Regression Code Along

Mila Pruiett

Statistical inference and regression analyses

Setting up the scenario

We want to build a road to our fishing site, while minimizing our impact on the delicate antarctic ecosystem. For today's lesson, we are going to focus on antarctic hairgrass, one of only two flowering species of plants on the continent.

https://www.researchgate.net/figure/Morphology-of-Antarctic-hair-grass-Deschampsia-antarctica-a-A-small-cluster-of-D_fig1_304660866

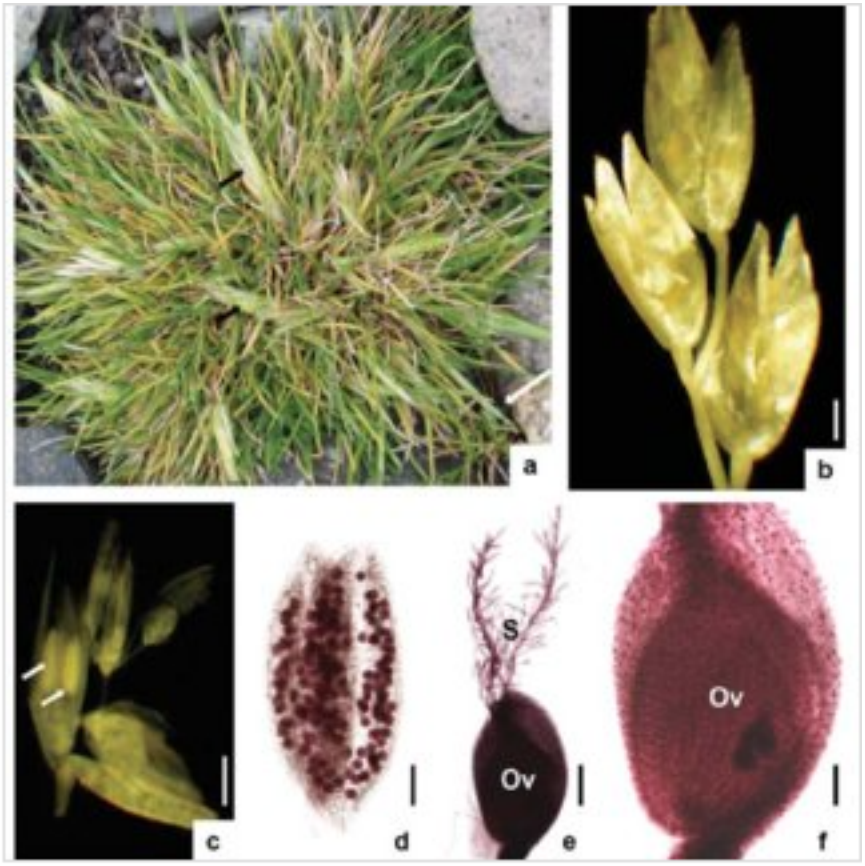


Figure 1: Hairgrass plant and flowering structures

We want to know what environmental conditions are associated with hairgrass, so we can build a road where those conditions are not. It would take far too long to survey every square inch of land between our base

and our fishing spot, so we are going to build a model based on some samples of where hairgrass is found to predict where else it might be.

We collected data for one month on key components of the hairgrass' environment

- soil pH : most plants prefer mildly acidic to neutral environments
- nitrogen content (as percentage per 100 mL soil sample) : important for plant growth and tissue building
- phosphorous content (as percentage per 100 mL soil sample) : important for plant growth and tissue building
- percent soil rock : rockiness of soil impacts water drainage and temperature
- max windpseed knots : extreme wind can pose a challenge to plants of all types
- average UV index : plants can get sunburned too
- average summer temperature
- average winter temperature
- penguin density within 100 m : the number of penguins per 5 m sq within 100 m of the sample quadrant for hairgrass
- hairgrass density (measured as number of individual clumps of hairgrass within 1 square meter)

(This data is based on this article: I.Yu. Parnikoza, N.Yu. Miryuta, D.N. Maidanyuk, S.A. Loparev, S.G. Korsun, I.G. Budzanivska, T.P. Shevchenko, V.P. Polischuk, V.A. Kunakh, I.A. Kozeretska, Habitat and leaf cytogenetic characteristics of Deschampsia antarctica Desv. in the Maritime Antarctica, Polar Science, Volume 1, Issues 2–4, 2007, Pages 121-128, ISSN 1873-9652, https://doi.org/10.1016/j.polar.2007.10.002.)

Exploring and analyzing our first variables of interest

There are many environmental conditions that may be associated with hairgrass density. For today's code along, we are going to focus on two: soil pH and nitrogen content.

Let's look at nitrogen content first.

We always should start with a data visualization and some descriptive statistics.

```
# load in the tidyverse
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# load in the data
hairgrass <- read_csv("hairgrass_data.csv")
```

```
## Rows: 480 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (11): location_ID, soil_pH, p_content, percent_soil_rock, max_windspeed...
## date  (1): date
##
## 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.
```

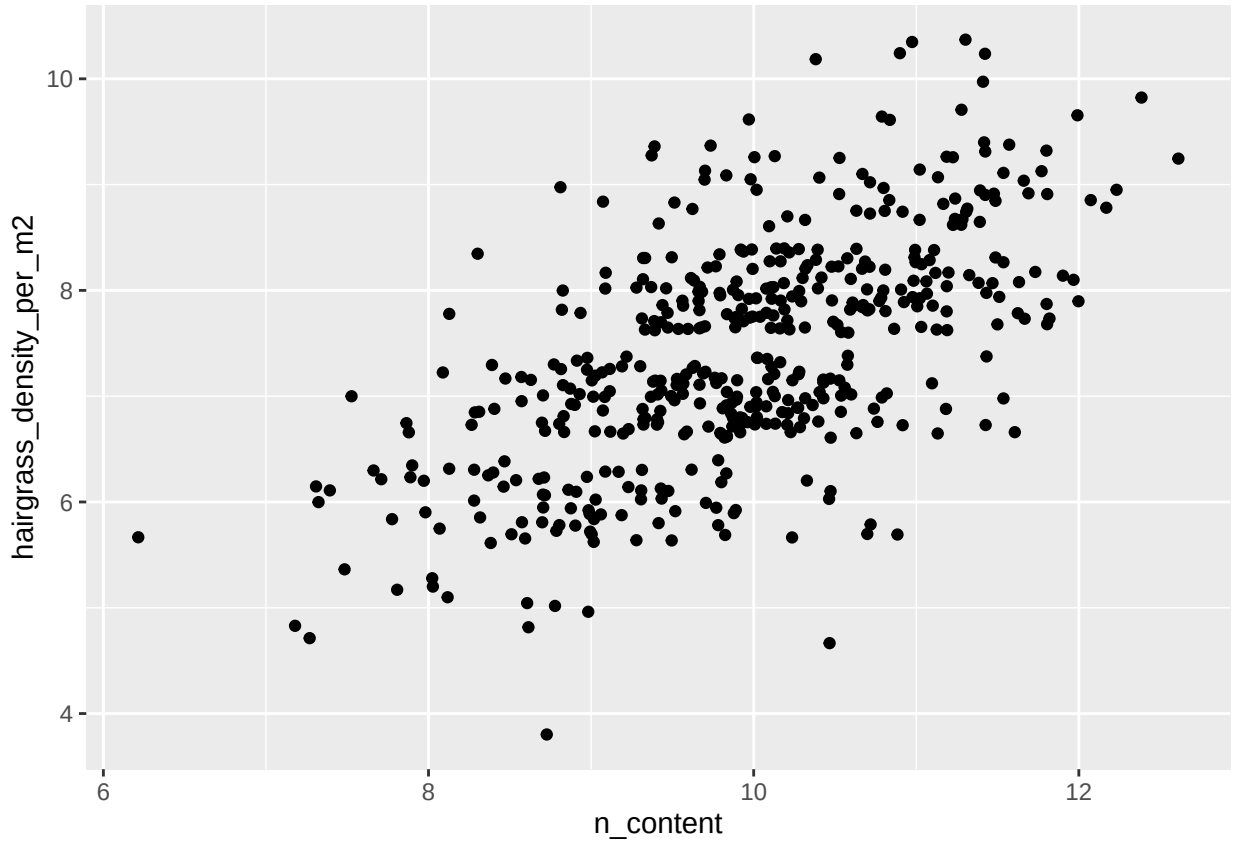
```
# What are our independent and dependent variables?
# What kind of variables are they?
# What kind of viz should we do?

hairgrass %>% summarize(max(n_content), min(n_content), mean(n_content), sd(n_content))
```

```
## # A tibble: 1 x 4
##   `max(n_content)` `min(n_content)` `mean(n_content)` `sd(n_content)`
##              <dbl>            <dbl>             <dbl>           <dbl>
## 1             12.6              6.2              9.93            1.02
```

```
# okay cool good to know, emphasize we are not doing group by because not categorical data

hairgrass %>% ggplot(aes(y = hairgrass_density_per_m2, x = n_content)) +
  geom_jitter()
```



```
  # could do jitter or not, what do other people think?

# do you see a pattern? Do you think these data are correlated? What do you think the correlation coeff
```

Now let's actually calculate the correlation coefficient, r. As a reminder, the correlation coefficient is a number between -1 and 1 that kooks at the relationship between two numeric variables. The greater the magnitude of the correlation coefficient, the stronger the correlation (All the points fall exactly on the line of best fit if r = 1 or -1).

```
r = cor(hairgrass$hairgrass_density_per_m2, hairgrass$n_content)
# What do we expect based on this correlation coefficient?
```

We often think about the correlation in terms of r-squared. All we have to do is square the value we calculated above. How do we interpret r-squared for this relationship?

```
r^2
```

```
## [1] 0.400296
```

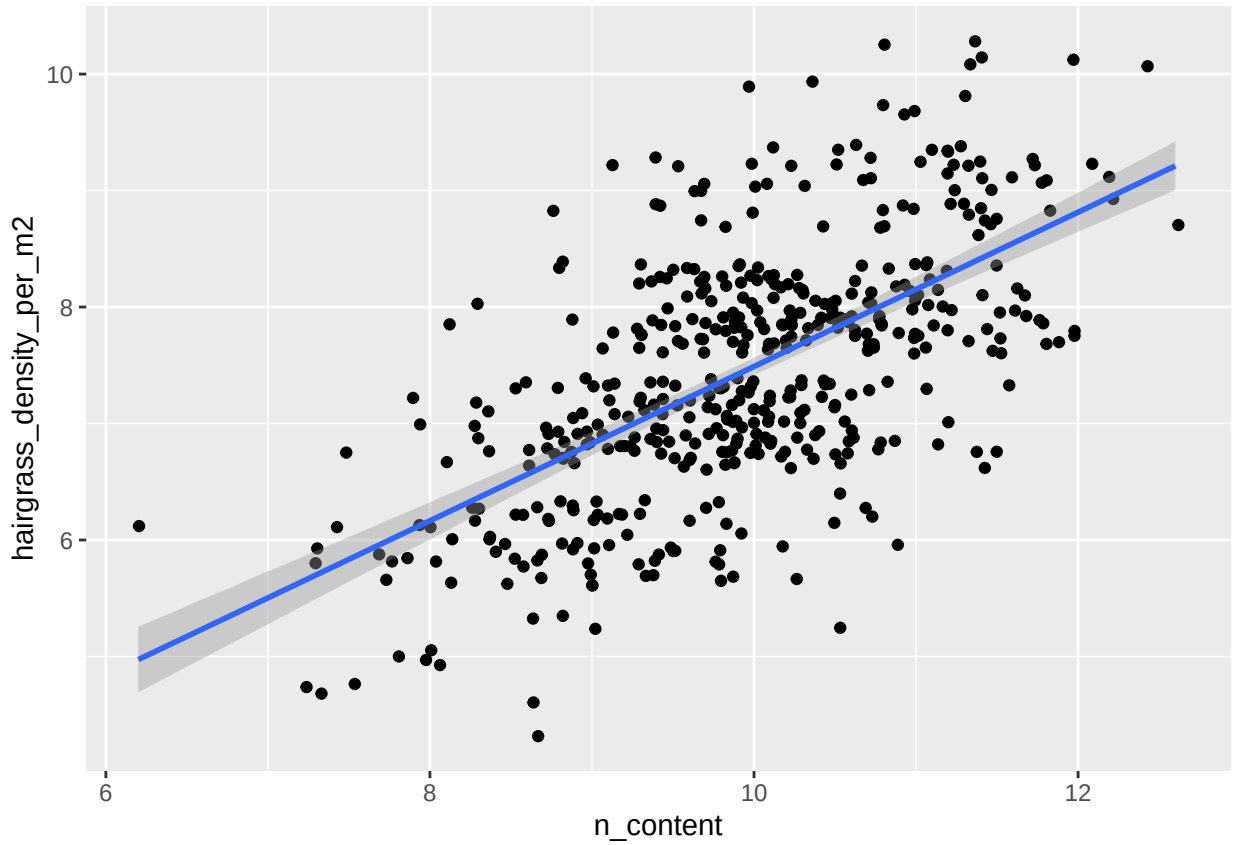
```
# Means that 40% of the variation in harigrass density can be explained by the variation in nitrogen co
```

Adding our line of best fit to the data

```
# is that what we expected based on that correlation coefficient?

hairgrass %>% ggplot(aes(y = hairgrass_density_per_m2, x = n_content)) +
  geom_jitter() +
  geom_smooth(method = "lm")
```

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



If we want to add statistical rigor, we need to use regression analysis. A regression analysis approximates the relationship between a dependent variable and one or more independent variables and evaluates the strength of that relationship (giving us a p-value).

We will use linear regressions in this unit. This simply means that the model will take the form of y = ax + b, where y is the dependent variable, x is the dependent variable, a is the slope, and b is the y-intercept.

What would the model for our question about nitrogen content be? (it's okay that we haven't yet calculated the values)

```
# hairgrass density = a *  n_content + b
```

What is the null hypothesis? What is the alternative hypothesis?

```
# null: There is no relationship between hairgrass density and n_content
# alt: There is a relationship between hairgrass density and n_content
```

R can actually calculate what this model would be for us. The formula for the line of best fit (y = mx+b) aims to minimize the distance between each observation (point) and the line. What is the model?

```
summary(lm(hairgrass_density_per_m2 ~ n_content, data = hairgrass))
```

```
##
## Call:
## lm(formula = hairgrass_density_per_m2 ~ n_content, data = hairgrass)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.82079 -0.55590 -0.02612  0.57654  2.51032
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.86739    0.37000   2.344   0.0195 *
## n_content    0.66223    0.03707  17.862   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8294 on 478 degrees of freedom
## Multiple R-squared:  0.4003, Adjusted R-squared:  0.399
## F-statistic: 319.1 on 1 and 478 DF,  p-value: < 2.2e-16
```

```
# model: hairgrass density = 0.87 + 0.66 *n_content
```

So what can we conclude about soil pH and hairgrass density?

```
# stats interpretation
# Because the p-value associated with the F statistic was 319, we reject the null hypothesis that there

# interpretation in light of scenario: we should pay attention to n content as we build our road

# REMIND everyone how to submit these words so they get counted
```

Moving on to soil pH

Data visualization, with the line of best fit, and summary statistics for soil pH

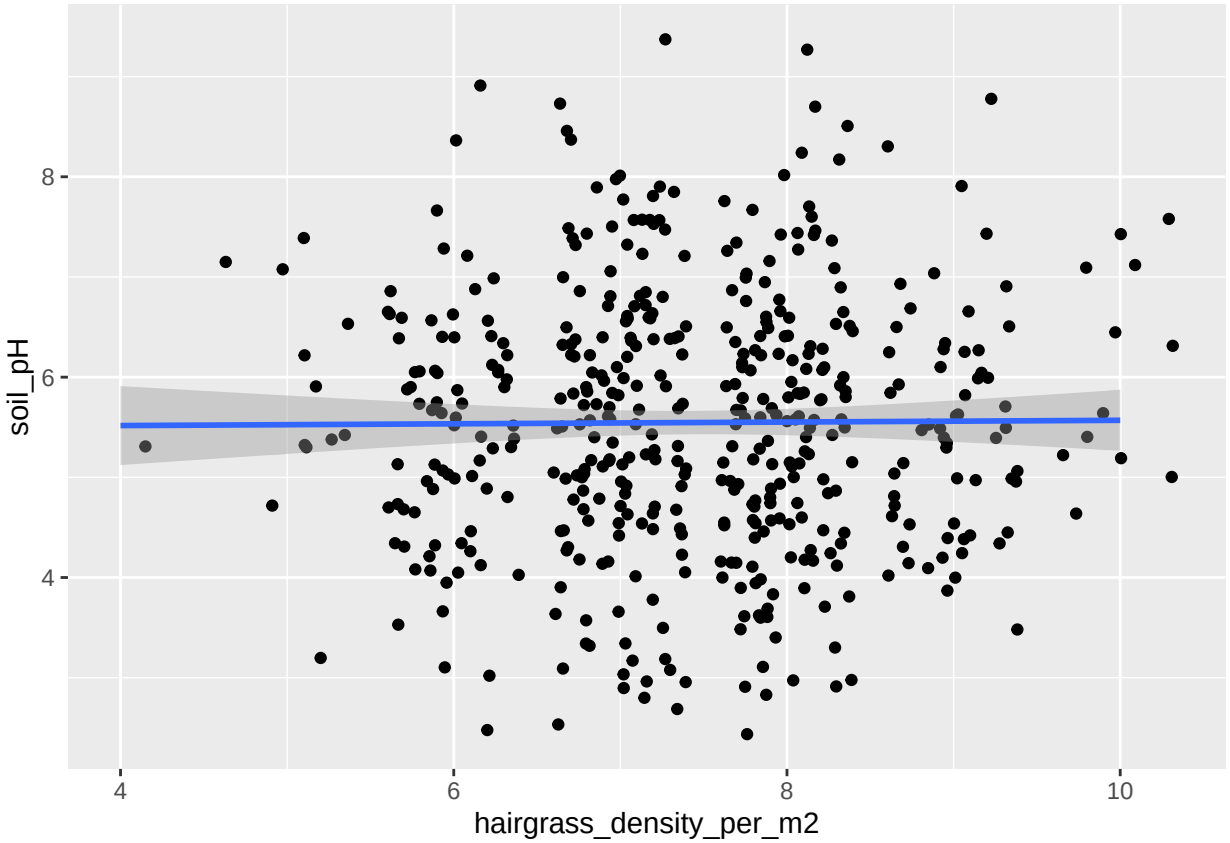
```
hairgrass %>% summarize(max(soil_pH), min(soil_pH), mean(soil_pH), sd(soil_pH))
```

```
## # A tibble: 1 x 4
##   `max(soil_pH)` `min(soil_pH)` `mean(soil_pH)` `sd(soil_pH)`
##            <dbl>          <dbl>           <dbl>         <dbl>
## 1           9.37           2.44            5.55          1.30
```

```
# okay cool good to know, emphasize we are not doing group by because not categorical data

hairgrass %>% ggplot(aes(x = hairgrass_density_per_m2, y = soil_pH)) +
  geom_jitter() +
  geom_smooth(method = "lm")
```

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



What is the correlation coefficient?

```
cor(hairgrass$hairgrass_density_per_m2, hairgrass$soil_pH)
```

```
## [1] 0.007200444
```

What is the model for our question about soil pH, without values?

```
# hairgrass density = a *  soil_pH + b
```

Create the model in R and calculate the values for a and b.

```
summary(lm(hairgrass_density_per_m2 ~ soil_pH, data = hairgrass))
```

```
##
## Call:
## lm(formula = hairgrass_density_per_m2 ~ soil_pH, data = hairgrass)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4403 -0.4491 -0.4271  0.5631  2.5637
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 7.408859   0.214051  34.613   <2e-16 ***
## soil_pH     0.005915   0.037570   0.157    0.875
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.071 on 478 degrees of freedom
## Multiple R-squared:  5.185e-05,  Adjusted R-squared:  -0.00204
## F-statistic: 0.02478 on 1 and 478 DF,  p-value: 0.875
```

```
# model: hairgrass density = 7.4 + 0.006 * soil pH
```

At alpha = 0.05, what do we conclude about the relationship between soil pH and hairgrass density and why?

```
# stats interpretation
# Because the p-value associated with the F statistic was 0.875, we accept the null hypothesis that the


# REMIND everyone how to submit these words so they get counted
```

What does this mean for the road we are building?

```
# interpretation in light of scenario: we shouldn't worry about soil pH as we think about where to buil
```