Homework 1 Suglia

Elena Suglia

10/15/2019

Question 1

Load the data set "CO2_HW1.txt", which describes the CO2 uptake rates of plants of the grass species Echinochloa crus-galli from Quebec and Mississippi.

```
d = read.table("CO2_HW1.txt", header = TRUE)
# Check the loaded data
head(d)
```

```
##     Type uptake  logconc
## 1 Quebec   14.2 4.553877
## 2 Quebec   24.1 5.164786
## 3 Quebec   30.3 5.521461
## 4 Quebec   34.6 5.857933
## 5 Quebec   32.5 6.214608
## 6 Quebec   35.4 6.514713
```

- Looks like it loaded correctly

Using a linear model for the analysis, investigate these questions:

How does the air concentration of CO2 ("logconc") affect a grass plant's CO2 uptake rate ("uptake")? Does this effect depend on the origin of the plant ("Type")?

In your answer, include some information on: What transformations if any you made on the data and why. What steps you took to check model assumptions and model performance. What the coefficients of the model are and how you interpret them.

Check structure of data frame

```
class(d) # data.frame
```

```
## [1] "data.frame"
```

```
str(d) # structure of the dataframe and data types in each column
```

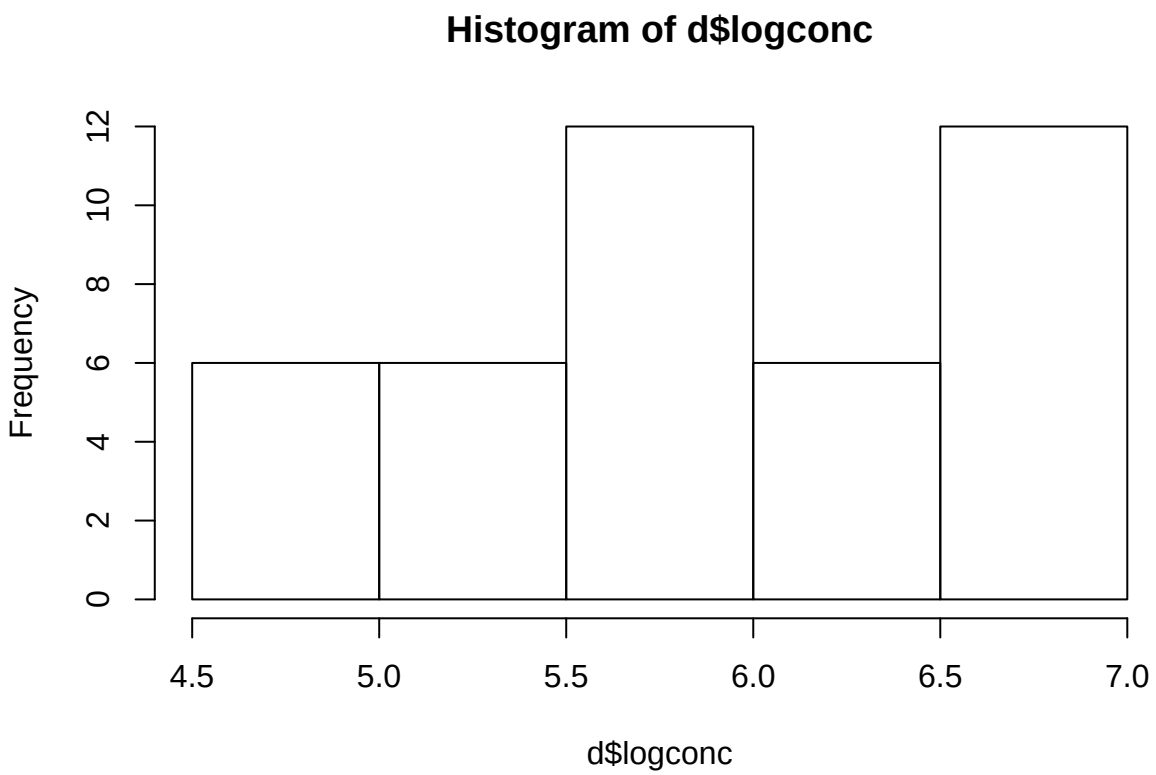
```
## 'data.frame':    42 obs. of  3 variables:
##  $ Type   : Factor w/ 2 levels "Mississippi",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ uptake : num  14.2 24.1 30.3 34.6 32.5 35.4 38.7 9.3 27.3 35 ...
##  $ logconc: num  4.55 5.16 5.52 5.86 6.21 ...
```

- Everything looks good

Visual checks for normality/distribution of the data

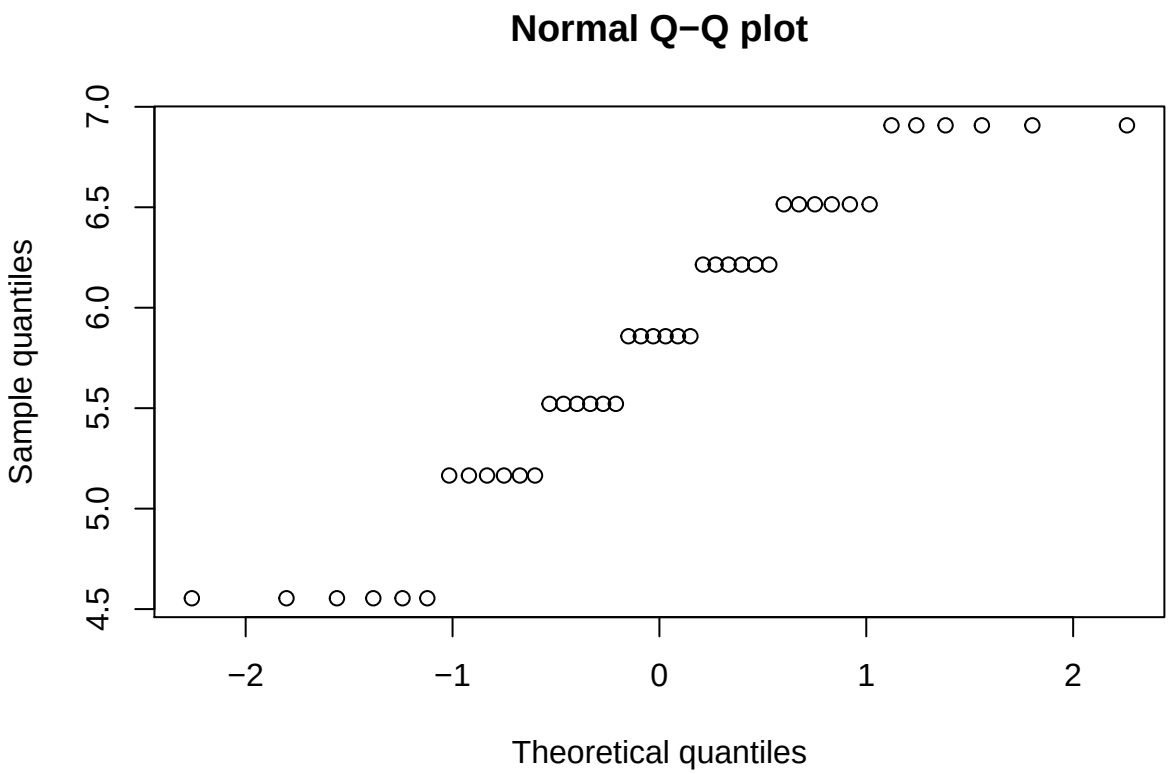
Histogram

```
hist(d$logconc)
```



Normal Q-Q plot

```
qqnorm(d$logconc, main = "Normal Q-Q plot", xlab = "Theoretical quantiles", ylab = "Sample quantiles")
```

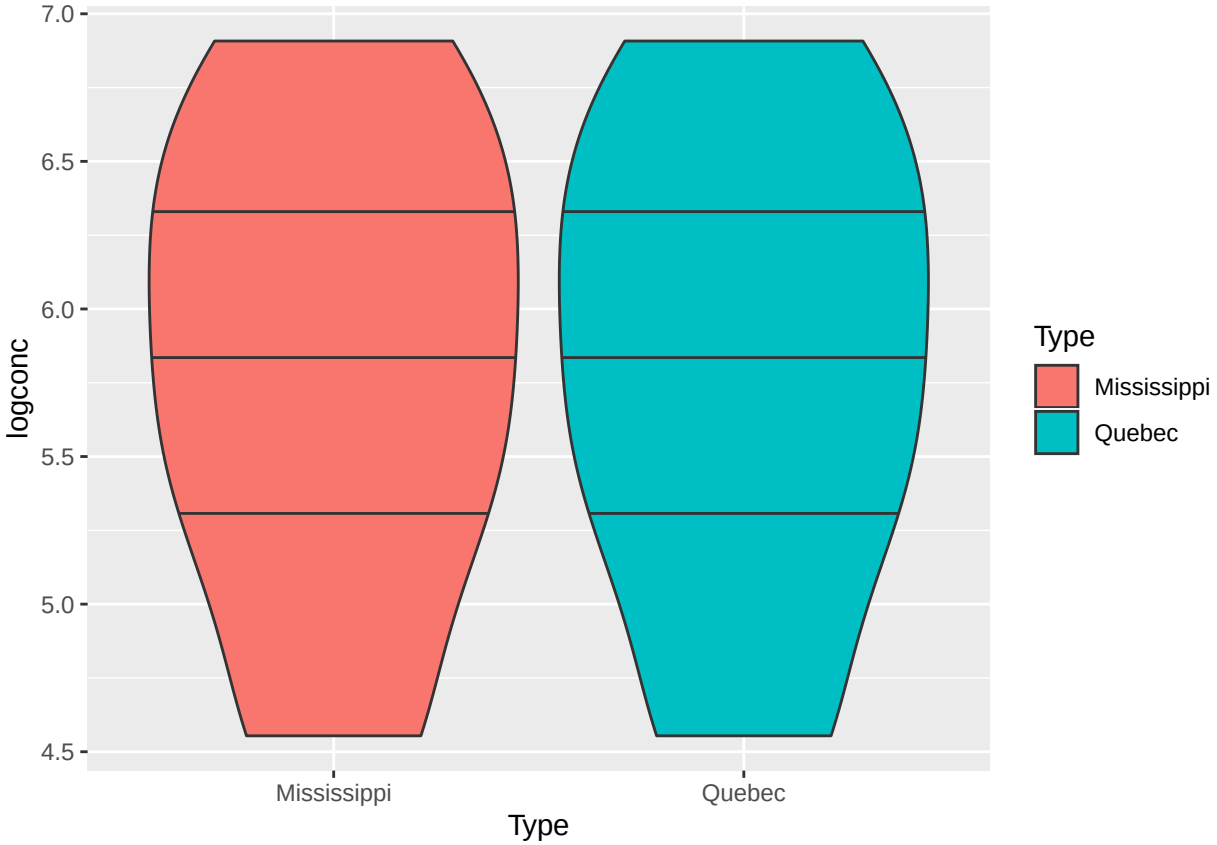


- Data does not appear to follow a normal distribution
- Skewed to the left based on histogram or perhaps overdispersed based on Q-Q plot

Violin plots to look at the spread of the data

First, look at spread of log concentration of CO2 uptake

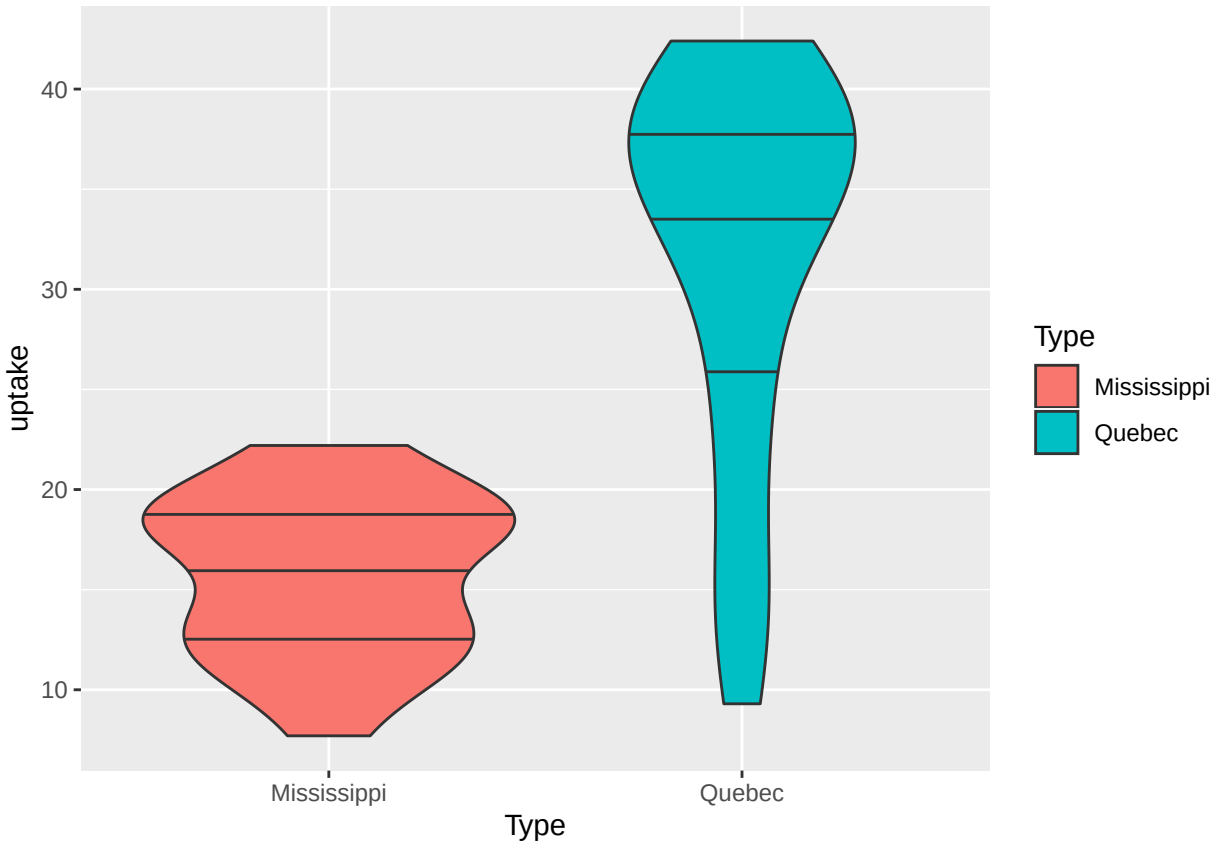
```
ggplot(d, aes(x=Type, y=logconc)) + geom_violin(draw_quantiles = c(0.25, 0.5, 0.75), aes(fill=Type))
```



- These plots tell us that there is virtually no difference in spread of log concentration of CO2 between the 2 sites

Now, let's look at the spread of the response variable, CO2 uptake rate

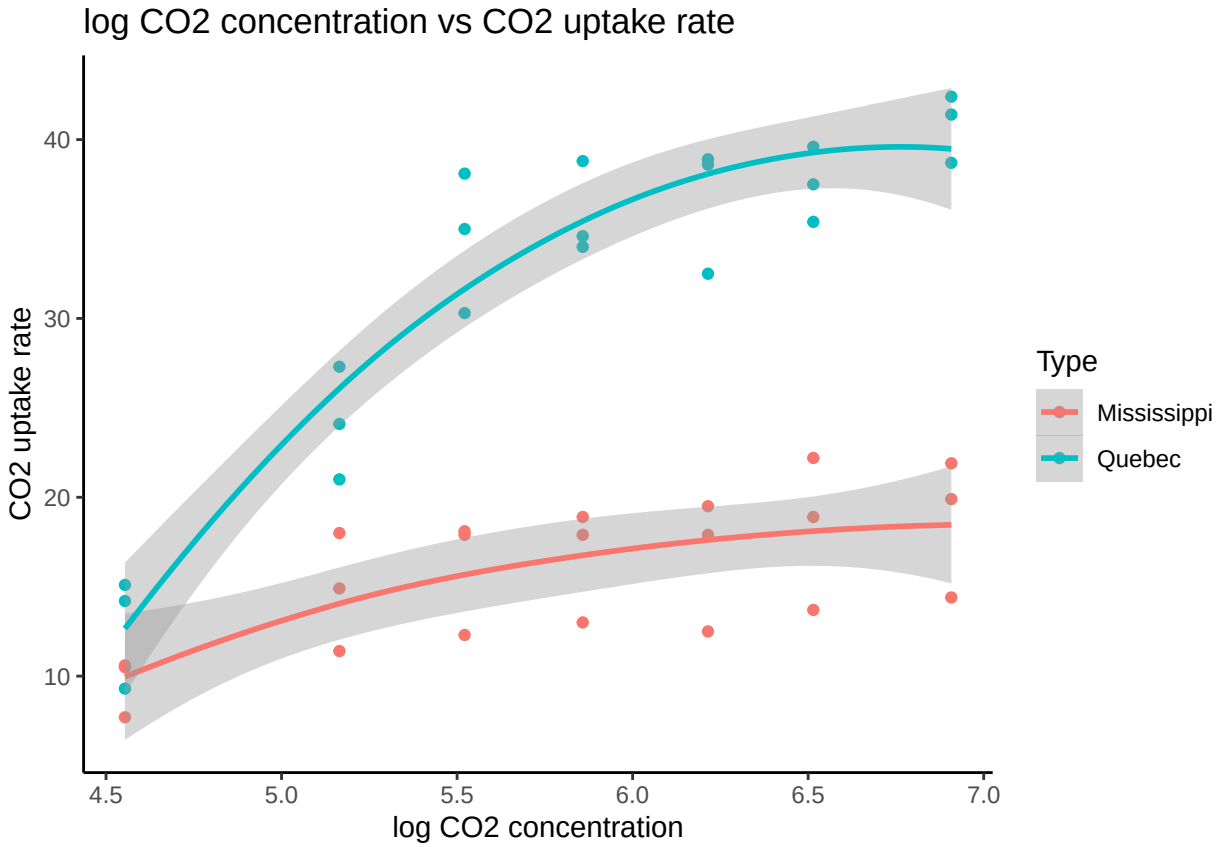
```
ggplot(d, aes(x=Type, y=uptake)) + geom_violin(draw_quantiles = c(0.25, 0.5, 0.75), aes(fill=Type))
```



- There is definitely a difference in CO2 uptake rates between sites (higher uptake rates in Quebec)

Relationship between response and explanatry variables: Scatter plot of the log CO2 concentration versus the rate of CO2 uptake

```
ggplot(d, aes(x=logconc, y=uptake, group = Type, color = Type)) +
  geom_point() +
  geom_smooth(span=2) +
  theme_classic() +
ggtitle("log CO2 concentration vs CO2 uptake rate") +
  xlab("log CO2 concentration") +
  ylab("CO2 uptake rate")
```



Higher log CO2 concentration appears positively correlated with uptake rate. There are higher overall uptake rates in Quebec. Here, you are able to see that there appears to be an interaction between location and response to CO2 concentration: Quebec plants increase their CO2 uptake rates more in response to an increase in log CO2 concentration than Mississippi plants do. Thus, the relationship between log CO2 concentration and uptake depends on location, or in other words, the association between the response and the explanatory variables depends on the level of a third variable.

Fit model

Let's compare model fit between several different models: one that has one explanatory variable (uptake rate), one that accounts for location, and one that has an interaction between uptake rate and location.

Only looks at relationship between log concentration CO2 and CO2 uptake rate

```
m1 = lm(uptake~logconc, d)
display(m1)
```

```
## lm(formula = uptake ~ logconc, data = d)
##             coef.est coef.se
## (Intercept) -18.82    11.42
## logconc       7.32     1.95
## ---
## n = 42, k = 2
## residual sd = 9.47, R-Squared = 0.26
```

Accounts for location

```
m2 = lm(uptake~logconc + Type, d)
display(m2)
```

```
## lm(formula = uptake ~ logconc + Type, data = d)
##             coef.est coef.se
## (Intercept) -26.79     5.91
## logconc       7.32     1.00
## TypeQuebec   15.94     1.50
## ---
## n = 42, k = 3
## residual sd = 4.86, R-Squared = 0.81
```

Includes interaction between uptake rate and location

```
m3 = lm(uptake~logconc*Type, d)
display(m3)
```

```
## lm(formula = uptake ~ logconc * Type, data = d)
##                     coef.est coef.se
## (Intercept)          -4.40     6.61
## logconc               3.47     1.13
## TypeQuebec          -28.85     9.35
## logconc:TypeQuebec    7.70     1.59
## ---
## n = 42, k = 4
## residual sd = 3.88, R-Squared = 0.88
```

The model that includes the interaction between uptake rate and location has the best fit: its residual standard deviation is lowest and the R-squared value is highest (this model explains the most variation in the response variable).

put here your interpretation of the coefficients of the model

Assumptions of a linear regression model

- Validity - our question is simply about the effects of one variable on the other; so in this case the regression model seems to be a valid test of this
- Linearity - relationships look fairly linear on the scatterplot
- Independence of errors
- Equal variance of errors
- Normality of errors

Add in a test of model fit: fitted vs observed residuals and a Q-Q plot

Script 2 has some useful code

Should we do any centering and standardization? P. 55 of G&H says it's especially useful when there are interactions

Question 2

Load the data set "ecdata_HW1.txt", which includes some growth and flowering time information on some Erodium cicutarium plants from serpentine and non-serpentine environments. The columns are: sourceSOILTYPE: soil type of source population, 1 = non-serpentine, 2 = serpentine earlylfno: count of leaves early in the plant's growth totallfno: count of total leaves at end of experiment ffdate: date of first flowering in days after germination

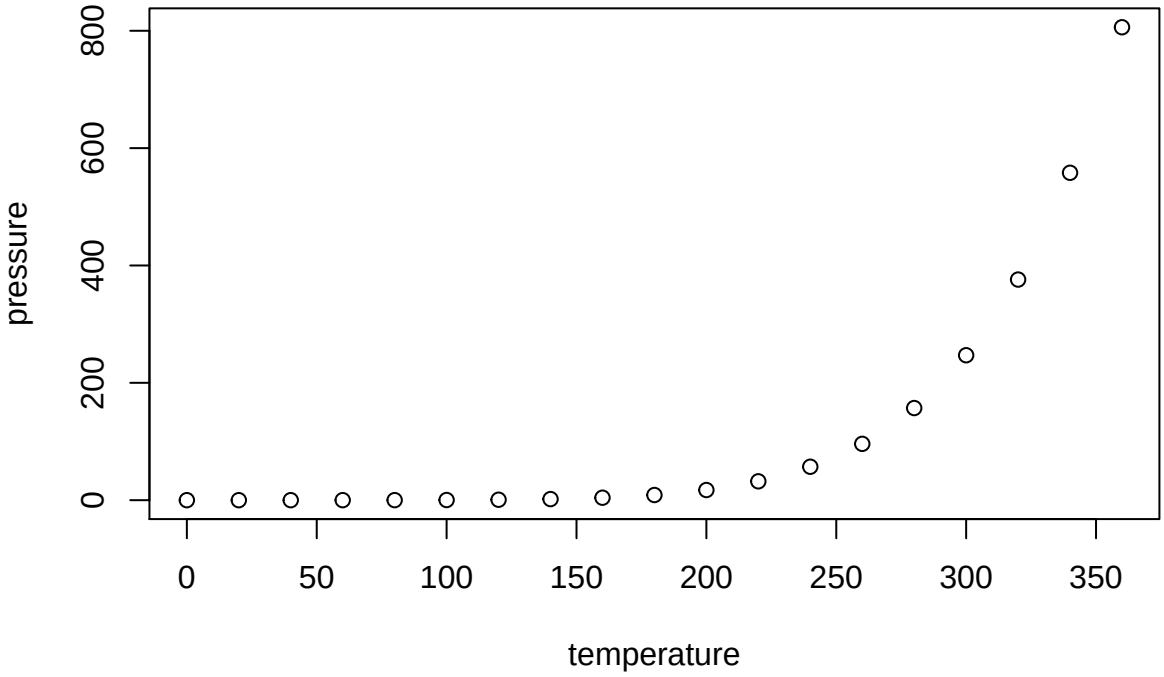
```
ec = read.table("ecdata_HW1.txt")
```

Fit a normal distribution to the Erodium ffdate data. Also fit a gamma distribution – does this distribution fit the data better or worse than the normal distribution does? Which is "better" by AIC score, or they both about the same?

Calculate the log-likelihood for the normal distribution at the fitted values of the parameters. Show (graphically or in numbers) that the log-likelihood of the data becomes more negative if you shift the mean parameter value away from its maximum-likelihood value.

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.