}$ we could write the regression equation symbolically as:

$$y_i = \alpha_{adelie} + \beta_{chin} \times x_{chin} + \beta_{gentoo} \times x_{gentoo}$$

What would the coefficient table and equation look like if Chinstrap penguins were lighter than Adelie penguins?

# 1-way ANOVA: ANOVA Table

We have examined the model coefficients and calculated the group means.

- The masses seem pretty different, but how could we assess the ANOVA *alternative* hypothesis?
  - "The body masses of penguins for *at least one* species are different from the masses of the other species"

```
## Analysis of Variance Table
##
## Response: body_mass_g
##            Df    Sum Sq  Mean Sq F value    Pr(>F)
## species     2 146864214 73432107  343.63 < 2.2e-16 ***
## Residuals 339  72443483   213698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# 1-way ANOVA: Model Coefficient Table

What can we learn from the model coefficient table?

The *intercept* and *speciesGentoo* coefficients have low p-values, but that's not exactly what we wanted to know!

- We wanted to know about the penguin species *in general*.

# 1-way ANOVA: ANOVA Table

The ANOVA table gives us a clue

```
Response: body_mass_g
           Df    Sum Sq  Mean Sq F value    Pr(>F)
species     2 146864214 73432107  343.63 < 2.2e-16 ***
Residuals 339  72443483   213698
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

# The ANoVA Table

- Note how the *species* predictor is now a single line.
  - There were model coefficients for each factor level.

```
Response: body_mass_g
           Df    Sum Sq  Mean Sq F value    Pr(>F)
species     2 146864214 73432107  343.63 < 2.2e-16 ***
Residuals 339  72443483   213698
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

# Model Coefficients and ANOVA Provide Complementary Information

We'll cover model coefficient interpretation, and the ANOVA table details in greater depth, but for now you should notice:

- Model slope/intercept coefficients: there is one coefficient for each *factor level* of a *categorical predictor*.
- The intercept coefficient corresponds to the *base case*.
- Model coefficient table characterizes the strength and significance of individual intercept and slope coefficients.
    - It *does not* tell us about the overall significance of the categorical predictor.
- The ANOVA table evaluates the ANOVA null hypothesis.
    - It *does not* tell us *which factor levels* are different
    - The two tables each provide part of the picture.

Neither the model coefficient table nor the ANOVA table tell us if a particular pair of factor levels are *significantly* different form one another!

# Model Coefficients and ANOVA Provide Complementary Information

Neither the model coefficient table nor the ANOVA table tell us whether a particular pair of factor levels are *significantly* different form one another!

- This is the realm of post-hoc testing.
  - Post-hoc testing is an analysis you perform after (post) you perform the initial analysis (hoc).
- The Tukey Honest Significant Difference is a common post-hoc method.

# Key Concepts

- Interpreting model coefficient tables for categorical variables
- Interpreting model coefficient tables for continuous variables
- Interpreting the ANOVA table
- Intro to dummy variables

# Chalkboard Model Art

**Dummy Variable Interpretation**

- Predictor variable adds one unit of Gentoo
- The coefficient is 1375
- One-unit increase in Gentoo corresponds to a 1375-unit increase in body mass