# Project: CO2, Carbon Dioxide Uptake in Grass Plants

Carayon Chloé - Taillieu Victor

16/11/2020

## Description of the dataset

The CO2 data frame is the result of an experiment based on the cold tolerance of the grass species Echinochloa crus-galli. We load the data and obtain some information thanks to those commands:

```
data(CO2)
dim(CO2)
```

```
## [1] 84  5
```

```
names(CO2)
```

```
## [1] "Plant"     "Type"      "Treatment" "conc"      "uptake"
```

The CO2 dataset is composed of 84 rows ans 5 columns of names "Plant", "Type", "Treatment", "conc" and "uptake". Using the help section, we learn that:

- Plant is a factor giving to each plant a unique identifier:
  - Q or M represents the origin of the plant, respectively Quebec or Mississippi
  - n or c gives information about the treatment
  - 1, 2 or 3 is the id of the plant

For example, Mc1 is plant 1 from Mississippi and was chilled.

- Type is a factor with levels Quebec and Mississippi
- Treatment is a factor with levels chilled and nonchilled
- conc is the ambient carbon dioxide concentrations' vector, measured in mL/L
- uptake is the carbon dioxide uptake rates' vector in umol/$m^2$ sec

The CO2 uptake rate of six plants from Quebec and six plants from Mississippi was measured at several levels of ambient CO2 concentration. Half the plants were chilled overnight before the experiment was conducted.

## Data exploration

### Summary

Let's display the dataset and its structure:

```
CO2
```

```
##    Plant        Type  Treatment conc uptake
## 1    Qn1      Quebec nonchilled   95   16.0
## 2    Qn1      Quebec nonchilled  175   30.4
## 3    Qn1      Quebec nonchilled  250   34.8
## 4    Qn1      Quebec nonchilled  350   37.2
## 5    Qn1      Quebec nonchilled  500   35.3
## 6    Qn1      Quebec nonchilled  675   39.2
## 7    Qn1      Quebec nonchilled 1000   39.7
## 8    Qn2      Quebec nonchilled   95   13.6
## 9    Qn2      Quebec nonchilled  175   27.3
## 10   Qn2      Quebec nonchilled  250   37.1
## 11   Qn2      Quebec nonchilled  350   41.8
## 12   Qn2      Quebec nonchilled  500   40.6
## 13   Qn2      Quebec nonchilled  675   41.4
## 14   Qn2      Quebec nonchilled 1000   44.3
## 15   Qn3      Quebec nonchilled   95   16.2
## 16   Qn3      Quebec nonchilled  175   32.4
## 17   Qn3      Quebec nonchilled  250   40.3
## 18   Qn3      Quebec nonchilled  350   42.1
## 19   Qn3      Quebec nonchilled  500   42.9
## 20   Qn3      Quebec nonchilled  675   43.9
## 21   Qn3      Quebec nonchilled 1000   45.5
## 22   Qc1      Quebec    chilled   95   14.2
## 23   Qc1      Quebec    chilled  175   24.1
## 24   Qc1      Quebec    chilled  250   30.3
## 25   Qc1      Quebec    chilled  350   34.6
## 26   Qc1      Quebec    chilled  500   32.5
## 27   Qc1      Quebec    chilled  675   35.4
## 28   Qc1      Quebec    chilled 1000   38.7
## 29   Qc2      Quebec    chilled   95    9.3
## 30   Qc2      Quebec    chilled  175   27.3
## 31   Qc2      Quebec    chilled  250   35.0
## 32   Qc2      Quebec    chilled  350   38.8
## 33   Qc2      Quebec    chilled  500   38.6
## 34   Qc2      Quebec    chilled  675   37.5
## 35   Qc2      Quebec    chilled 1000   42.4
## 36   Qc3      Quebec    chilled   95   15.1
## 37   Qc3      Quebec    chilled  175   21.0
## 38   Qc3      Quebec    chilled  250   38.1
## 39   Qc3      Quebec    chilled  350   34.0
## 40   Qc3      Quebec    chilled  500   38.9
## 41   Qc3      Quebec    chilled  675   39.6
## 42   Qc3      Quebec    chilled 1000   41.4
## 43   Mn1 Mississippi nonchilled   95   10.6
## 44   Mn1 Mississippi nonchilled  175   19.2
## 45   Mn1 Mississippi nonchilled  250   26.2
## 46   Mn1 Mississippi nonchilled  350   30.0
## 47   Mn1 Mississippi nonchilled  500   30.9
## 48   Mn1 Mississippi nonchilled  675   32.4
## 49   Mn1 Mississippi nonchilled 1000   35.5
## 50   Mn2 Mississippi nonchilled   95   12.0
## 51   Mn2 Mississippi nonchilled  175   22.0
```

```
## 52   Mn2 Mississippi nonchilled  250   30.6
## 53   Mn2 Mississippi nonchilled  350   31.8
## 54   Mn2 Mississippi nonchilled  500   32.4
## 55   Mn2 Mississippi nonchilled  675   31.1
## 56   Mn2 Mississippi nonchilled 1000   31.5
## 57   Mn3 Mississippi nonchilled   95   11.3
## 58   Mn3 Mississippi nonchilled  175   19.4
## 59   Mn3 Mississippi nonchilled  250   25.8
## 60   Mn3 Mississippi nonchilled  350   27.9
## 61   Mn3 Mississippi nonchilled  500   28.5
## 62   Mn3 Mississippi nonchilled  675   28.1
## 63   Mn3 Mississippi nonchilled 1000   27.8
## 64   Mc1 Mississippi    chilled   95   10.5
## 65   Mc1 Mississippi    chilled  175   14.9
## 66   Mc1 Mississippi    chilled  250   18.1
## 67   Mc1 Mississippi    chilled  350   18.9
## 68   Mc1 Mississippi    chilled  500   19.5
## 69   Mc1 Mississippi    chilled  675   22.2
## 70   Mc1 Mississippi    chilled 1000   21.9
## 71   Mc2 Mississippi    chilled   95    7.7
## 72   Mc2 Mississippi    chilled  175   11.4
## 73   Mc2 Mississippi    chilled  250   12.3
## 74   Mc2 Mississippi    chilled  350   13.0
## 75   Mc2 Mississippi    chilled  500   12.5
## 76   Mc2 Mississippi    chilled  675   13.7
## 77   Mc2 Mississippi    chilled 1000   14.4
## 78   Mc3 Mississippi    chilled   95   10.6
## 79   Mc3 Mississippi    chilled  175   18.0
## 80   Mc3 Mississippi    chilled  250   17.9
## 81   Mc3 Mississippi    chilled  350   17.9
## 82   Mc3 Mississippi    chilled  500   17.9
## 83   Mc3 Mississippi    chilled  675   18.9
## 84   Mc3 Mississippi    chilled 1000   19.9
```

```
class(CO2)
```

```
## [1] "nfnGroupedData" "nfGroupedData"  "groupedData"    "data.frame"
```

```
str(CO2)
```

```
## Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame':   84 obs. of  5 variables
##  $ Plant    : Ord.factor w/ 12 levels "Qn1"<"Qn2"<"Qn3"<..: 1 1 1 1 1 1 1 2 2 2 ...
##  $ Type     : Factor w/ 2 levels "Quebec","Mississippi": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Treatment: Factor w/ 2 levels "nonchilled","chilled": 1 1 1 1 1 1 1 1 1 1 ...
##  $ conc     : num  95 175 250 350 500 675 1000 95 175 250 ...
##  $ uptake   : num  16 30.4 34.8 37.2 35.3 39.2 39.7 13.6 27.3 37.1 ...
##  - attr(*, "formula")=Class 'formula'  language uptake ~ conc | Plant
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
##  - attr(*, "outer")=Class 'formula'  language ~Treatment * Type
##   .. ..- attr(*, ".Environment")=<environment: R_EmptyEnv>
##  - attr(*, "labels")=List of 2
##   ..$ x: chr "Ambient carbon dioxide concentration"
##   ..$ y: chr "CO2 uptake rate"
```

```
##  - attr(*, "units")=List of 2
##   ..$ x: chr "(uL/L)"
##   ..$ y: chr "(umol/m^2 s)"
```

The output confirms what we have seen before thanks to the functions and help section.

This dataset was originally part of another package (nlme) which has its own grouped-data classes. That is why there are some unusual classes like nfnGroupedData, nfGroupedData, groupedData and attr.

```
summary(CO2)
```

```
##      Plant             Type         Treatment       conc          uptake
##  Qn1    : 7   Quebec     :42   nonchilled:42   Min.   :  95   Min.   : 7.70
##  Qn2    : 7   Mississippi:42   chilled   :42   1st Qu.: 175   1st Qu.:17.90
##  Qn3    : 7                                    Median : 350   Median :28.30
##  Qc1    : 7                                    Mean   : 435   Mean   :27.21
##  Qc3    : 7                                    3rd Qu.: 675   3rd Qu.:37.12
##  Qc2    : 7                                    Max.   :1000   Max.   :45.50
##  (Other):42
```

```
levels(CO2$Plant)
```

```
##  [1] "Qn1" "Qn2" "Qn3" "Qc1" "Qc3" "Qc2" "Mn3" "Mn2" "Mn1" "Mc2" "Mc3" "Mc1"
```

Thanks to the summary, we obtain information about each column. We execute the levels function on CO2$Plant as we don't have all the information displayed in the summary. We see that 3 plants were put in several conditions (with their name corresponding to the type, treatment and number) and measurements were made on their carbon dioxide uptake rates according to the concentration of ambient carbon dioxide.

### Characteristics

- By grouping plants by their type (Quebec - Mississipi), we can see some characteristics:

```
tapply(CO2$uptake, CO2$Type, summary)
```

```
## $Quebec
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.30   30.32   37.15   33.54   40.15   45.50
##
## $Mississippi
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.70   13.88   19.30   20.88   28.05   35.50
```

Plants from Quebec have an uptake rate between 9.30 and 45.50 umol/$m^2$ sec whereas for Mississippi, it is between 7.70 and 35.50 umol/$m^2$ sec. Moreover, for Quebec, half of the plants have an uptake rate greater than 37.15 while for Mississippi, it is 19.30. There are huge differences between Quebec and Mississippi minimums, maximums, medians and also quartiles. A quarter of Quebec plants have an uptake rate under 30.32 while for Mississipi it is 13.88. Finally, for 75% of Mississippi plants, the uptake rate is less than 28.05 whereas for Quebec it is less than 40.15.

So it appears that Quebec plants seem to have higher uptake rates than Mississippi plants.

- By grouping plants by their treatment (chilled - nonchilled), we can also see some characteristics:

```
tapply(CO2$uptake, CO2$Treatment, summary)
```

```
## $nonchilled
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.60   26.48   31.30   30.64   38.70   45.50
##
## $chilled
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.70   14.53   19.70   23.78   34.90   42.40
```

Non chilled plants have an uptake rate between 10.60 and 45.50 umol/$m^2$ sec whereas for chilled ones, it is between 7.70 and 42.40 umol/$m^2$ sec. Moreover, for non chilled, half of the plants have an uptake rate greater than 31.30 while for chilled, it is 19.70. A quarter of non chilled plants have an uptake rate under 26.48 while for chilled it is 14.53. Finally, for 75% of chilled plants, the uptake rate is less than 34.90 whereas for non chilled it is less than 38.70.

So it appears that non chilled plants seem to have higher uptake rates than chilled ones.

Now let's combine our two characteristics, type and treatment, and take a look on two plants from Quebec, one chilled and one non chilled:

```
tapply(CO2$uptake, CO2$Plant, summary)
```

```
## $Qn1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   32.60   35.30   33.23   38.20   39.70
##
## $Qn2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.60   32.20   40.60   35.16   41.60   44.30
##
## $Qn3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.20   36.35   42.10   37.61   43.40   45.50
##
## $Qc1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.20   27.20   32.50   29.97   35.00   38.70
##
## $Qc3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.10   27.50   38.10   32.59   39.25   41.40
##
## $Qc2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.30   31.15   37.50   32.70   38.70   42.40
##
## $Mn3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.30   22.60   27.80   24.11   28.00   28.50
##
```

```
## $Mn2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   26.30   31.10   27.34   31.65   32.40
##
## $Mn1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.60   22.70   30.00   26.40   31.65   35.50
##
## $Mc2
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.70   11.85   12.50   12.14   13.35   14.40
##
## $Mc3
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.60   17.90   17.90   17.30   18.45   19.90
##
## $Mc1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.5    16.5    18.9    18.0    20.7    22.2
```

Let's take the examples of Qn3 and Qc3:

- Qn3 (third plant of Quebec nonchilled) have an uptake rate between 16.20 and 45.50 with an average of 37.61. Half of the Qn3 have a rate less than 42.10. In those plants, 25% have an uptake rate less than 36.35. Moreover, 75% have a rate less than 43.40. In fact 50% of the population uptake rates are between 36.35 and 43.40 mL/L.
- Qc3 (third plant of Quebec chilled): have an uptake rate between 15.10 and 41.40 with an average of 32.59. Half of the Qc3 have a rate less than 38.10. In those plants, 25% have an uptake rate less than 27.50. Moreover, 75% have a rate less than 39.25. In fact 50% of the population uptake rates are between 27.50 and 39.25 mL/L.

So we see that for two plants numbered 3 from the same origin (Quebec) but one chilled and the other non chilled, we obtain a significant difference in uptake rates.

We can sum up these differences using the summarize function from dplyr library.

```
library(dplyr)
summarize(group_by(CO2, Type), average_uptake = mean(uptake))
```

```
## # A tibble: 2 x 2
##   Type        average_uptake
##   <fct>                <dbl>
## 1 Quebec                33.5
## 2 Mississippi           20.9
```

```
summarize(group_by(CO2, Treatment), average_uptake = mean(uptake))
```

```
## # A tibble: 2 x 2
##   Treatment  average_uptake
##   <fct>               <dbl>
## 1 nonchilled           30.6
## 2 chilled              23.8
```

```
summarize(group_by(CO2, conc), average_uptake = mean(uptake))
```

```
## # A tibble: 7 x 2
##    conc average_uptake
##   <dbl>          <dbl>
## 1    95           12.3
## 2   175           22.3
## 3   250           28.9
## 4   350           30.7
## 5   500           30.9
## 6   675           32.0
## 7  1000           33.6
```

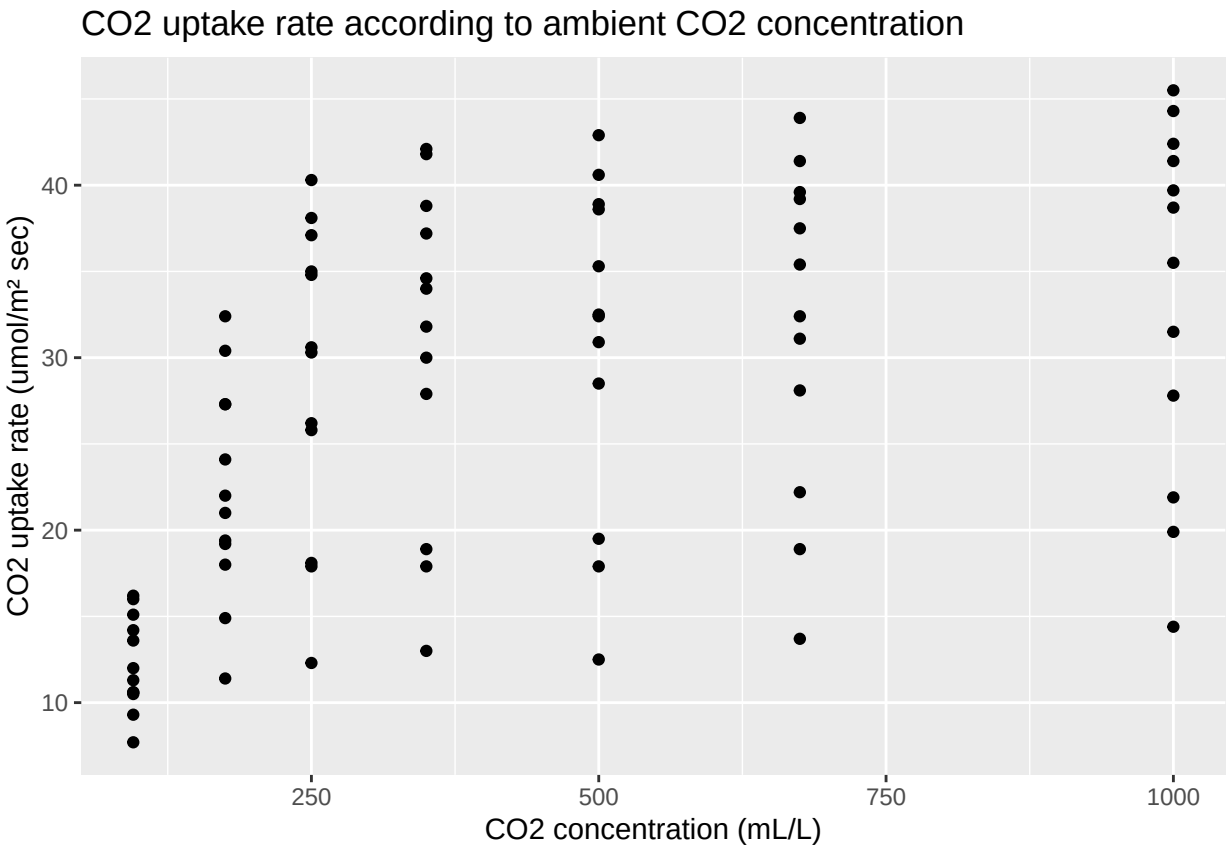
From these results, we observe that the average uptake for plants from Quebec is much higher than the one for plants from Mississippi (33.5 and 20.9 umol/$m^2$ sec respectively). If we compare them according to the treatment, chilled plants have a smaller average uptake rate compared to nonchilled ones (23.8 as opposed to 30.6 umol/$m^2$ sec). Finally, it makes sense to see that when we increase the concentration of ambient carbon dioxide, the plants tend to have higher carbon dioxide uptake rates.

## Data visualization

### Global overview

To begin with, let's just plot the data without any filtering:

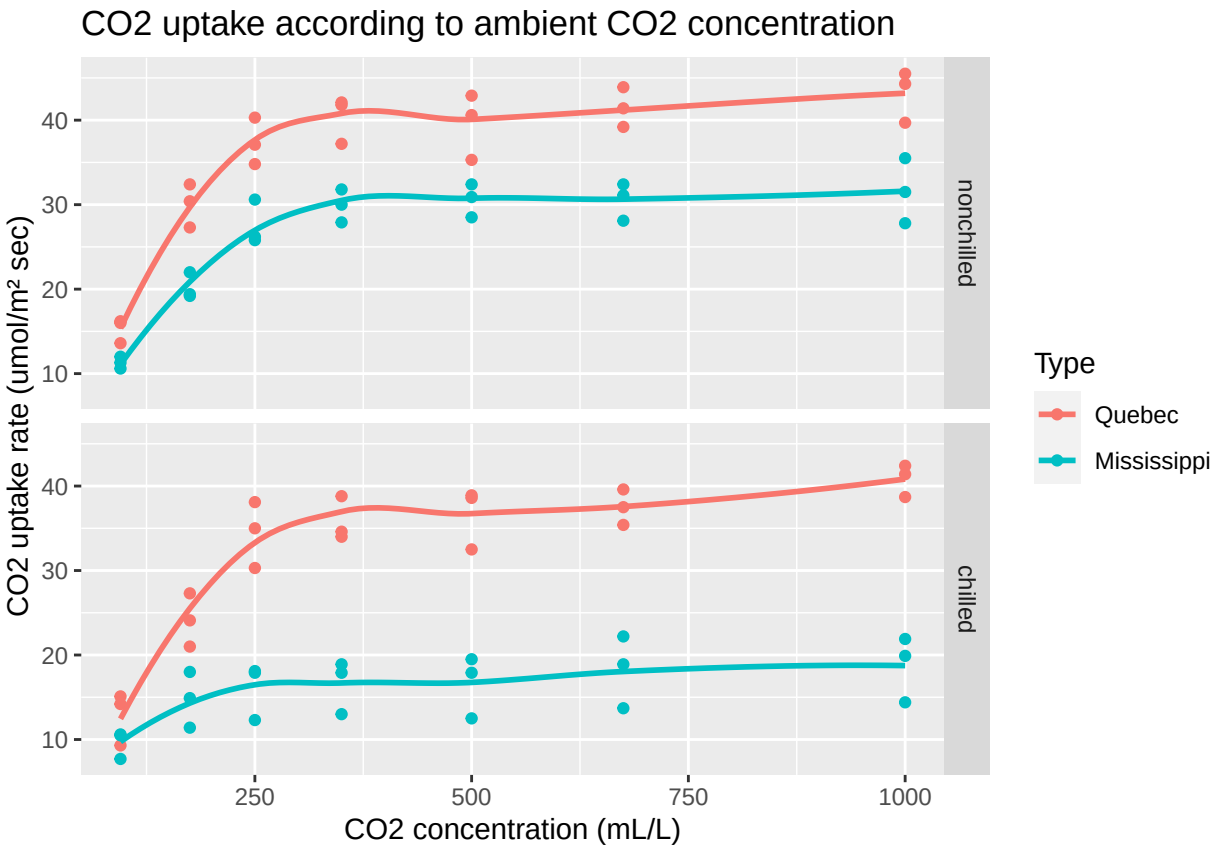
```
library(ggplot2)
qplot(conc, uptake, data = CO2, geom = "point",
      main = "CO2 uptake rate according to ambient CO2 concentration",
      xlab = "CO2 concentration (mL/L)", ylab = "CO2 uptake rate (umol/m² sec)")
```




The only thing we can say is that CO2 uptake rates seem to increase as ambient CO2 concentration increase.

Then, we can separate plants by their characteristics to get some insights:

```
qplot(conc, uptake, facets = Treatment ~ ., data = CO2, color = Type,
      geom = c("point", "smooth"), se = FALSE,
      main = "CO2 uptake according to ambient CO2 concentration",
      xlab = "CO2 concentration (mL/L)", ylab = "CO2 uptake rate (umol/m² sec)")
```

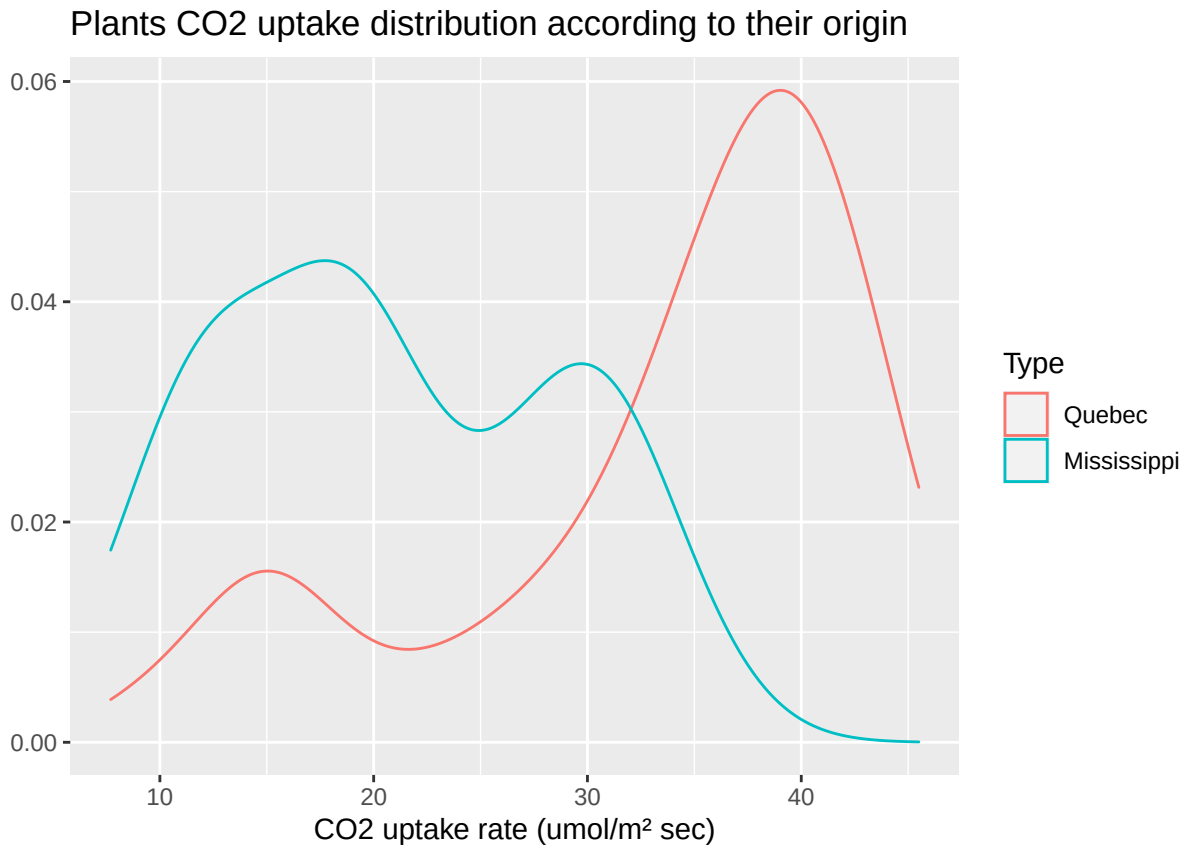



Here we can observe several interesting things:

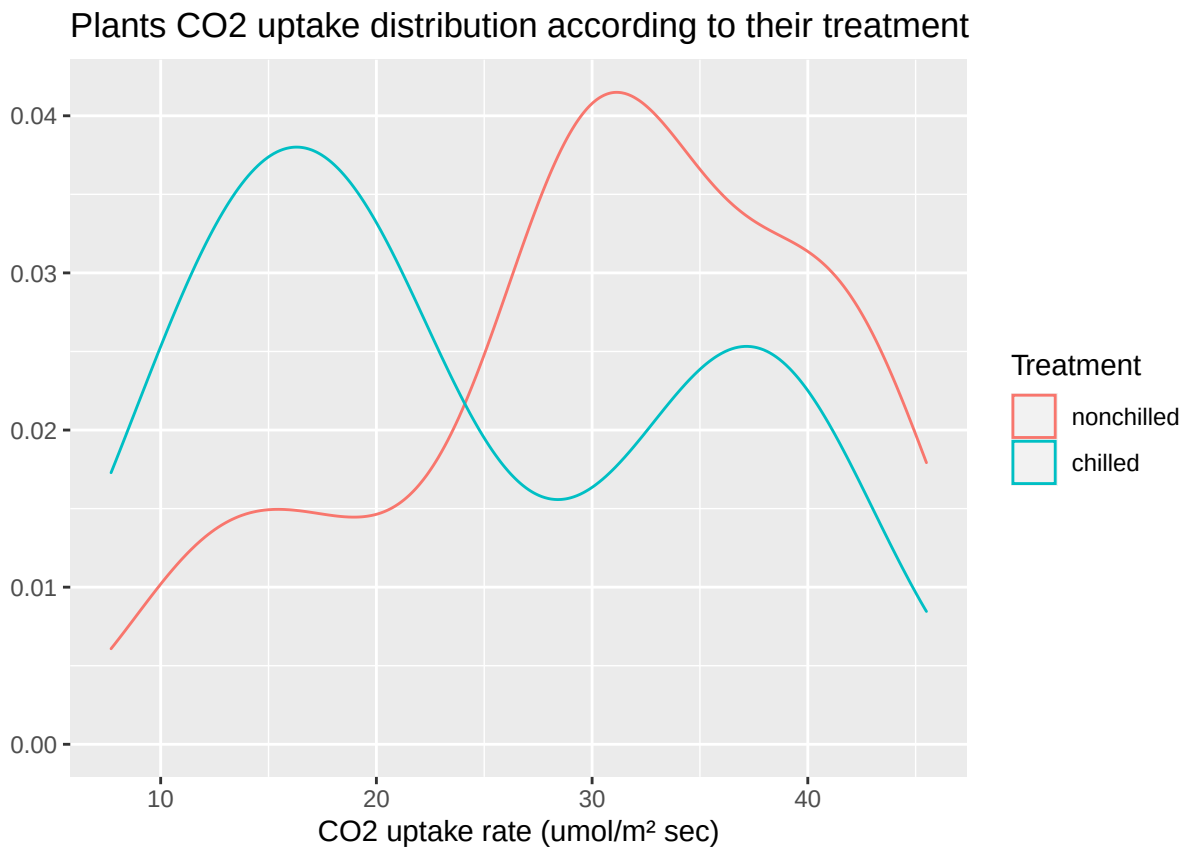
- As we noticed in the previous part, plants from Quebec (in red) almost always have a higher CO2 uptake rate compared to plants from Mississippi for the same ambient concentration. The smooth tendency lines clearly show that.
- Moreover, nonchilled plants (in the upper graph) seem to jump higher in terms of uptake rates that chilled ones.

We can say that this visualization confirms what we pointed out in the data exploration part. Now, let's go further by analyzing the distribution of uptake rates according to the country of origin and treatment.

```
qplot(uptake, data = CO2, color = Type, geom = "density",
      main = "Plants CO2 uptake distribution according to their origin",
      xlab = "CO2 uptake rate (umol/m² sec)")
```




```
qplot(uptake, data = CO2, color = Treatment, geom = "density",
      main = "Plants CO2 uptake distribution according to their treatment",
      xlab = "CO2 uptake rate (umol/m² sec)")
```



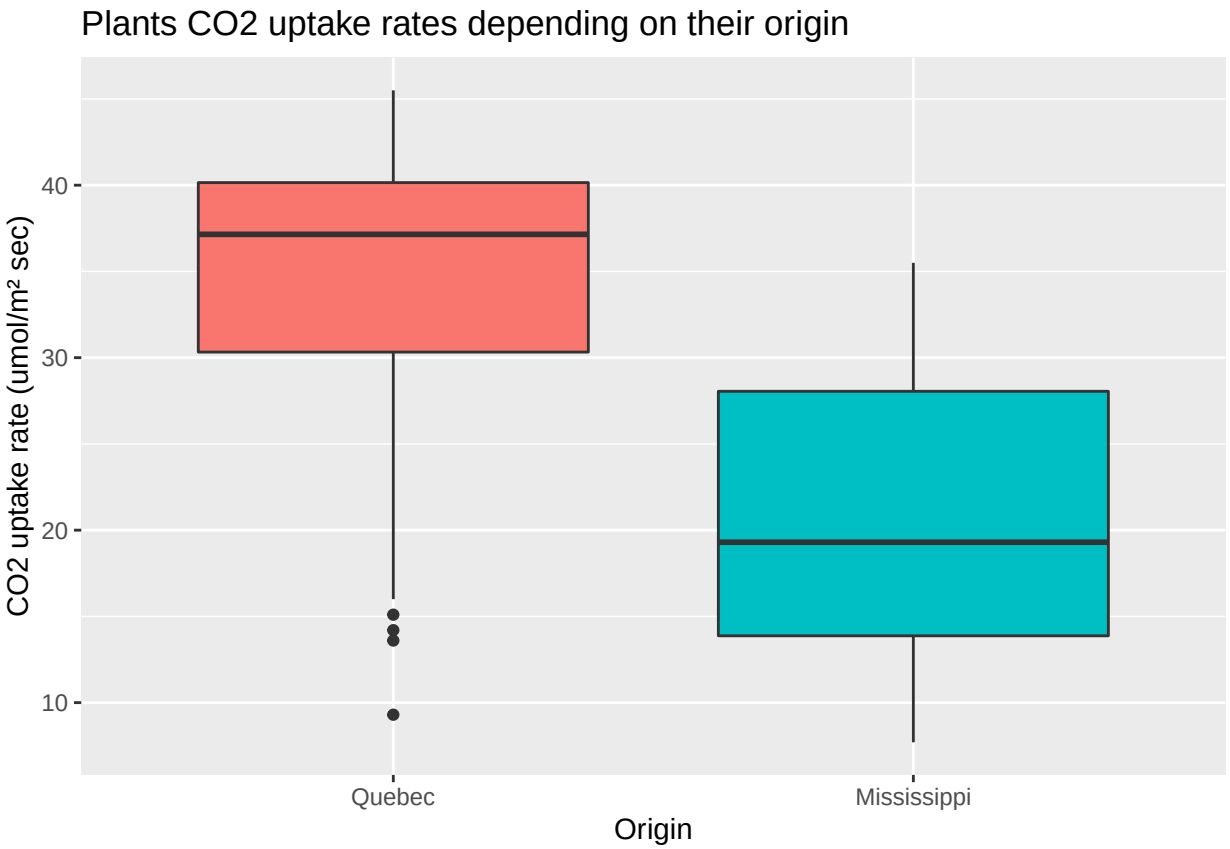

Again, these two plots confirm our hypothesis made above. An important part of Mississippi plants have an uptake rate between 15 and 20 umol/m$^2$ sec whereas for Quebec plants the peak is at about 40 umol/m$^2$ sec. For the treatment, a big part of chilled plants have an uptake rate between 15 and 20 umol/m$^2$ sec whereas for non chilled plants the peak is higher, at about 30 umol/m$^2$ sec.

### Boxplots

Now, we are going to use box plots to visualize the spreading of the uptake rates.

**Boxplot of uptake by origin** We can group our plants by type (Quebec - Mississippi):

```
ggplot(CO2, aes(x = Type, y = uptake, fill = Type)) +
  geom_boxplot(show.legend = FALSE) +
  ggtitle("Plants CO2 uptake rates depending on their origin") +
  xlab("Origin") + ylab("CO2 uptake rate (umol/m² sec)")
```




This figure confirms what we saw in the data exploration part. There is a huge difference in the uptake rates between plants from Quebec and plants from Mississippi.

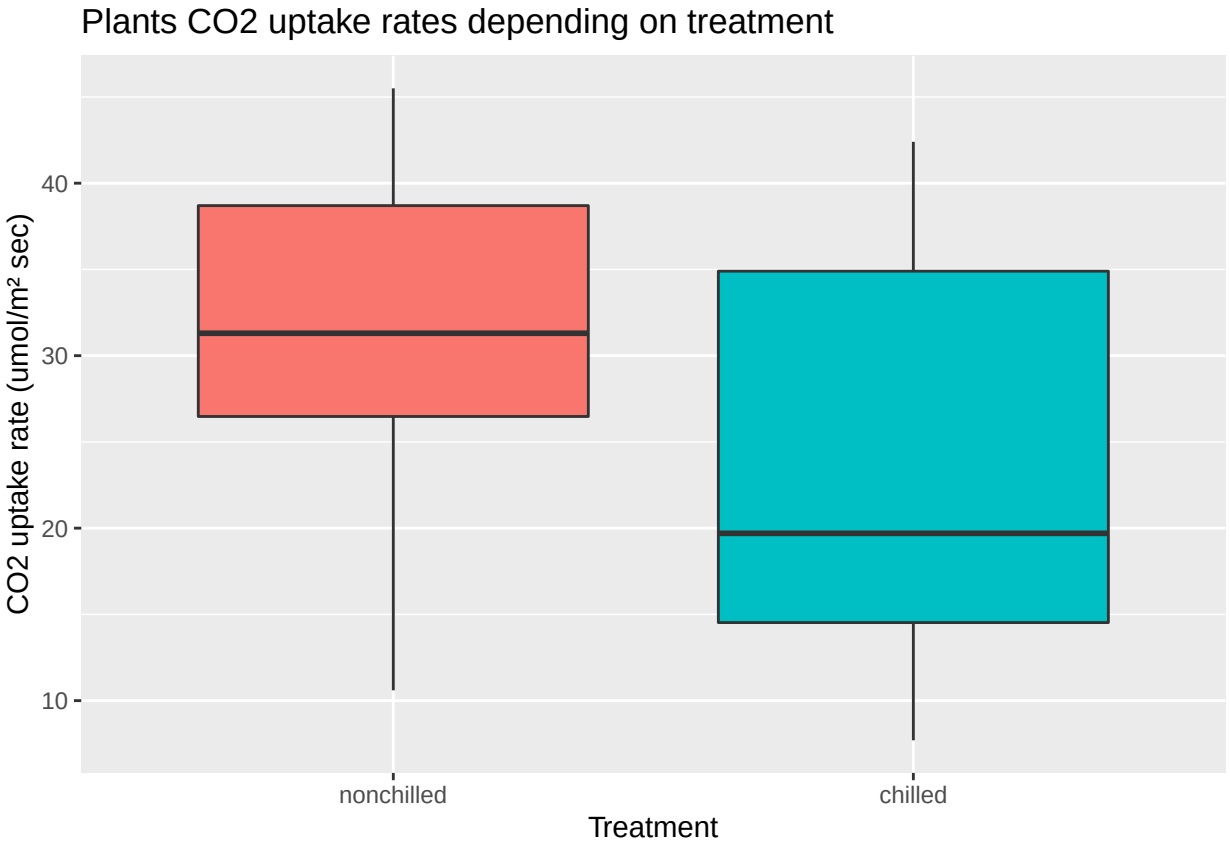
As we know 50% of the data is in the box and by visualizing those box plots, we clearly see that on average, plants from Quebec have higher uptake rates than Mississippi plants. The Quebec box is smaller and higher compared to the Mississippi one.

Half of Quebec plants are over 37 umol/m² sec for the uptake rates while for Mississippi it is around 19 umol/m² sec. In fact 25% of Quebec plants have less than 30 umol/m² sec for their uptake rate whereas more than 75% of Mississippi plants are under 29 umol/m² sec for their uptake rates. This means that the lower quartile of Quebec box plot is greater than the upper quartile of Mississippi box plot.

So doing models, we can suppose that we will easily draw the line between Quebec and Mississippi plants.

**Boxplots of uptake depending on treatment** Now, let's take a look at the importance of the treatment on those plants. Let's compare the uptake rate depending on the treatment.

```
ggplot(CO2, aes(x = Treatment, y = uptake, fill = Treatment)) +
  geom_boxplot(show.legend = FALSE) +
  ggtitle("Plants CO2 uptake rates depending on treatment") +
  xlab("Treatment") + ylab("CO2 uptake rate (umol/m² sec)")
```



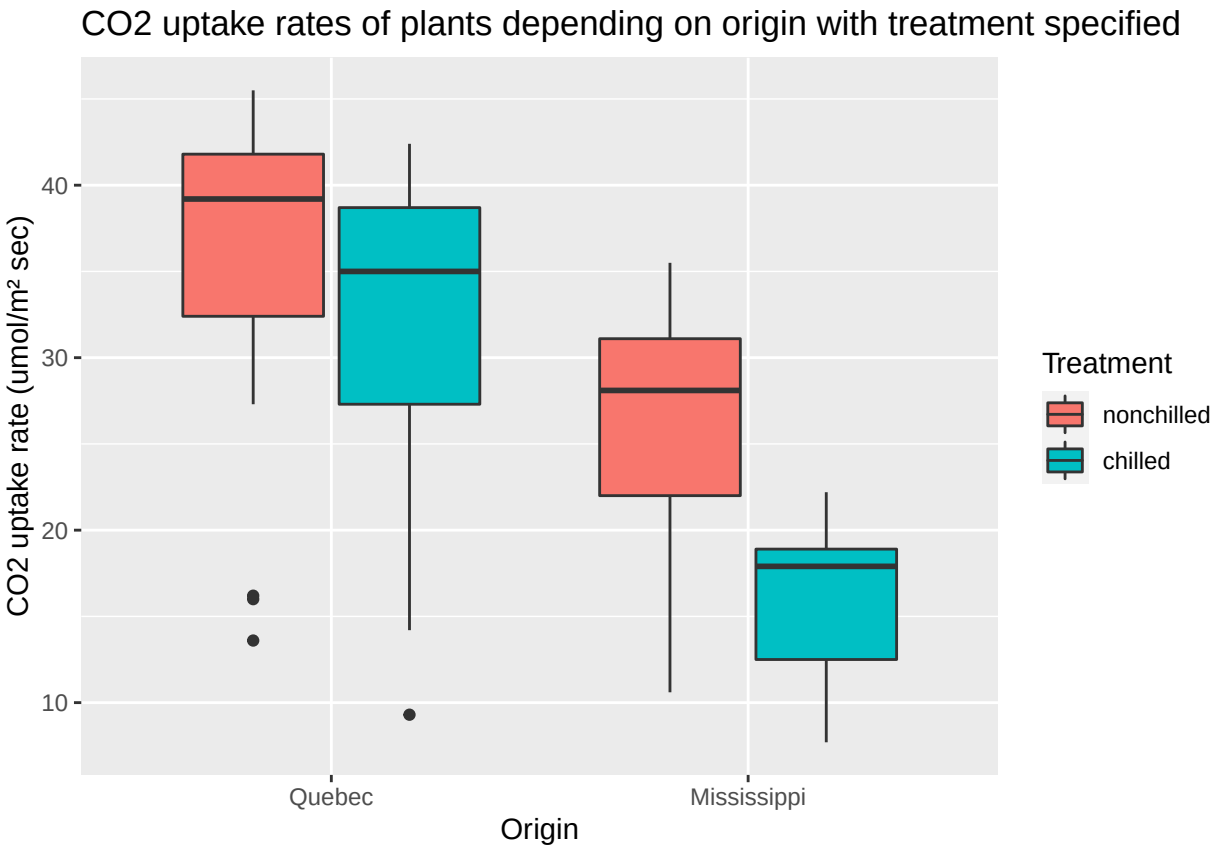

The difference between non chilled and chilled plants is less important than between Quebec and Mississippi data from the previous graph.

The chilled box is lower and more spread out than the non chilled one. We clearly see that half non chilled plants have an uptake rate over 30 umol/m² sec while for chilled it is around 20 umol/m² sec. 25% of non chilled plants have an uptake rate less than 26 umol/m² sec.

The spread out of chilled plants can be explain by the difference of values from Quebec and Mississippi plants that we will see in the next box plot.

**Boxplots of uptake depending on treatment and origin** If we combine our plots, we will obtain the following result:

```
ggplot(CO2, aes(x = Type, y = uptake, fill = Treatment)) + geom_boxplot() +
  ggtitle("CO2 uptake rates of plants depending on origin with treatment specified") +
  xlab("Origin") + ylab("CO2 uptake rate (umol/m² sec)")
```




- For Quebec

In the previous part, we studied two plants from Quebec with different treatments and find out that there was a significant difference in their uptake rates. It appears clearly here as less than 75% of chilled plants are under 39 umol/m² sec while half of non chilled have an uptake rate of more than 39 umol/m² sec.

- For Mississippi

Those differences are even more relevant with Mississippi plants. The chilled box is shorter and lower compared to the non chilled one: the box of non chilled Mississippi plants shows that 25% are upper 22 umol/m² sec whereas for chilled plants, a bit less than 75% of them can reach 19.

Thanks to those box plots, we can clearly see that depending on the type of plants and the treatment, there are some chance to obtain very different uptake rates.

## Testing hypothesis

Let's work on hypothesis regarding the treatment and the type.

### Hypothesis on type

Firstly, we make an assumption on the **type** of the plants. Our null hypothesis is that type will not affect the uptake rates. Our alternative hypothesis is that type of plants may affect the uptake rates. We take a 99% confidence interval:

```r
t.test(CO2$uptake ~ CO2$Type, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  CO2$uptake by CO2$Type
## t = 6.5969, df = 78.533, p-value = 4.451e-09
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##   7.593543 17.725505
## sample estimates:
##      mean in group Quebec mean in group Mississippi
##                  33.54286                  20.88333
```

Looking at the p-value for type, which is 4.451e-9, we see that it is smaller than 0.01. So, we can reject our null hypothesis and conclude that the type may make a difference on the uptake rates.

This very low p-value confirms what we saw in the box plot on type. We can easily draw the line between the two types of plants.

### Hypothesis on treatment

Secondly, we make an assumption on the **treatment** of the plants. Our null hypothesis is that treatment on plant will not affect the uptake rates. Our alternative hypothesis is that treatment on plants may affect the uptake rates. We take a 99% confidence interval:

```r
t.test(CO2$uptake ~ CO2$Treatment, conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  CO2$uptake by CO2$Treatment
## t = 3.0485, df = 80.945, p-value = 0.003107
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##   0.9237369 12.7953107
## sample estimates:
## mean in group nonchilled    mean in group chilled
##                 30.64286                 23.78333
```

Looking at the p-value for treatment, which is 3.107e-3, we see that it is smaller than 0.01.

So, we can reject our null hypothesis and conclude that the treatment may make a difference on the uptake rates.

The p-value for treatment is bigger than the one for type. So we can suppose that we could encounter more mistakes doing prediction on treatment than on type.

### Combining our hypothesis

So we saw that independently, type and treatment may make a difference on the uptake rate.

We can combine our two hypothesis and make an assumption that the relationship between type and treatment will not make a difference on the uptake rates. We combine our hypothesis and use an ANOVA test. T-test method determines whether two groups are statistically different whereas ANOVA allows us to determines on three or more groups, that is why we have to use it here.

```
anova(lm(CO2$uptake ~ CO2$Type * CO2$Treatment))
```

```
## Analysis of Variance Table
##
## Response: CO2$uptake
##                          Df Sum Sq Mean Sq F value    Pr(>F)
## CO2$Type                  1 3365.5  3365.5 52.5086 2.378e-10 ***
## CO2$Treatment             1  988.1   988.1 15.4164 0.0001817 ***
## CO2$Type:CO2$Treatment    1  225.7   225.7  3.5218 0.0642128 .
## Residuals                80 5127.6    64.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both t-test and ANOVA look at the difference in means and the spread of the distributions. But they don't determine the statistical significance in the same way. That is why, our p-value for respectively treatment and type are quite different from the ones we obtain using t-test.

With a confidence interval at 99%: We obtain 2.378e-10 for type and 1.817e-4 for treatment for p-value. Those values are smaller than 0.01, so we can again reject the null hypothesis and conclude that type and treatment may make a difference on the uptake rates.

But looking at the p-value for relationship between treatment and type, we obtain 0.0642 which is bigger than 0.01. So, with an interval at 99%, we can't reject the null hypothesis and we conclude that the relationship between type and treatment did not seriously impact the uptake rates.

If we take an interval at 90%, we can reject the null hypothesis but it will deteriorate the accuracy. We may encounter some errors in our prediction for example for Mississippi chilled and non chilled plants as their uptake rates are very similar.

### Correlation with concentration

In the general plots above, we noticed that the relationship between concentration and uptake was not perfectly linear. So let's take a look at the correlation between them:

```
cor.test(CO2$uptake, CO2$conc)
```

```
##
##  Pearson's product-moment correlation
##
## data:  CO2$uptake and CO2$conc
## t = 5.0245, df = 82, p-value = 2.906e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3022189 0.6336595
## sample estimates:
##       cor
## 0.4851774
```

The correlation value is under 0.5. We can improve it by using a log:

```
cor.test(CO2$uptake, log(CO2$conc))
```

```
##
##  Pearson's product-moment correlation
##
## data:  CO2$uptake and log(CO2$conc)
## t = 6.6591, df = 82, p-value = 2.915e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4330388 0.7158975
## sample estimates:
##       cor
## 0.5924317
```

Using a log, we lower the impact of the concentration compared to the uptake and improve our linear correlation between those two values. It will be useful for our linear regression model.

## Fitting model

In this part, we are going to fit several models (logistic and linear ones) in order to predict the origin, the treatment and the CO2 uptake rate according to several variables.

### Model to predict the origin of a plant

The first model aims to predict the origin of the plant. The predicted value is discrete and has two outcomes: Mississippi or Quebec. In this case, it seems reasonable to fit a **logistic regression**. We tried several combination of predictor variables and we obtained these accuracy values:

- uptake only: 72.6 %
- uptake + Treatment: 83.3 %
- uptake + Treatment + conc: 85.7 %
- uptake + Treatment + log(conc): 90.4 %

We remark that the more predictors we use, the better the accuracy is. Moreover, using the log of the concentration improves the prediction because it has a better correlation than the concentration alone as we have seen in the previous part.

```
type_model <- glm(Type ~ uptake + Treatment + log(conc), family = binomial, data = CO2)
summary(type_model)
```

```
##
## Call:
## glm(formula = Type ~ uptake + Treatment + log(conc), family = binomial,
##     data = CO2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.07518  -0.17019  -0.00239   0.20740   2.26358
```

```
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -21.5346     7.5749  -2.843 0.004470 **
## uptake            -0.6746     0.1766  -3.820 0.000134 ***
## Treatmentchilled  -3.5180     1.1761  -2.991 0.002778 **
## log(conc)          7.1891     2.1559   3.335 0.000854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116.449  on 83  degrees of freedom
## Residual deviance:  35.011  on 80  degrees of freedom
## AIC: 43.011
##
## Number of Fisher Scoring iterations: 7
```

```
type_probs <- predict(type_model, type = "response")
type_pred <- ifelse(type_probs > 0.5, "Mississippi", "Quebec")
table(Predicted = type_pred, Observed = CO2$Type)
```

```
##              Observed
## Predicted     Quebec Mississippi
##   Mississippi      5          39
##   Quebec          37           3
```

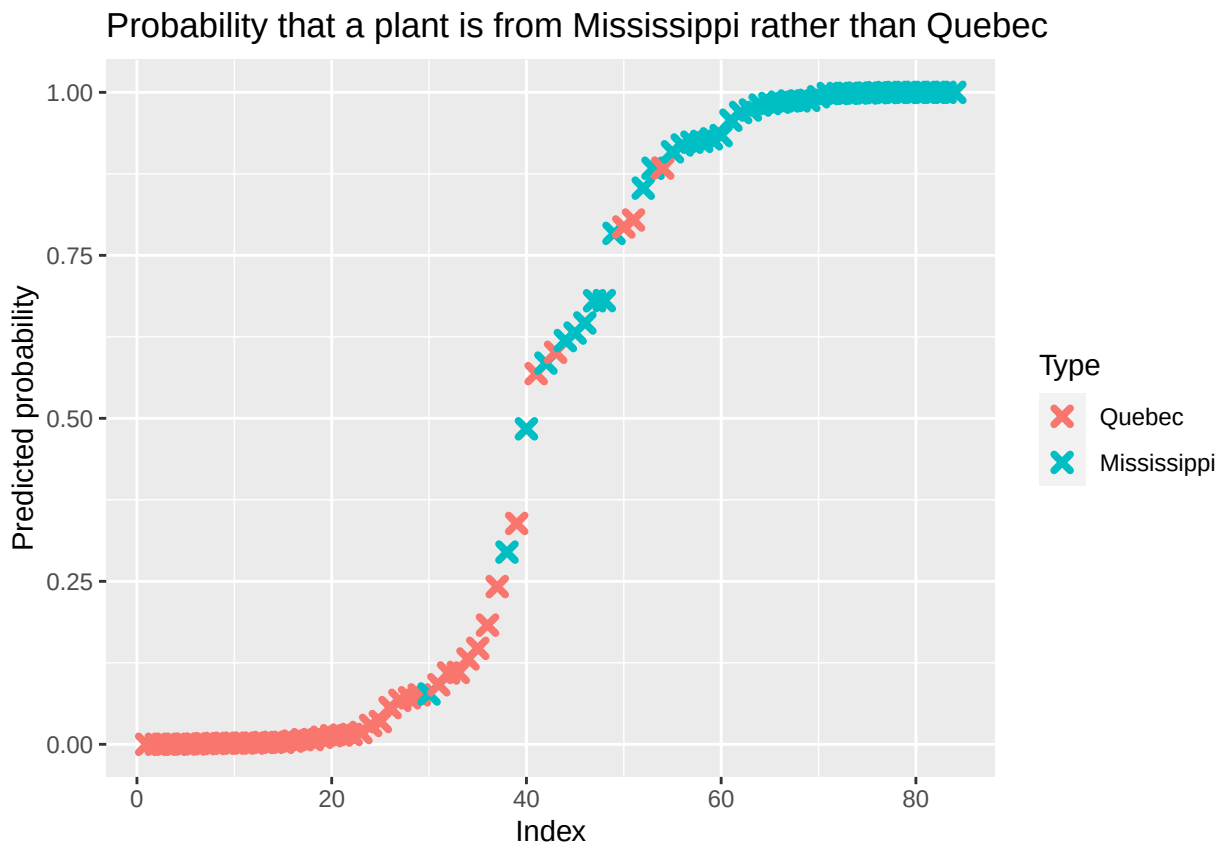
```
mean(type_pred == CO2$Type)
```

```
## [1] 0.9047619
```

Here, we chose the combination offering the best accuracy. We can see that the model is wrong for only 8 observations out of 84. The following graph allows us to visualize the model and its errors:

```
type_pred_data <- data.frame(prob = type_model$fitted.values, Type = CO2$Type)
type_pred_data <- type_pred_data[order(type_pred_data$prob), ]
type_pred_data$rank <- 1:nrow(type_pred_data)

ggplot(data = type_pred_data, aes(x = rank, y = prob)) +
  geom_point(aes(color = Type), shape = 4, stroke = 2) +
  ggtitle("Probability that a plant is from Mississippi rather than Quebec") +
  xlab("Index") +
  ylab("Predicted probability")
```




**Model to predict the treatment of a plant**

For this model, we use the same reasoning as for the previous one, except that we want to predict the treatment in this case. The two outcomes are chilled or nonchilled. Below are the accuracy obtained for several combinations:

- uptake only: 67.8 %
- uptake + Type: 70.2 %
- uptake + Type + conc: 76.1 %
- uptake + Type + log(conc): 79.7 %

```
treat_model <- glm(Treatment ~ uptake + Type + log(conc), family = binomial, data = CO2)
summary(treat_model)
```

```
##
## Call:
## glm(formula = Treatment ~ uptake + Type + log(conc), family = binomial,
##     data = CO2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.36897  -0.75100  -0.05638   0.78298   2.09323
##
## Coefficients:
```

```
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -4.94305    2.50522  -1.973 0.048484 *
## uptake          -0.29843    0.06758  -4.416 1.01e-05 ***
## TypeMississippi -3.72143    1.00260  -3.712 0.000206 ***
## log(conc)        2.57919    0.69459   3.713 0.000205 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 116.449  on 83  degrees of freedom
## Residual deviance:  82.137  on 80  degrees of freedom
## AIC: 90.137
##
## Number of Fisher Scoring iterations: 5
```

```
treat_probs <- predict(treat_model, type = "response")
treat_pred <- ifelse(treat_probs > 0.5, "chilled", "nonchilled")
table(Predicted = treat_pred, Observed = CO2$Treatment)
```

```
##             Observed
## Predicted    nonchilled chilled
##   chilled             8      33
##   nonchilled         34       9
```

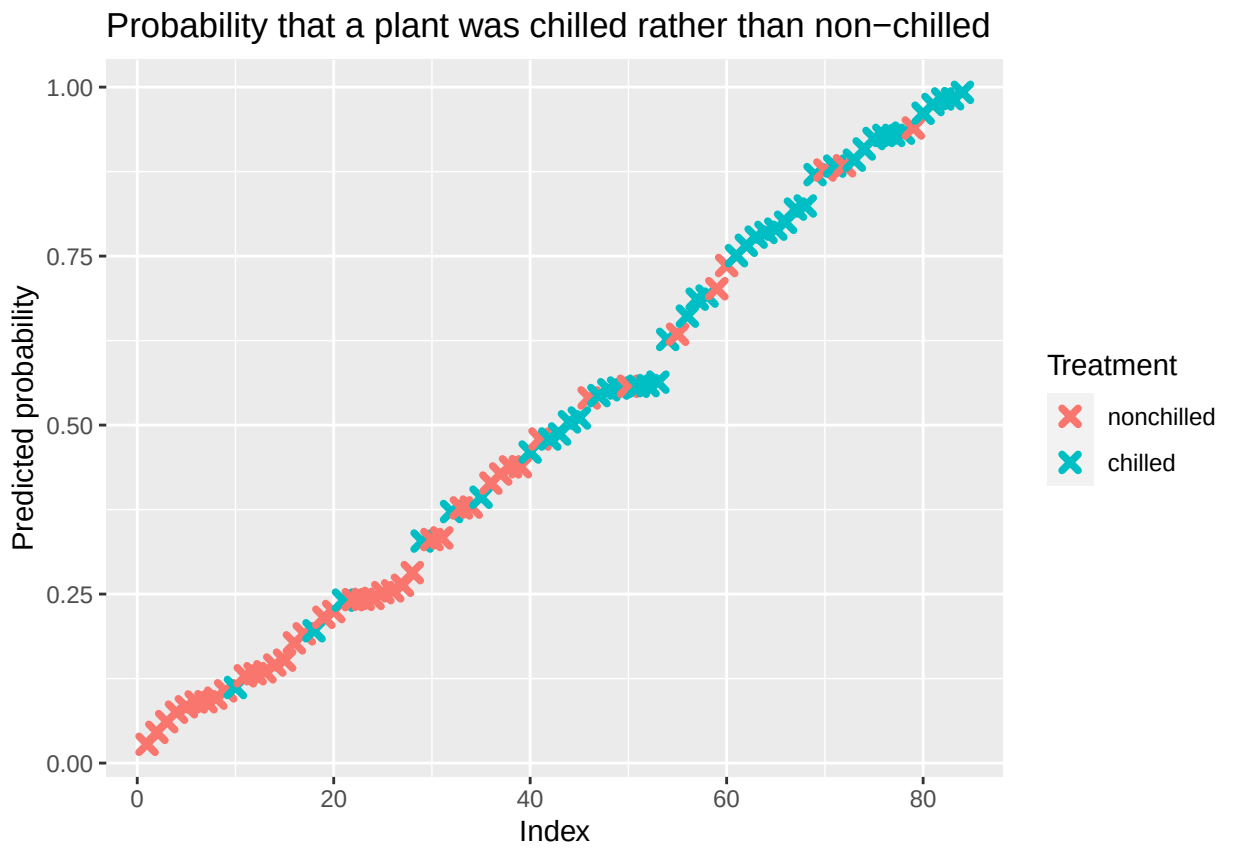
```
mean(treat_pred == CO2$Treatment)
```

```
## [1] 0.797619
```

We can see that the accuracy is worse than the one of the previous model (79.7 % against 90.4 %). Indeed, the treatment is more difficult to predict than the origin because the difference of the means is smaller (see data visualization part). The model is wrong for 17 observations out of 84. This graph helps to visualize the model and errors made:

```
treat_pred_data <- data.frame(prob = treat_model$fitted.values, Treatment = CO2$Treatment)
treat_pred_data <- treat_pred_data[order(treat_pred_data$prob), ]
treat_pred_data$rank <- 1:nrow(treat_pred_data)

ggplot(data = treat_pred_data, aes(x = rank, y = prob)) +
  geom_point(aes(color = Treatment), shape = 4, stroke = 2) +
  ggtitle("Probability that a plant was chilled rather than non-chilled") +
  xlab("Index") +
  ylab("Predicted probability")
```




**Model to predict the CO2 uptake rate of a plant**

Finally, we fit a **linear regression** model to predict the CO2 uptake. The outcome is thus a continuous value. As before, we can try several combinations of predictors and compare the $R^2$ value this time:

- conc only: 0.235
- log(conc) only: 0.351
- log(conc) + Type: 0.697
- log(conc) + Type + Treatment: 0.799

```
uptake_model <- lm(uptake ~ log(conc) + Type + Treatment, data = CO2)
summary(uptake_model)
```

```
##
## Call:
## lm(formula = uptake ~ log(conc) + Type + Treatment, data = CO2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.6369  -2.6968   0.5398   3.2446  10.5138
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -12.3976     4.2750  -2.900  0.00482 **
```

```
## log(conc)          8.4839     0.7169  11.833  < 2e-16 ***
## TypeMississippi  -12.6595     1.0764 -11.761  < 2e-16 ***
## Treatmentchilled  -6.8595     1.0764  -6.373 1.11e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.933 on 80 degrees of freedom
## Multiple R-squared:  0.7995, Adjusted R-squared:  0.792
## F-statistic: 106.3 on 3 and 80 DF,  p-value: < 2.2e-16
```

We see that our fitted model is correct but not very good with a $R^2$ value of about 0.8. To get this value closer to 1, we would need more observations or other predictor variables to better predict the uptake.

To illustrate this model, we can input some test values and compare with the predictions:

```
v1 = c(95, 250, 250, 250, 500, 1000)
v2 = c("Quebec", "Quebec", "Mississippi", "Mississippi", "Mississippi", "Quebec")
v3 = c("chilled", "nonchilled", "chilled", "nonchilled", "chilled", "nonchilled")
v4 = c(12.8, 37.4, 16.1, 27.5, 16.6, 43.1)
# v4 values obtained with summarize(group_by(CO2, Type, Treatment, conc), mean(uptake))

testData <- data.frame(conc = v1, Type = v2, Treatment = v3, uptake = v4)
testData["predicted_uptake"] <- predict(uptake_model, testData)
testData
```

```
##   conc        Type  Treatment uptake predicted_uptake
## 1   95      Quebec    chilled   12.8         19.37736
## 2  250      Quebec nonchilled   37.4         34.44575
## 3  250 Mississippi    chilled   16.1         14.92670
## 4  250 Mississippi nonchilled   27.5         21.78622
## 5  500 Mississippi    chilled   16.6         20.80728
## 6 1000      Quebec nonchilled   43.1         46.20690
```

The actual uptake values and the predicted ones are quite similar even if the prediction is far from being perfect. Our model and the dataset are not optimal so we are satisfied with these results.

## Conclusion

In our analysis, we saw that type and treatment can both influence the CO2 uptake rates of plants. This first approach on CO2 data set gave us the ability to predict treatment and type depending on uptake rates and inversely uptake rates depending on treatment and type.

The most precise **logistic model** on type confirms what we saw in the data visualization part where box plot from Quebec and Mississippi had huge differences in their quartiles and median. This model can easily predict the type of plants based on CO2 uptake rates, CO2 concentration and treatment. The second one is the **linear regression model** which predicts uptake rates based on CO2 concentration, type and treatment. The results are not as precise as we want them to be.

In order to improve our models, we would need more observations or predictor variables.