4.1: Roads and Regressions

The Scenario

Last module, we tried to figure out which bays we should prioritize for fishing based on the following data sets:

- Radio collar data
- Fish catch data
- Leopard seal abundance data

Based on our results from Assignment 2, it looks like Sulzberger Bay and Hope Bay are our best bets. Now, we need to make these bays accessible over land to transport the fish we've caught to our home base.

We want to build a road while minimizing our impact on the delicate antarctic ecosystem.

Our first order of business is to make sure that we avoid areas with that are well-suited for Antarctica hair grass (*Deschamsia antarctica*), one of only two flowering species of plants on the continent.

We want to know what environmental conditions are associated with hair grass. This way, we can avoid areas where these conditions are met and not destroy precious habitat for the hair grass.

It would take far too long to survey every bit of land between our base and our fishing spots, so we are going to build a **model** based on some samples of where hair grass is found to help us predict where else it might be.

What is a model?

A model is a way for us to take complex systems and break them down into small, more understandable bits. We can use models to help us understand the relationship between different variables. We can then use those model to make predictions based on those relationships.

Data

Knowing that we would soon be building roads, we asked our botanists to collect data for us on key components of the hairgrass' environment. Since it would take too long to sample everywhere hairgrass grows, they collected data from a sample of the hairgrass population.

Our botanists collected data for the following variables:

- ***soil pH*:** most plants prefer mildly acidic to neutral environments
- ***nitrogen (N) content*:** as percentage per 100 mL soil sample; important for plant growth and tissue building
- ***phosphorous (P) content*:** as percentage per 100 mL soil sample; important for plant growth and tissue building
- ***percent rock*:** how rocky the location is; rocks in soil impact water drainage and temperature

- ***max windspeed (knots per hour)*:** extreme wind can pose a challenge to plants of all types
- ***average summer temperature (C)*:** temperature in the growing season
- ***penguin density*:** the number of penguins per 5 m^2 within 100 m of the sample quadrant for hair grass; penguin poop increases nitrogen content in the system
- ***hairgrass density*:** number of individual clumps (tussocks) of hairgrass per 1 m^2

This data is based on this article: Parnikoza, et al. 2007

Summarize and Visualize

As we see above, there are many environmental conditions that may be associated with hair grass density.

For this lesson, we are going to focus on two: *nitrogen (N) content* and *soil pH*.

Set-Up

As usual, we start with loading our packages and our data.

```
# Load library
library(tidyverse)

# Load data
hairgrass <- read_csv("data/hairgrass_data.csv")
```

Let's take a look at the data in the data set. What does each row represent?

```
# View first few rows
head(hairgrass)
```

```
# A tibble: 6 x 10
  location_ID soil_pH p_content percent_soil_rock max_windspeed_knots
        <dbl>   <dbl>     <dbl>             <dbl>               <dbl>
1           1    4.9     5.49               44.5                14.9
2           2    6.94    8.53               50.5                11.0
3           3    4.36    0.0801             88.5                26.5
4           4    5.41    1.98               61                  23.6
5           5    5.32    6.6                67.1                27.4
6           6    6.49    4.09               42.8                22.6
# i 5 more variables: avg_uv_index <dbl>, avg_summer_temp <dbl>,
#   n_content <dbl>, hairgrass_density_m2 <dbl>, penguin_density_5m2 <dbl>
```

```r
# View last few rows
tail(hairgrass)
```

```
# A tibble: 6 x 10
  location_ID soil_pH p_content percent_soil_rock max_windspeed_knots
        <dbl>   <dbl>     <dbl>             <dbl>               <dbl>
1         475    4.57     7.96               96.2               25.5
2         476    6.1      0.448              94.5                5.00
3         477    4.72     4.86               70.3               12.3
4         478    6.41     3.8                86.9                4.86
5         479    6.02     4.62               12                  7.23
6         480    3.64     2.32               94.9                3.94
# i 5 more variables: avg_uv_index <dbl>, avg_summer_temp <dbl>,
#   n_content <dbl>, hairgrass_density_m2 <dbl>, penguin_density_5m2 <dbl>
```

Nitrogen Content

Let's start by investigating any relationship between hair grass density and nitrogen content.

First, we should spend a little time thinking about our variables. Spend a few minutes in small groups discussing the answer to the questions below.

- Which columns are we interested in right now? **N_content, hairgrass_density**
- Which one is the independent variable and which is the dependent variable?
  - independent: **N_content**
  - dependent: **hairgrass_density_m2**
- Are the variables categorical or continuous? **Both are continuous**
- Which type of data visualization should we use? **Scatterplot**

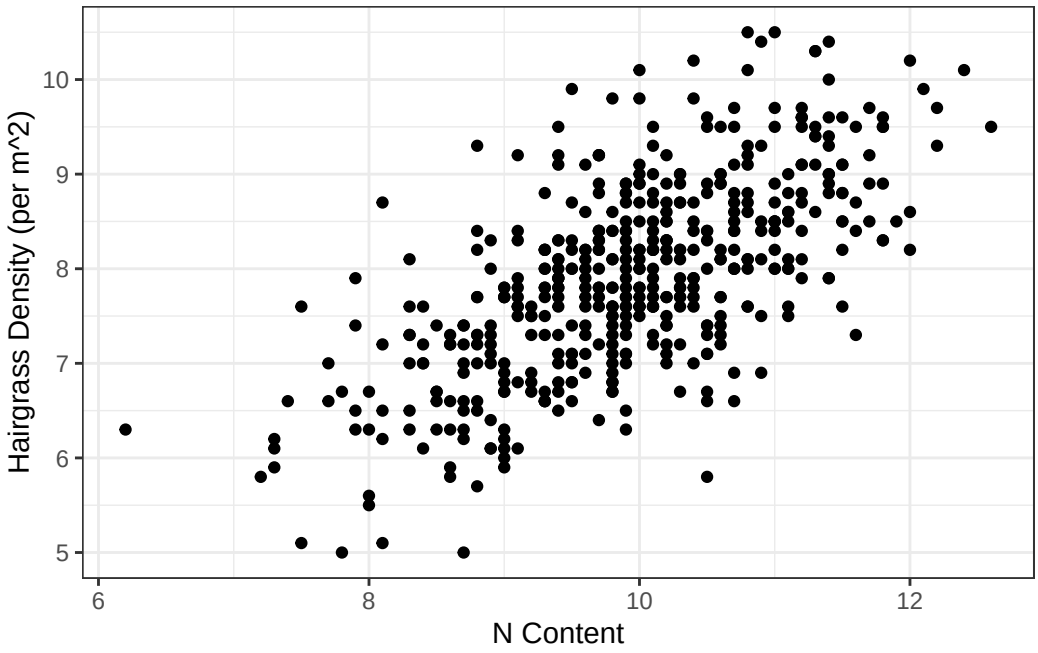
Next, calculate the mean (`mean()`) and standard deviation (`sd()`) for the nitrogen content.

```r
# Mean and standard deviation of nitrogen content
hairgrass %>%
  summarise(mean_N = mean(n_content),
            sd_N = sd(n_content))
```

```
# A tibble: 1 x 2
  mean_N  sd_N
   <dbl> <dbl>
1   9.93  1.02
```

Now that we have summarized the data, let's take a look at the data visually.

```
# Scatterplot
ggplot(hairgrass, aes(n_content, hairgrass_density_m2)) +
  geom_point() +
  labs(x = "N Content",
       y = "Hairgrass Density (per m^2)") +
  theme_bw()
```



Do you see a pattern? What type of relationship do you see?

How do we analyze this type of relationship statistically?

Statistical Analysis

We are going to use two (related) statistical methods to understand the relationship between two continuous (numeric) variables:

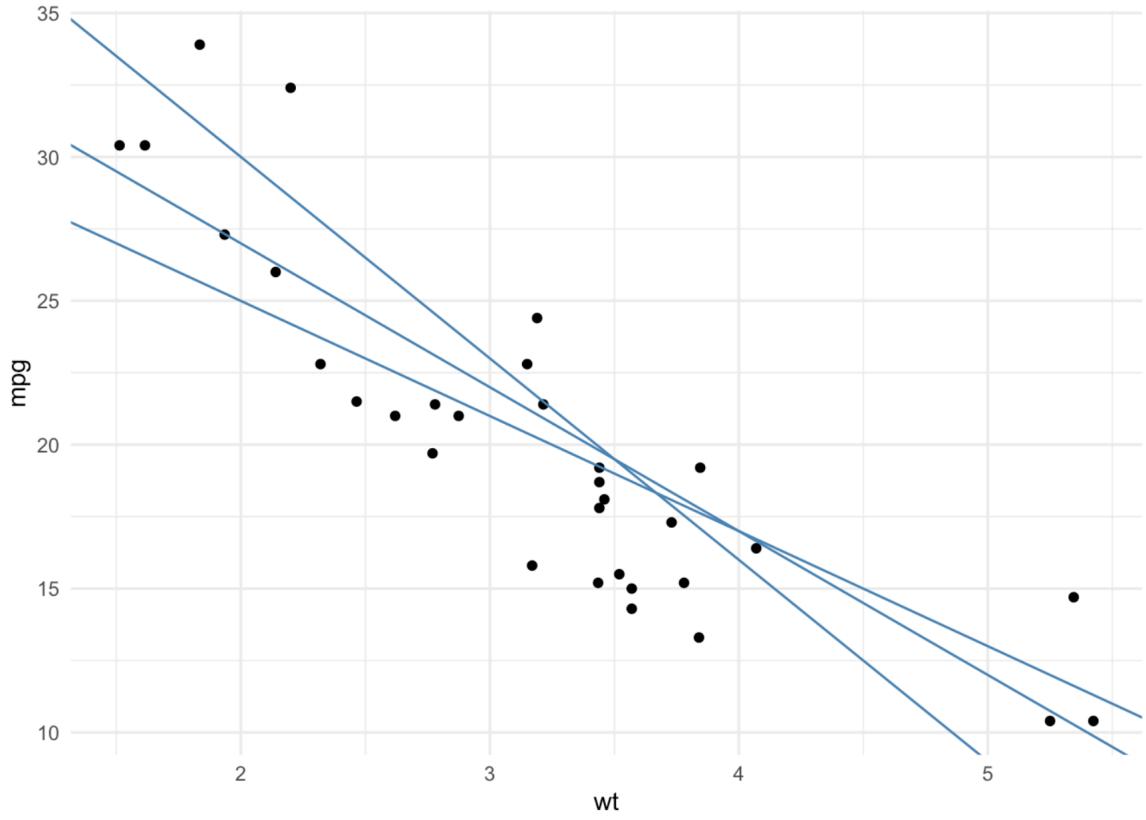
- Correlation coefficient (r) and R-squared (R^2)
- Linear regression

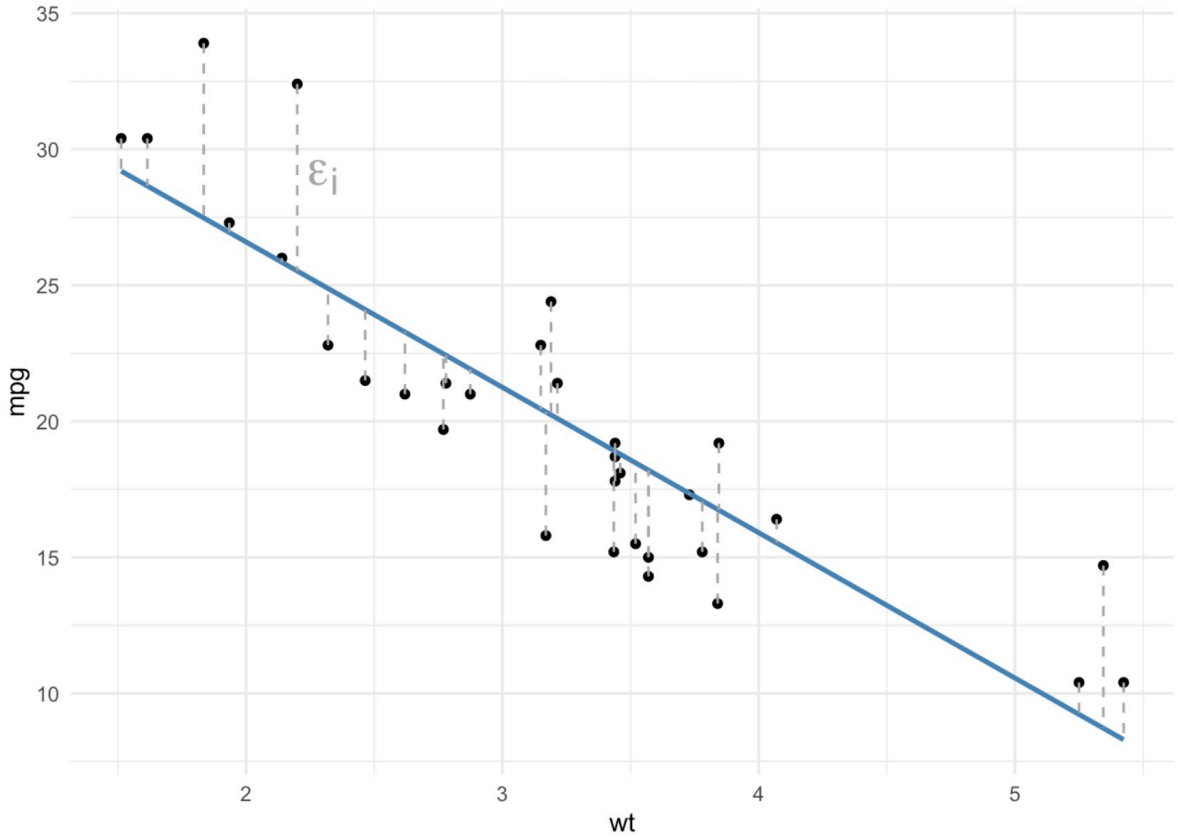
Line of Best Fit

To understand correlation, R-squared, and linear regression, we first need to talk about something we call the *line of best fit*.

The line of best fit aims to minimize the distance between each observation (points) and the line. The distances between each observation and any line are called *residuals* (the dotted gray lines). The line of best fit is the line that has the smallest *residuals*.

In the first step, there are many potential lines. Three of them are plotted:

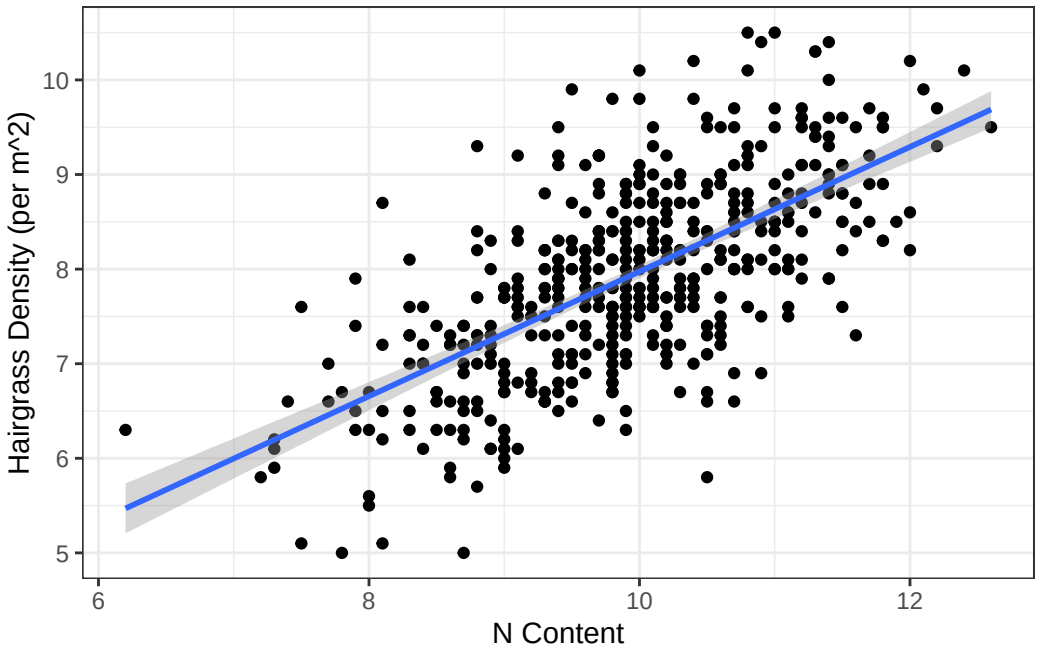




We can add a line of best fit to a ggplot by using the `geom_smooth()` function. We need to specify that the method we want should produce a straight line ("linear model").

```
# Scatterplot
ggplot(hairgrass, aes(n_content, hairgrass_density_m2)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "N Content",
       y = "Hairgrass Density (per m^2)") +
  theme_bw()
```

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



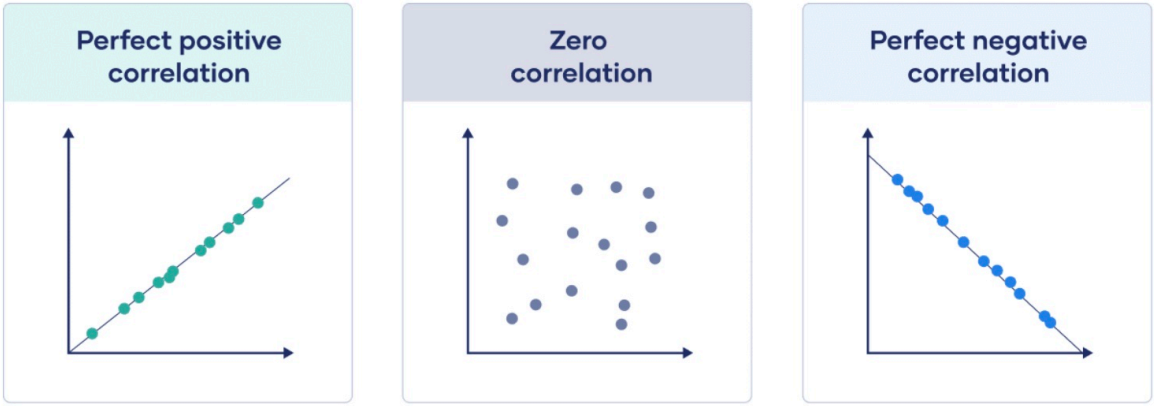
Correlation Coefficient

The correlation coefficient, `r`, is a measurement of the strength of the relationship between two continuous variables.

The correlation coefficient is a number between -1 and 1 that looks at the relationship between two numeric variables. If the value is negative, there is a negative relationship between the two variable; if the value is positive, there is a positive relationship.

If all the points fall exactly on the line of best fit, r = 1 or -1. If there is no relationship between the variables, `r` is 0 (or something very close to it).

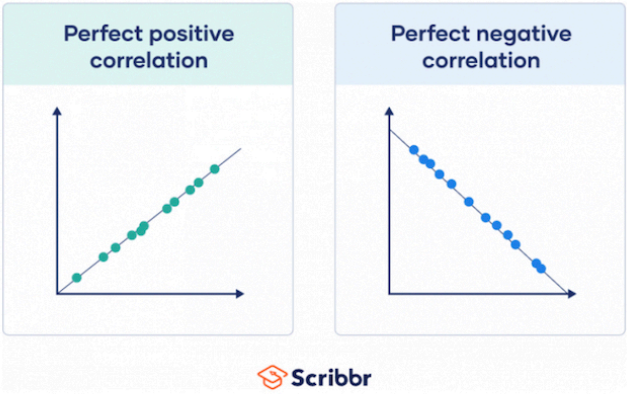
Correlation coefficient value	Correlation type	Meaning
1	Perfect positive correlation	When one variable changes, the other variables change in the same direction.
0	Zero correlation	There is no relationship between the variables.
-1	Perfect negative correlation	When one variable changes, the other variables change in the opposite direction.



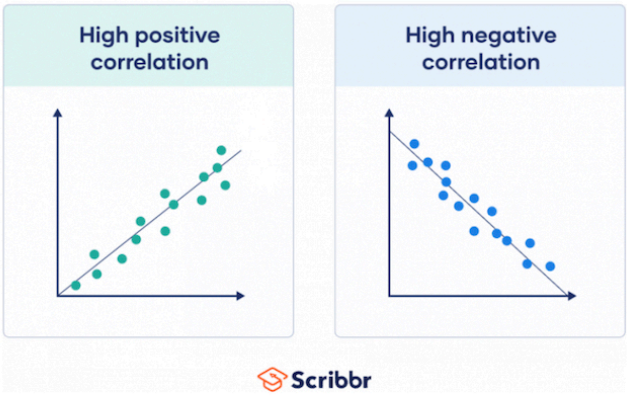
The greater the magnitude (size) of the correlation coefficient, the stronger the relationship between the two variables.

The closer your points are to this line, the higher the absolute value of the correlation coefficient and the stronger your linear correlation.

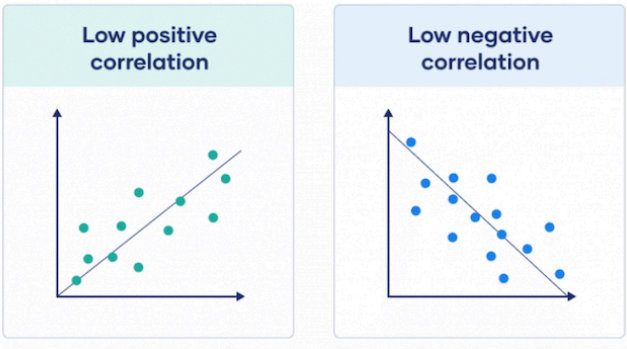
If all points are perfectly on this line, you have a **perfect** correlation.



If all points are close to this line, the absolute value of your correlation coefficient is **high**.

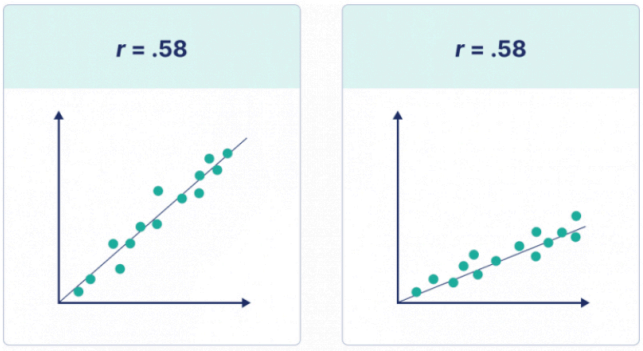


If these points are spread far from this line, the absolute value of your correlation coefficient is **low**.



It is also important to recognize that the correlation coefficient has nothing to do with the slope of the line (we use linear regression to assess that!).

> Note that the steepness or slope of the line isn't related to the correlation coefficient value. The correlation coefficient doesn't help you predict how much one variable will change based on a given change in the other, because two datasets with the same correlation coefficient value can have lines with very different slopes.



Based on the hairgrass plot we did above, do you expect the correlation coefficient to be positive or negative? For this class, we'll say that any `r` value with 0.1 of 0 means there is no relationship.

To calculate the correlation coefficient, we need to go back to base R and indicate which columns we are referencing with the `$` operator. We use the `cor()` function.

```
# Find the correlation coefficient
r <- cor(x = hairgrass$n_content, y = hairgrass$hairgrass_density_m2)

# View result
r
```

```
[1] 0.6556456
```

R-squared (R^2)

The R-squared (R^2) value is a representation of how much variation is explained by the line of best fit.

When we have a dependent variable and an independent variable, the R-squared value tells us how much of the variation in the *dependent* variable is explained by the *independent* variable.

Sometimes we don't have obvious dependent and independent variables, but we can still use similar language (R^2 tells us how much of the variation in the data is explained).

To calculate R-squared, we square the correlation coefficient value we calculated above. It is always positive because we are squaring it `r` value; that means that R-squared values range from 0 to 1. The closer to 1, the more variation is explained.

```
# What is r^2?
r^2
```

```
[1] 0.4298711
```

How would we interpret this `r^2` value?

Linear Regression Analysis

A regression analysis approximates the relationship between a dependent variable and one or more independent variables. It evaluates the strength of that relationship, ultimately giving us a p-value.

Since we are using linear regressions in this course, the regression model will take the form of a line: `y = mx + b`

- `y` = dependent variable (y-axis)
- `x` = independent variable (x-axis)
- `m` = slope of the line is the slope
- `b` = y-intercept

Using our variables, what would our linear regression model look like (we don't know `m` or `b` yet…)?

Hypothesis Testing

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

Null Hypothesis (H_0): There is *no* relationship between hair grass density and nitrogen content.

Alternative Hypothesis (H_A): There *is* a relationship between hair grass density and nitrogen content.

What do you think this means for the slopes?

Regression Analysis

Thankfully, R can calculate the slope (`m`) and y-intercept (`b`) of the line of best fit for us.

Let's find the equation for our line of best fit, our test statistic, and our p-value. To do this, we use a function called `lm()`: this stands for "linear model."

Like with ANOVA, we will then want to use the `summary()` function to get out the values we need.

```
# Linear model of hairgrass density v. nitrogen content
nitrogen <- lm(hairgrass_density_m2 ~ n_content, hairgrass)

# Summary of results
summary(nitrogen)
```

```
Call:
lm(formula = hairgrass_density_m2 ~ n_content, data = hairgrass)

Residuals:
     Min       1Q   Median       3Q      Max
-2.50324 -0.53853  0.01917  0.49500  2.25562

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.38516    0.34636   3.999 7.36e-05 ***
n_content    0.65886    0.03471  18.984  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7765 on 478 degrees of freedom
Multiple R-squared:  0.4299,    Adjusted R-squared:  0.4287
F-statistic: 360.4 on 1 and 478 DF,  p-value: < 2.2e-16
```

As with our other statistical tests — t-tests and ANOVAs — the results give us some important values:

- `b` (y-intercept): part of our line of best fit equation
  - The "intercept estimate", in this case 1.385, is our y-intercept in our line of best fit
- `m` (slope): also part of our line of best fit equation

  - This is the estimate for our independent variable (N_content in this case)
  - In this model, m = 0.659

- *F-statistic*: this is our test statistic
- *p-value*: calculated from our regression model, used to determine significance (0.05 cut-off, as usual)
  - There are multiple p-values here. Focus either on the p-value for the independent variable (N_content) or the overall p-value displayed in the last line of the results summary (p < 2.2e-16)
- *R-squared*: this is our R-squared value that we calculated earlier
  - You can report either the "multiple" or the "adjusted"
  - The "multiple" will typically match the one we calculate with code

This means our equation for the line of best fit is: y = 0.659x + 1.385 and there is a statistically significant relationship between nitrogen content and hairgrass density.

Soil pH

Let's do the same series of steps to determine how soil pH impacts hair grass densities. Work on this in your small groups, and we will go over it in about 10 minutes.

Start with summarizing the data: mean and standard deviation.

```
# Mean and standard deviation of soil pH
hairgrass %>%
  summarize(mean_pH = mean(soil_pH),
            sd_pH = sd(soil_pH))
```

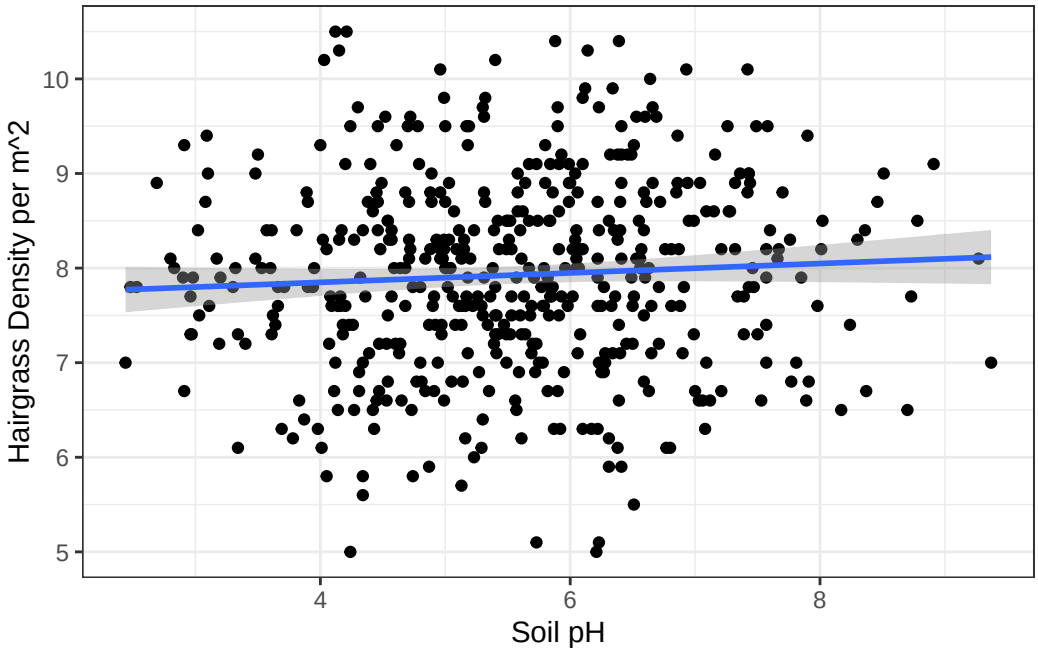
```
# A tibble: 1 x 2
  mean_pH sd_pH
    <dbl> <dbl>
1    5.54  1.30
```

Visualize the data. Remember to add in the line of best fit using `ggplot2`. Also add labels and a theme for practice.

```
# Scatterplot
ggplot(hairgrass, aes(soil_pH, hairgrass_density_m2)) +
  geom_point() +
  geom_smooth(method = "lm") +
```

```
    labs(x = "Soil pH",
         y = "Hairgrass Density per m^2") +
    theme_bw()
```

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



Calculate the correlation coefficient (`r`)? What does this tell us?

```
# Correlation coefficient
r <- cor(x = hairgrass$soil_pH, y = hairgrass$hairgrass_density_m2)
```

How much variation does soil pH explain in the hair grass density data?

```
# What is r^2?
r^2
```

```
[1] 0.003962606
```

Write out the model for our question about soil pH (without values)?

```
# hairgrass density  = m(soil_pH) + b
```

Run the regression model and write out the equation for the line of best fit.

```
# Linear model of hairgrass density v. soil pH
soil_pH <- lm(hairgrass_density_m2 ~ soil_pH, data = hairgrass)
summary(soil_pH)
```

```
Call:
lm(formula = hairgrass_density_m2 ~ soil_pH, data = hairgrass)

Residuals:
     Min       1Q   Median       3Q      Max
-2.95914 -0.66321  0.02364  0.65938  2.64477

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.65039    0.20531  37.263   <2e-16 ***
soil_pH      0.04972    0.03605   1.379    0.169
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.026 on 478 degrees of freedom
Multiple R-squared:  0.003963,  Adjusted R-squared:  0.001879
F-statistic: 1.902 on 1 and 478 DF,  p-value: 0.1685
```

Equation for line of best fit: y = 0.005x + 7.65

Interpret the results of the regression (our p-value cutoff is still = 0.05). What do we conclude about the relationship between soil pH and hair grass density? Why?

Because p > 0.05 for the regression model, we fail to reject the null hypothesis and conclude that there is no relationship between soil pH and hairgrass density.

Data-driven Decision Making

The reason we are using regression analysis is to inform where we should (or should not) build our road so we don't harm the sensitive hair grass or take away their prime habitat.

What do results above for nitrogen content and soil pH mean for the road we are building?