Regression Homework

Mila Pruiett

Statistical inference and regression analyses homework

Your name:

Date:

Score out of 20 points:

The assignment

In this homework, you are going to look at the relationship between hairgrass density and another variable, the average summer temperature. The sites we sampled were along a gradient of summer temperatures, but we don't know if that is at all related to the growth of the hairgrass.

1. First load the tidyverse and read in the data. (1 pts)

```
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()
```

```
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.
```

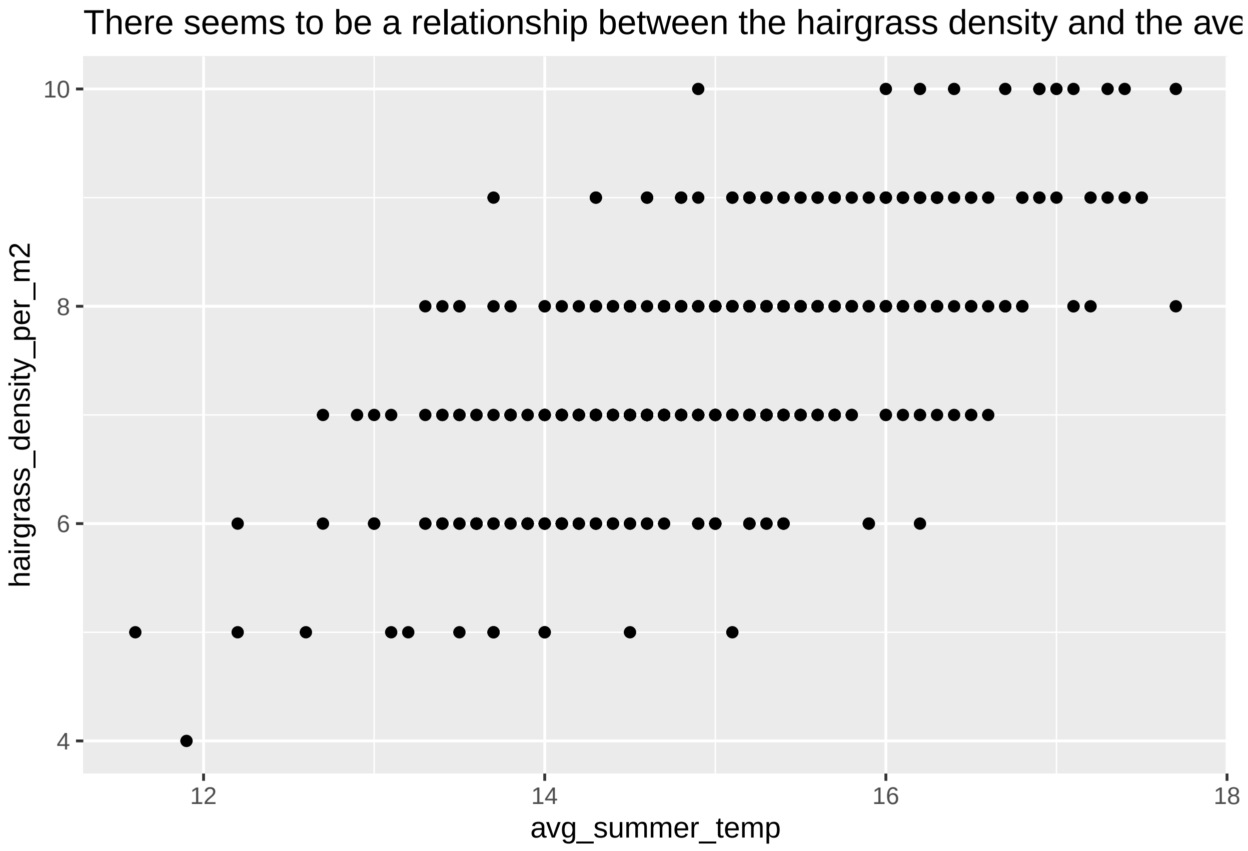
2. Calculate descriptive statistics for the temperature (mean and standard deviation). As a reminder, this should NOT be grouped by another categorical variable (our data set doesn't even have categorical variables) (2pts)

```
hairgrass %>% summarize(mean(avg_summer_temp), sd(avg_summer_temp))
```

```
## # A tibble: 1 x 2
##   `mean(avg_summer_temp)` `sd(avg_summer_temp)`
##                     <dbl>                 <dbl>
## 1                    15.0                  1.01
```

3. Create a scatter plot of hairgrass density and average summer temperature. Think carefully about which variable is the independent (x axis) and which is the dependent (y axis). Give your plot a title. hint: ggtitle("title here") Write a few sentences about what you think the relationship between hairgrass density and avg summer temperature is (is it related at all? are they positively related? negatively related?) (4 pts)

```
hairgrass %>%
  ggplot(aes(x = avg_summer_temp, y = hairgrass_density_per_m2)) +
  geom_point() +
  ggtitle("There seems to be a relationship between the hairgrass density and the average summer temp")
```



```
# I think there is a positive relationship. As temp increases, so does hairgrass density (anything is f
```

4. Calculate the correlation coefficient, r. Calculate r-squared. Interpret what r squared means for this relationship. (3pts)

```
r <- cor(y = hairgrass$hairgrass_density_per_m2, x = hairgrass$avg_summer_temp)
r^2
```

```
## [1] 0.3988874
```

```
# This means that 40% of the variation in hairgrass density can be explained by variation in the averag
```

5. What would the model for hairgrass density and summer temperature be? (Write it without the numbers for the coefficients) (1 pt)

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

6. 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 avg_summer_temp
# alt: There is a relationship between hairgrass density and avg_summer_temp
```

7. Create the model in R and obtain the summary of it. What is the model? What is the p-value associated with the F-statistic? Do we accept or reject the null hypothesis regarding the relationship between these two variables? What can we conclude then about building a road? (5 pts)

```
mod <- lm(hairgrass$hairgrass_density_per_m2 ~hairgrass$avg_summer_temp)
summary(mod) # 1pt
```

```
##
## Call:
## lm(formula = hairgrass$hairgrass_density_per_m2 ~ hairgrass$avg_summer_temp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.49670 -0.56407  0.03632  0.57001  2.63672
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -2.57656    0.56378   -4.57 6.21e-06 ***
## hairgrass$avg_summer_temp  0.66710    0.03746   17.81  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8304 on 478 degrees of freedom
## Multiple R-squared:  0.3989, Adjusted R-squared:  0.3976
## F-statistic: 317.2 on 1 and 478 DF,  p-value: < 2.2e-16
```

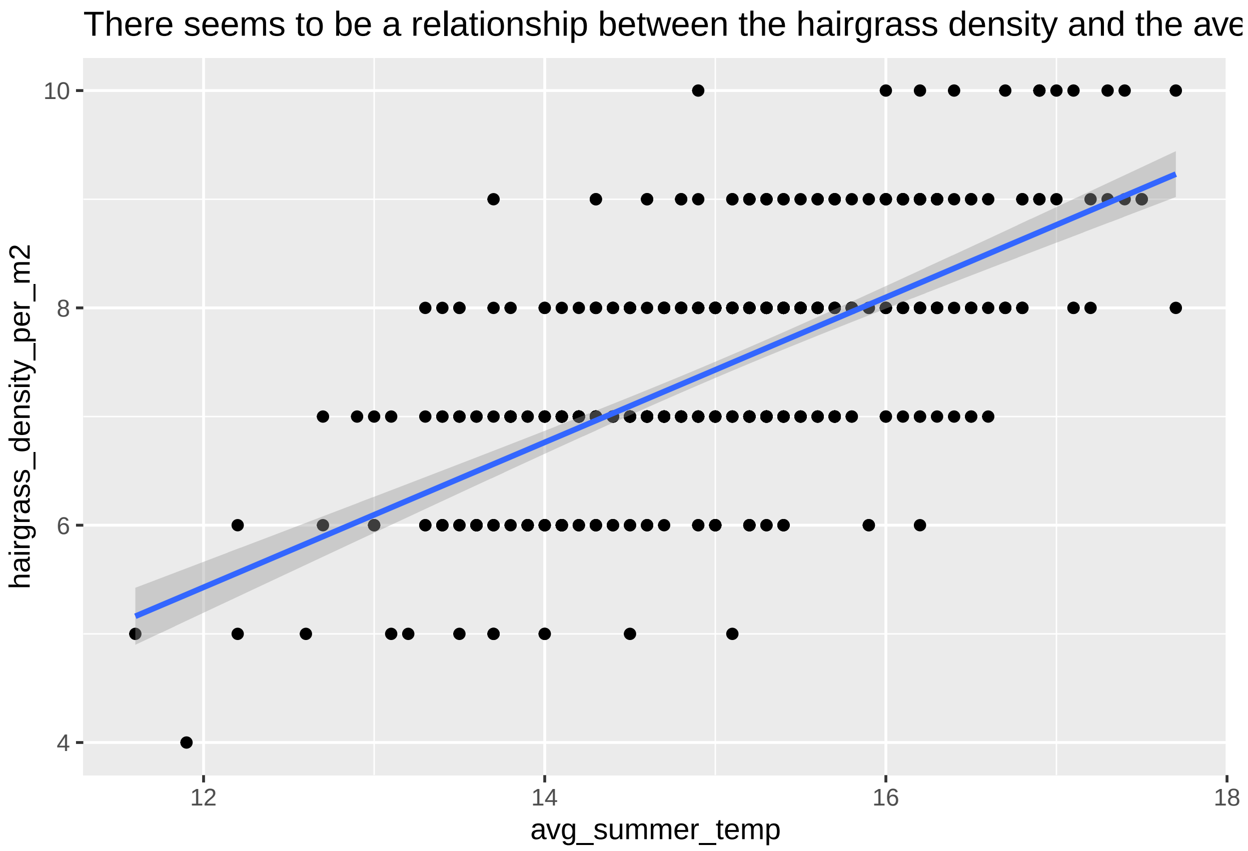
```
# the p-value is <2.2e-16
# We can reject the null hypothesis
# There is a relationship between average summer temperature and hairgrass ensity
# Because it is a positie value, we should consider building a road in areas where it is colder where t

# if they put in the wrong variables and get the incorrect model, they can still get the rest of the po
```

8. Create the scatter plot that includes the line of best fit (of which you now know the formula for)! (2 pts)

```
hairgrass %>%
  ggplot(aes(x = avg_summer_temp, y = hairgrass_density_per_m2)) +
  geom_point() +
  ggtitle("There seems to be a relationship between the hairgrass density and the average summer temp")
  geom_smooth(method = "lm")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```


There seems to be a relationship between the hairgrass density and the ave
hairgrass_density_per_m2
avg_summer_temp