Module 4 Assignment 1

Ellen Bledsoe

2023-05-04

In this assignment, we are going to continue using the hair grass data set from class. The first lesson (Roads and Regressions) will be particularly helpful to you in completing this assignment.

We are going to look at the relationship between hair grass density and two other variables: phosphorus content and the average summer temperature.

Set-Up

As always, we must get organized before we can do anything!

First load the tidyverse and read in the hair grass data set.

```r
library("tidyverse")
hairgrass <- read_csv("../data/hairgrass_data.csv")
```

Phosphorus Content

1. Calculate the mean and standard deviation of the measured phosphorus content.

```r
hairgrass %>%
  summarize(mean_P = mean(p_content),
            stdev_P = sd(p_content))
```

```
## # A tibble: 1 x 2
##   mean_P stdev_P
##    <dbl>   <dbl>
## 1   4.76    2.75
```

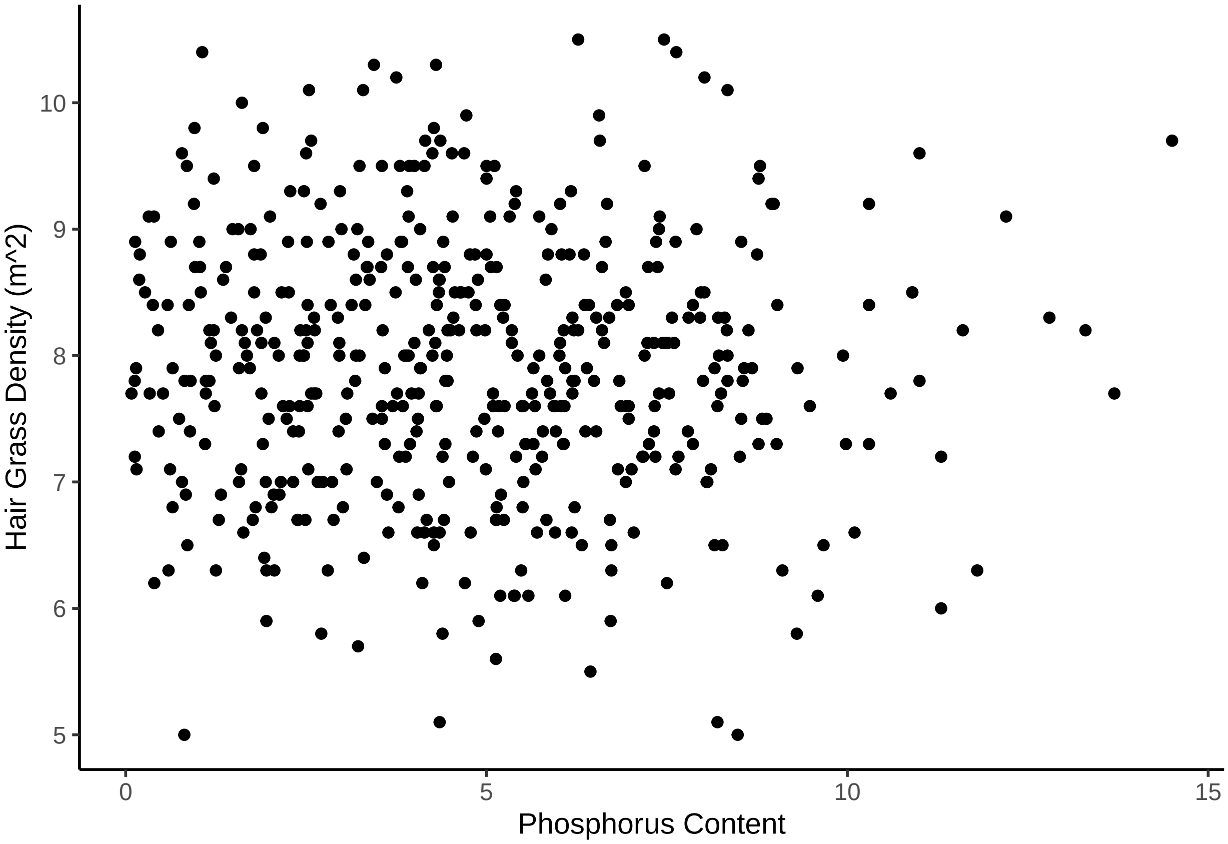
2. Which variable is the independent variable? Which is the dependent?

   Independent:

   Dependent:

3. Create a scatter plot of hair grass density and phosphorus content. Be sure to make the labels easier to understand and add a theme.

```r
ggplot(hairgrass, aes(x = p_content, y = hairgrass_density_m2)) +
  geom_point() +
  labs(x = "Phosphorus Content",
       y = "Hair Grass Density (m^2)") +
  theme_classic()
```



4. Write 1-2 sentences interpreting the plot above. Is this a positive relationship, negative relationship or no relationship at all? Based on your prediction, do you think the correlation coefficient will be positive, negative, or zero?

   Answer:

5. Calculate the correlation coefficient, `r`.

```
r <- cor(y = hairgrass$hairgrass_density_m2, x = hairgrass$p_content)
r
```

```
## [1] -0.02708762
```

6. Calculate the `r^2` value. Write a one sentence interpretation of what the `r^2` value means in the context of these two variables.

```
r^2
```

```
## [1] 0.0007337394
```

Interpretation: This means that less than 1% of the variation in hair grass density can be explained by variation in phosphorus content.

7. What are the null and alternative hypothesis regarding the relationship between these two variables? (2 pts)

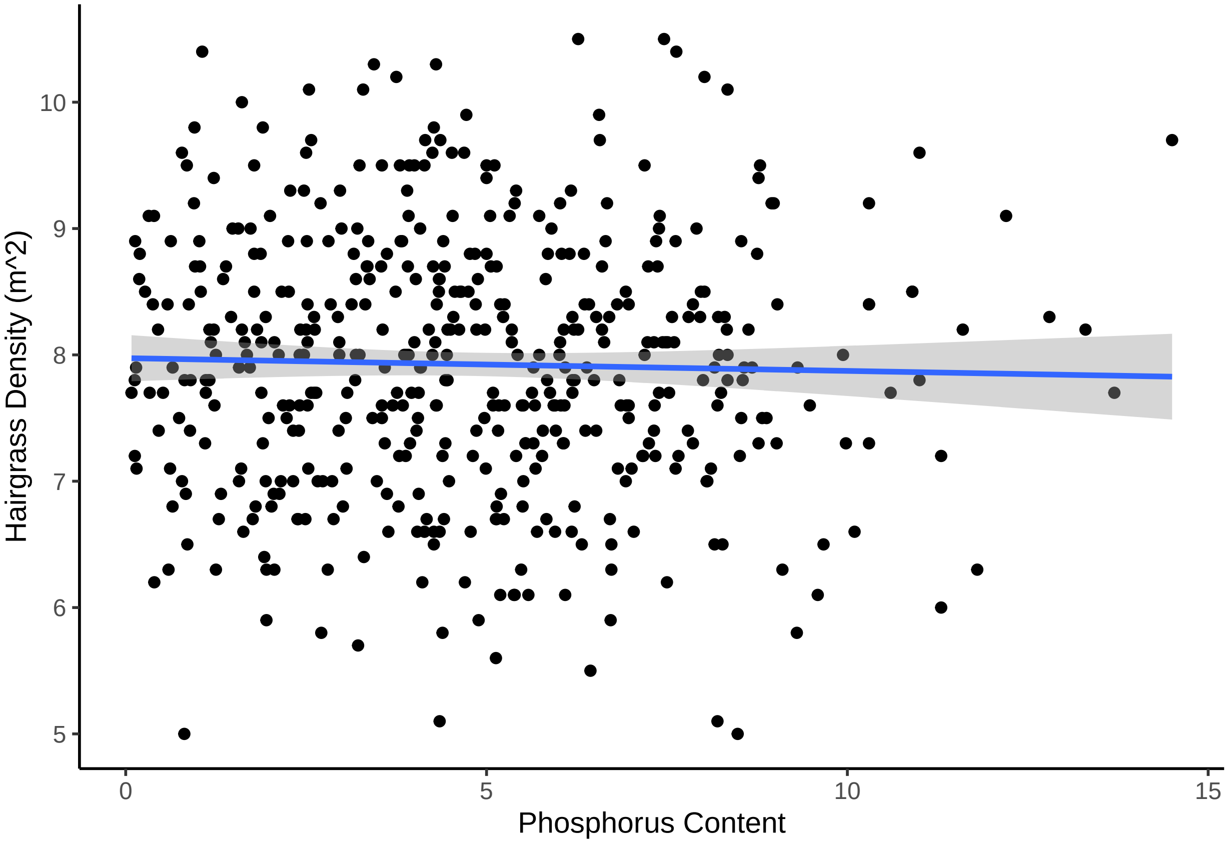
   Null: There is no relationship between hairgrass density and phosphorus content

   Alternative: There is a relationship between hairgrass density and phosphorus content

8. Create the scatter plot that includes the line of best fit (have `ggplot2` calculate the linear equation for you). Again, make the labels clearer and add a theme

```
hairgrass %>%
  ggplot(aes(x = p_content, y = hairgrass_density_m2)) +
  geom_point() +
  geom_smooth(method = "lm") +
  ylab("Hairgrass Density (m^2)") +
  xlab("Phosphorus Content") +
  theme_classic()
```

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



9. Using code, create the regression model in R and obtain the summary of it.

```
mod <- lm(hairgrass$hairgrass_density_m2 ~ hairgrass$p_content)
summary(mod) # 1pt
```

```
##
## Call:
## lm(formula = hairgrass$hairgrass_density_m2 ~ hairgrass$p_content)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.96603 -0.66070  0.01303  0.66970  2.60124
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.97425    0.09393  84.898   <2e-16 ***
## hairgrass$p_content -0.01012    0.01708  -0.592    0.554
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.028 on 478 degrees of freedom
## Multiple R-squared:  0.0007337,  Adjusted R-squared:  -0.001357
## F-statistic: 0.351 on 1 and 478 DF,  p-value: 0.5538
```

10. Write out the equation for the line of best fit using the values from the results above.

    Answer: y = 0.030x + 7.31

11. Interpret the model summary. What is the p-value for our variable of interest? Do we accept or reject the null hypothesis regarding the relationship between these two variables? What can we conclude then about building a road? (2 pts)

    Answer:

the p-value is 0.0943, so we do not reject the null hypothesis There is a not a relationship between phosphorus content and hairgrass density, so we don't need to take it into account for our road

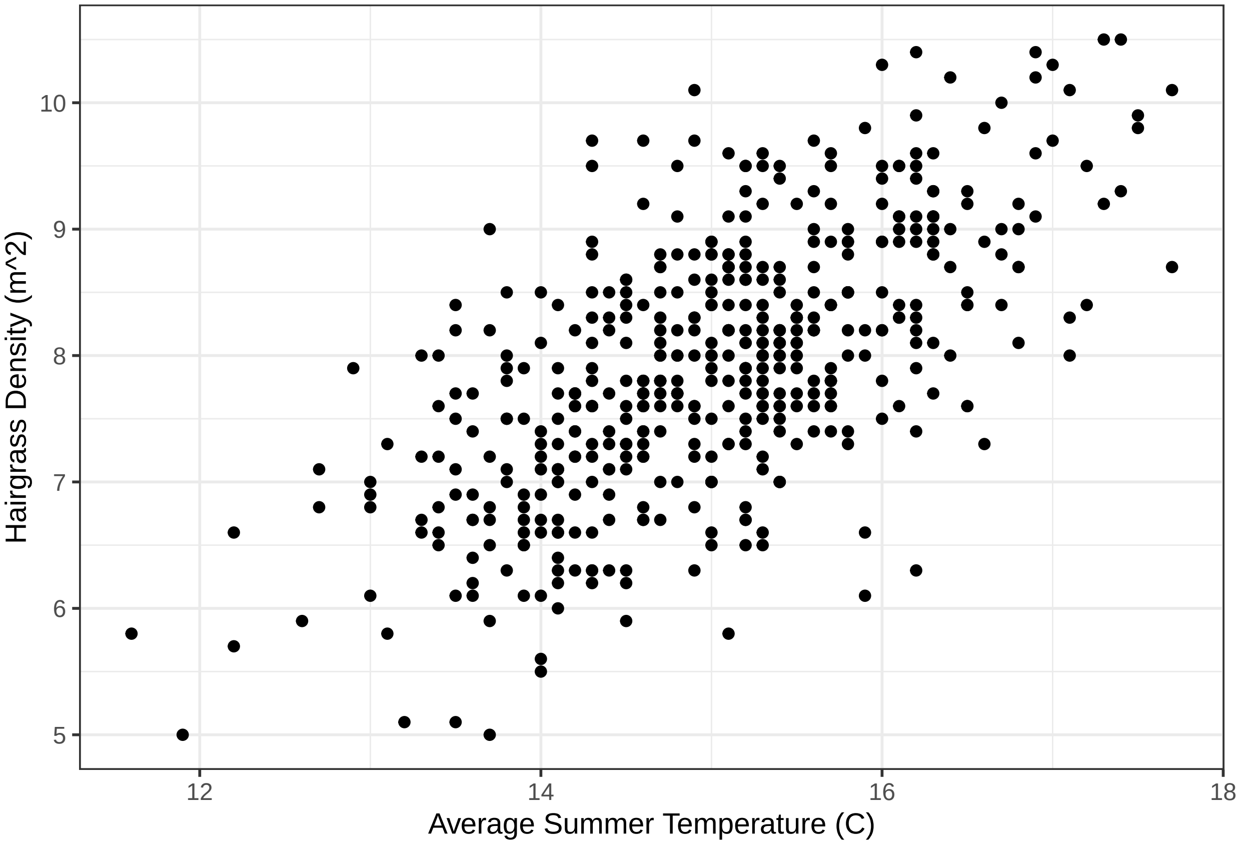
if they put in the wrong variables and get the incorrect model, they can still get the rest of the points as long as they interpret the one they made correctly

Summer Temperature

Now let's do the same thing for the average summer temperatures.

12. Create a scatter plot of hair grass density and average summer temperature. Remember to improve the axes labels and add a theme!

```
ggplot(hairgrass, aes(x = avg_summer_temp, y = hairgrass_density_m2)) +
  geom_point() +
  labs(x = "Average Summer Temperature (C)",
       y = "Hairgrass Density (m^2)") +
  theme_bw()
```



13. Write 1-2 sentences interpreting the plot above. Is this a positive relationship, negative relationship or no relationship at all? Based on your prediction, do you think the correlation coefficient will be positive, negative, or zero?

    Answer: positive relationship; r will be positive

14. Calculate the correlation coefficient, `r`.

```
r <- cor(y = hairgrass$hairgrass_density_m2, x = hairgrass$avg_summer_temp)
r
```

```
## [1] 0.643731
```

15. Calculate the `r^2` value. Write a one sentence interpretation of what the `r^2` value means in the context of these two variables.

```
r^2
```

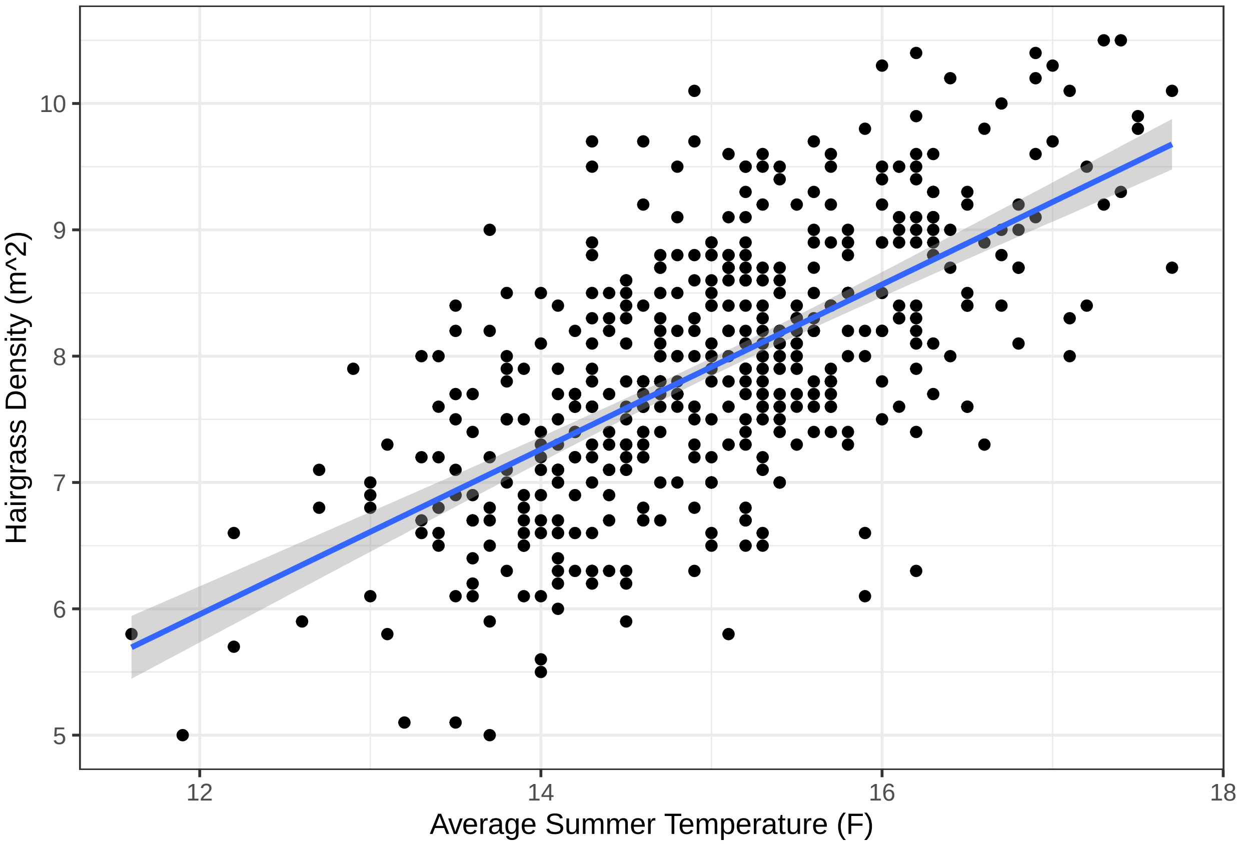
```
## [1] 0.4143896
```

Interpretation: This means that 40% of the variation in hair grass density can be explained by variation in average summer temperature.

16. Create the scatter plot that includes the line of best fit (have `ggplot2` calculate the linear equation for you). Make the labels easier to interpret and add a theme.

```
ggplot(hairgrass, aes(x = avg_summer_temp, y = hairgrass_density_m2)) +
  geom_point() +
  geom_smooth(method = "lm") +
  ylab("Hairgrass Density (m^2)") +
  xlab("Average Summer Temperature (F)") +
  theme_bw()
```

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



17. Using code, create the regression model in R and obtain the summary of it.

```
mod <- lm(hairgrass$hairgrass_density_m2 ~ hairgrass$avg_summer_temp)
summary(mod)
```

```
##
## Call:
## lm(formula = hairgrass$hairgrass_density_m2 ~ hairgrass$avg_summer_temp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.40214 -0.49902 -0.01046  0.51215  2.25066
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1.8774     0.5343  -3.514 0.000483 ***
## hairgrass$avg_summer_temp   0.6528     0.0355  18.391  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7869 on 478 degrees of freedom
## Multiple R-squared:  0.4144, Adjusted R-squared:  0.4132
## F-statistic: 338.2 on 1 and 478 DF,  p-value: < 2.2e-16
```

18. Interpret the model summary. What is the p-value for our variable of interest? Do we accept or reject the null hypothesis regarding the relationship between these two variables? What can we conclude then about building a road? (2 points)

    Answer:

p-value is <2e-16, so super small and highly significant; reject the null; should think about temperature when we decide where to build the road