Module 4 Assignment 2

Ellen Bledsoe

2022-12-02

Assignment Details

Purpose

The goal of this assignment is to assess your ability to perform and interpret multiple regressions.

Task

Write R code which produces the correct answers and correctly interpret the results of visualizations and statistical tests.

Criteria for Success

- Code is within the provided code chunks
- Code is commented with brief descriptions of what the code does
- Code chunks run without errors
- Code produces the correct result
  - Code that produces the correct answer will receive full credit
  - Code attempts with logical direction will receive partial credit
- Written answers address the questions in sufficient detail

Due Date

April 27 at midnight MST

Assignment Questions

In this assignment, we will continue exploring data that will inform where we should build our fishing roads.

We've looked at data for Antarctic hair grass, a vascular plant. Now we are going to take into consideration two groups of non-vascular plants: mosses and liverworts.

Penguins are important players in the Antarctic ecosystem because they cycle nutrients from the ocean onto the land (or ice). Penguin poop is high in nutrients that plants need, such as nitrogen and phosphorus. We are curious to discover if the density of penguins at a given site corresponds to how much area of each site is covered in plants.

Set-Up

1. As always, let's load the `tidyverse` to get started.

```
library(tidyverse)
```

2. Now read in the data set, which is called `nonvascular_plants.csv`. Name the data frame `plants`.

```
plants <- read_csv("../data/nonvascular_plants.csv")
```

```
## Rows: 200 Columns: 4
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): plant_type
## dbl (3): site, percent_plant_cover, penguin_density_m2
##
## 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.
```

3. Take a look at the data; use the `head()` and `tail()` functions to look at the beginning of the data set and the end of the data set respectively. (2 points)

```
head(plants)
```

```
## # A tibble: 6 x 4
##    site plant_type percent_plant_cover penguin_density_m2
##   <dbl> <chr>                    <dbl>              <dbl>
## 1     1 moss                      47.5              1.89
## 2     2 moss                      39.5              1.18
## 3     3 moss                      39.3              1.81
## 4     4 moss                      40.9              1.63
## 5     5 moss                      45.4              0.843
## 6     6 moss                      36.7              0.613
```

```
tail(plants)
```

```
## # A tibble: 6 x 4
##    site plant_type percent_plant_cover penguin_density_m2
##   <dbl> <chr>                    <dbl>              <dbl>
## 1   195 liverwort                 22.4               2.38
## 2   196 liverwort                 24.6               2.56
## 3   197 liverwort                 19.1               2.68
## 4   198 liverwort                 31.6               2.12
## 5   199 liverwort                 20.6               1.98
## 6   200 liverwort                 11.4               1.99
```

The data set has 4 columns: the number of the site, which type of non-vascular plant is found at the site, how much of the ground is covered by plants, and the density of penguins.

Regression

Our first goal is to determine if there is a relationship between the amount of plant cover and penguin density.

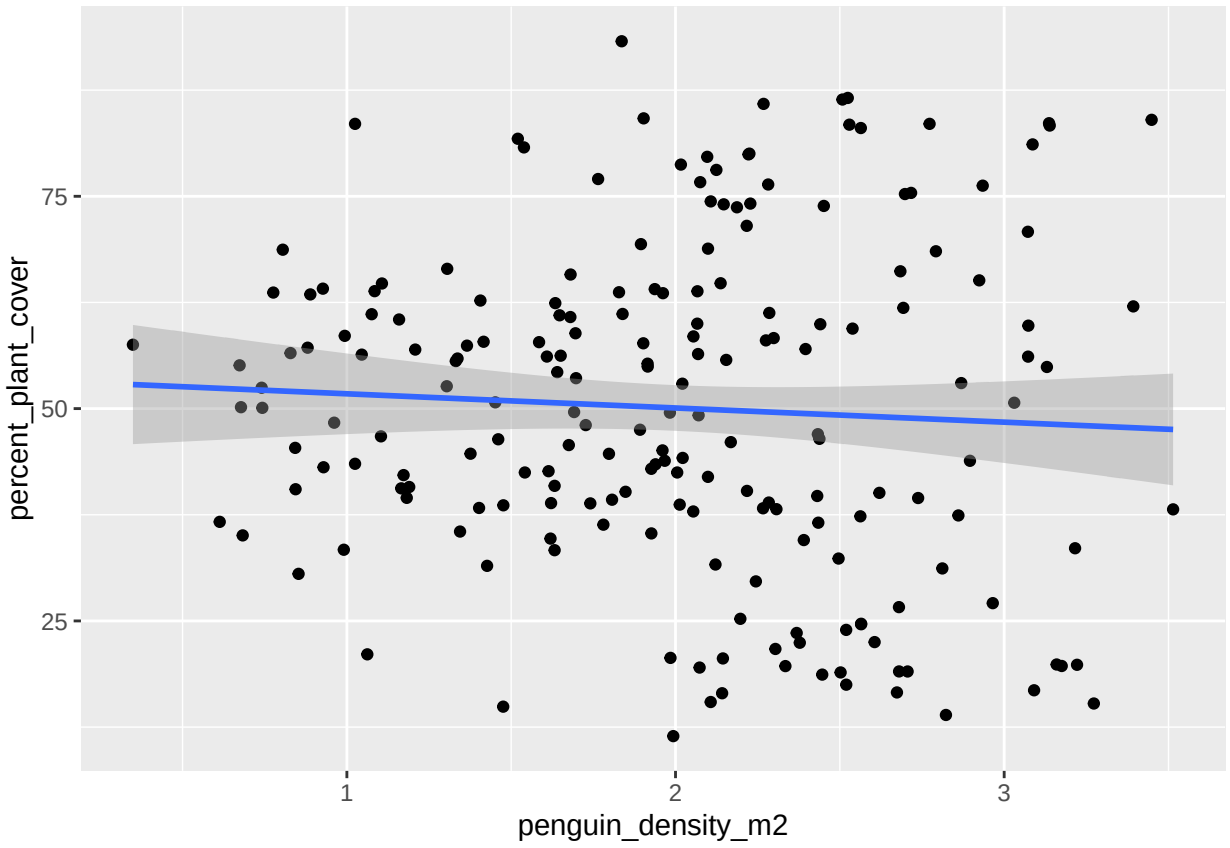
4. Which of our two variables is dependent and which is independent (hint: re-read the introduction if you're feeling confused). Determine whether each is continuous or categorical. (2 points)

- `penguin_density_m2`: independent, continuous
- `percent_plant_cover`: dependent, continuous

5. Plot the data (disregard plant type for now) using the appropriate plot from `ggplot2`. Remember to add a line of best fit (and remember to make that line a straight line using the `method = "lm"` argument!). (2 points)

```
ggplot(plants, aes(penguin_density_m2, percent_plant_cover)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



6. Calculate the correlation coefficient. What does this tell us about the relationship? (2 points)

Answer: perhaps a slight negative relationship but probably none at all

```r
r <- cor(plants$penguin_density_m2, plants$percent_plant_cover)
r
```

```
## [1] -0.05922145
```

7. Calculate R^2. How much variation does our line of best fit explain (report in %)? (2 points)

   Answer: 0.3% aka not very much

```r
r^2
```

```
## [1] 0.00350718
```

8. Run a linear regression model for these days and interpret the results. Does it seem like penguin density is a significant driver of plant cover? (3 points)

   Answer: $p > 0.05$ so not a likely driver of plant cover

```r
summary(lm(data = plants, percent_plant_cover ~ penguin_density_m2))
```

```
##
## Call:
## lm(formula = percent_plant_cover ~ penguin_density_m2, data = plants)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.647 -12.154  -0.661  12.233  42.906
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          53.413      4.212  12.682   <2e-16 ***
## penguin_density_m2   -1.670      2.001  -0.835    0.405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.29 on 198 degrees of freedom
## Multiple R-squared:  0.003507,   Adjusted R-squared:  -0.001526
## F-statistic: 0.6969 on 1 and 198 DF,  p-value: 0.4048
```

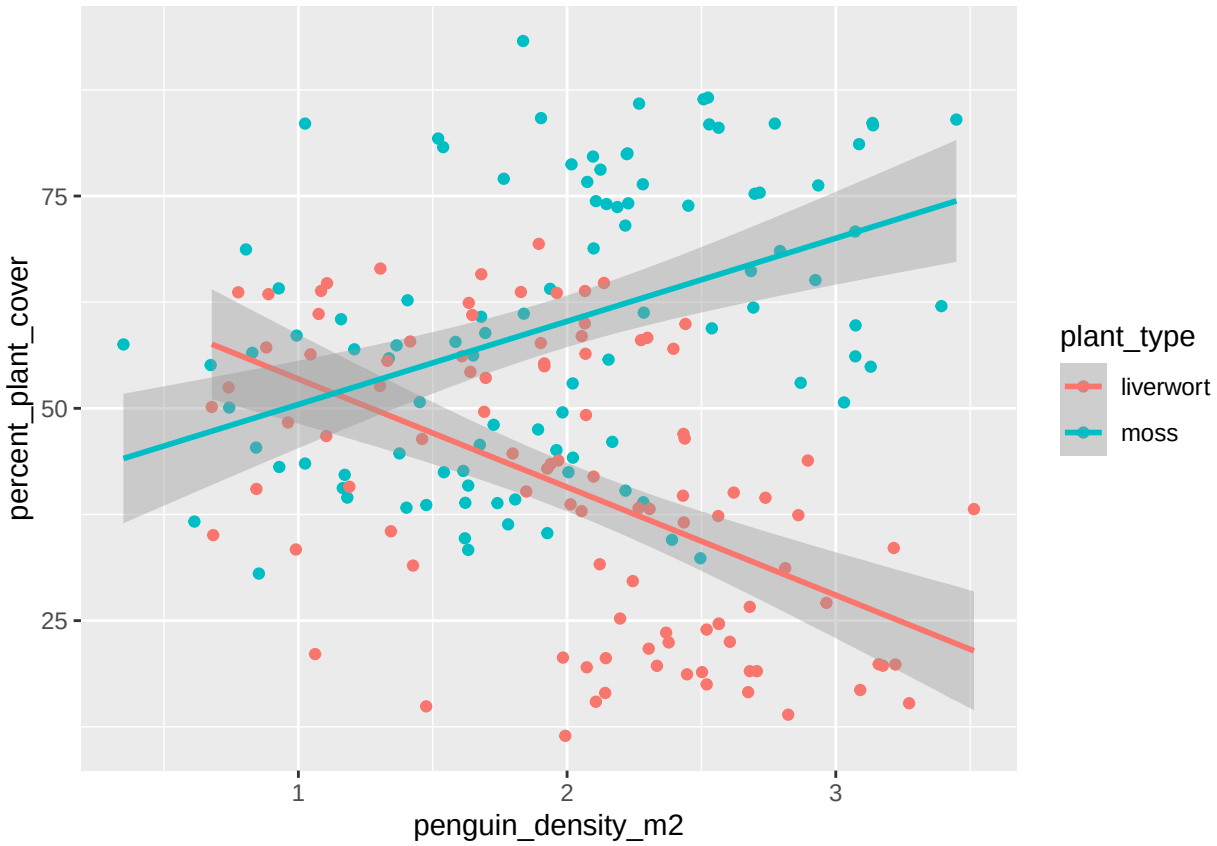
Multiple Regression and Interaction

Maybe we can get more information if we include the plant type in the model.

9. First, let's plot the data again, but this time make the color different for each type of plant. (2 points)

```r
ggplot(plants, aes(penguin_density_m2, percent_plant_cover, color = plant_type)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Woah, that's quite different! Our interpretation of whether penguin density affects plant cover might need to change...

10. Run a multiple regression model, incorporating the plant type into the model using the * notation. Write 2-3 sentences interpreting the results. (3 points)

- Which variables are significant? How do we know?
- Is the interaction term significant? How do we know?

  Answer: all variables are highly significant including the interaction term–all $p < 0.05$

```
summary(lm(data = plants, percent_plant_cover ~ penguin_density_m2 * plant_type))
```

```
##
## Call:
## lm(formula = percent_plant_cover ~ penguin_density_m2 * plant_type,
##     data = plants)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.757 -12.508   1.299  11.637  34.609
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       66.183      4.776  13.858  < 2e-16 ***
## penguin_density_m2               -12.730      2.238  -5.688 4.62e-08 ***
```

```
## plant_typemoss                      -25.508      6.508  -3.920 0.000123 ***
## penguin_density_m2:plant_typemoss   22.516      3.089   7.289 7.51e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.86 on 196 degrees of freedom
## Multiple R-squared:  0.4151, Adjusted R-squared:  0.4061
## F-statistic: 46.37 on 3 and 196 DF,  p-value: < 2.2e-16
```

Why might different plants respond differently to penguin densities?

Perhaps one group is more sensitive to trampling or perhaps penguins like to snack on them. Any number of causes could be at play. What is important to recognize is that there is an interactive effect here: different plants respond to differently to penguin densities!