Stomatal conductance of *Sorghum bicolor*

This project stems from a larger study of *Sorghum bicolor*'s hydraulics under Salt stress on the short term, this analysis is focused on the stomatal conductance, one of the many variables measured in the study. The stomata are small pores found in the epidermis of some plant leaves that allow the exit of water vapor and the entry of CO_2 (gas exchange), in most plants they are closed during the night and open during the day, when photosynthesis is carried out. In this case, the stomatal conductance is our response variable of interest, measured in mmol of water vapor per m^2 and second, it basically represents how many of these pores are open and how many are closed.

Our hypothesis for this work was that the stomatal conductance decreases as the medium Salt concentration increases and time passes (1, 4 and 24 hours were used in this study but for a shorter analysis only the 1 and 24 hours will be shown). And also, that this decline differs for each genotype, as seen in previous studies done under water stress.

First of all, we import the necessary libraries for the analysis and create a tiny function to calculate the standard error of the mean (S.E.M.).

```
library("readxl")
library("dplyr")
library("ggplot2")
library("broom")
se <- function(x) sd(x)/sqrt(length(x))
```

Then, we import the dataset and use the function `str()` to take a quick look at the dataset.

```
df <- read_excel("Data_gs.xlsx")
str(df)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    262 obs. of  6 variables:
##  $ Set     : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ Genotype: chr  "9530" "9530" "B2" "B2" ...
##  $ Plant   : num  4 5 4 5 4 5 4 5 4 5 ...
##  $ Time    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Salt    : num  0 0 0 0 100 100 100 100 200 200 ...
##  $ gs      : num  40 48.4 38.9 67.2 51.2 31.9 49.2 33.8 34.9 16.8 ...
```

As we can see, some columns were not imported in the most convenient type (such as the Time explanatory variable as numeric), we just convert them to factors accordingly:

```
df[1:5] <- lapply(df[1:5], as.factor)
```

Exploratory data analysis

Missing values, outliers and general description

We use the `summary()` built-in function to obtain some general statistic descriptive values for each column:

```
summary(df)
```

```
##  Set    Genotype       Plant     Time      Salt          gs
##  2:40   9530:130   19     : 17   1 :112   0  :65   Min.   : 5.50
##  3:41   B2  :132   5      : 16   24:150   100:63   1st Qu.:22.30
##  4:40              9      : 16            200:66   Median :28.20
##  5:40              13     : 16            300:68   Mean   :29.94
```

```
##  6:35              14      : 16                    3rd Qu.:36.80
##  7:33              8       : 15                    Max.   :67.20
##  8:33              (Other):166                     NA's   :7
```

As we can see, the data is considerably unbalanced (for example: time, 112 values for 1 hour and 150 for 24 hours). In our response variable's column (gs), we detect 13 missing values and considering the median is approximately the same as the mean, we can assume outliers won't be a big deal (although we'll take care of them shortly).

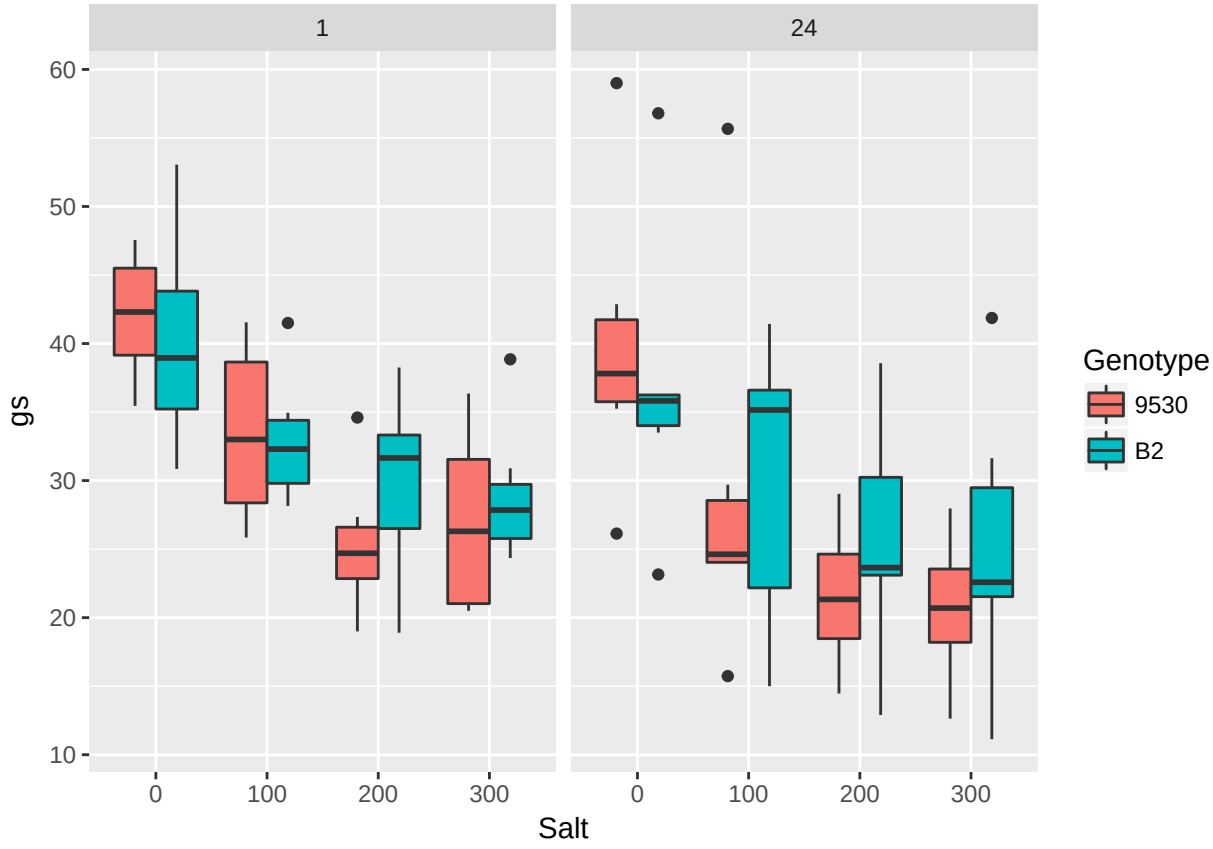
We'll take the mean of the response variable for each set (the Set variable refers to each group of seeds sown in the same day then taken to hydroponics on another same day and finally measured at the same time), since the number of replicates was not the same for each combination of factors (although the number of independent experiments was).

```
df <- df %>% group_by(Set, Genotype, Time, Salt) %>% summarize(gs = mean(gs))
```

Since this is a dataset coming from experiments, it wouldn't be correct to fill NAs with the mean because considering the small sample for each genotype, Salt, time combination as it would bias our results too much. So, we shall omit them and then use a boxplot to check outliers:

```
df = na.omit(df)

ggplot(df, aes(x = Salt, y = gs, fill = Genotype))+
  geom_boxplot()+
  facet_wrap(~Time)
```

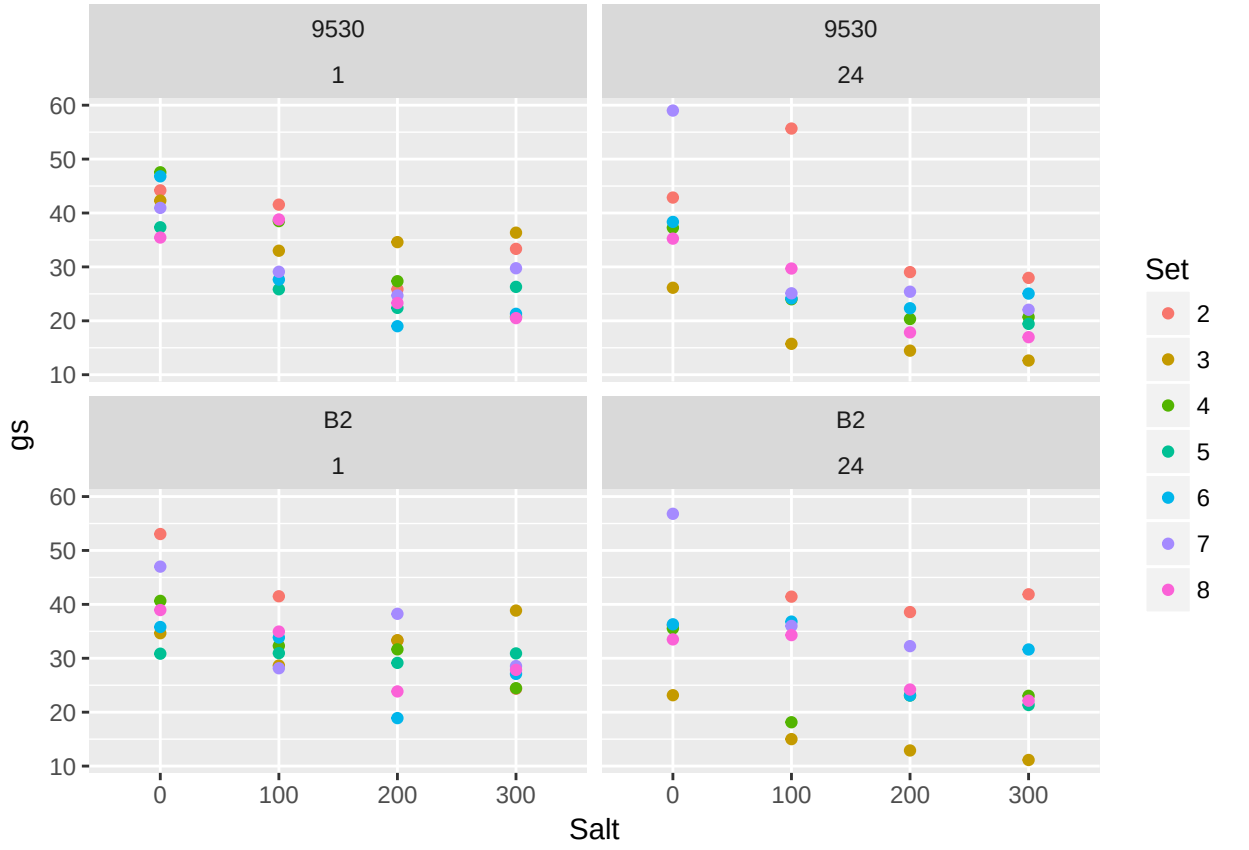


For both time treatments, we see a declining trend until 200 mM where the values seem to be equal to those seen at 300 mM. And also, the presence of outliers seems to be higher for the 24 hours treatment.

We'll go one step further and use the *Set* variable in the dataset to determine if a certain Set is more prone

to outliers:

```
ggplot(df, aes(x = Salt, y = gs, col = Set))+
  geom_point()+
  facet_wrap(Genotype~Time)
```



First, we take a look at the 24 hour treatment data points. Compared to the previous plot we can see that the Sets 2 and 7 are where the extreme values at 100 and 200 mM stem from. Considering this we won't remove them since it would have a considerably impact in our number of samples that isn't very high (7 independent experiments). For the 1 hour treatment we see that the variability is not a big deal and stems naturally from the process observed.

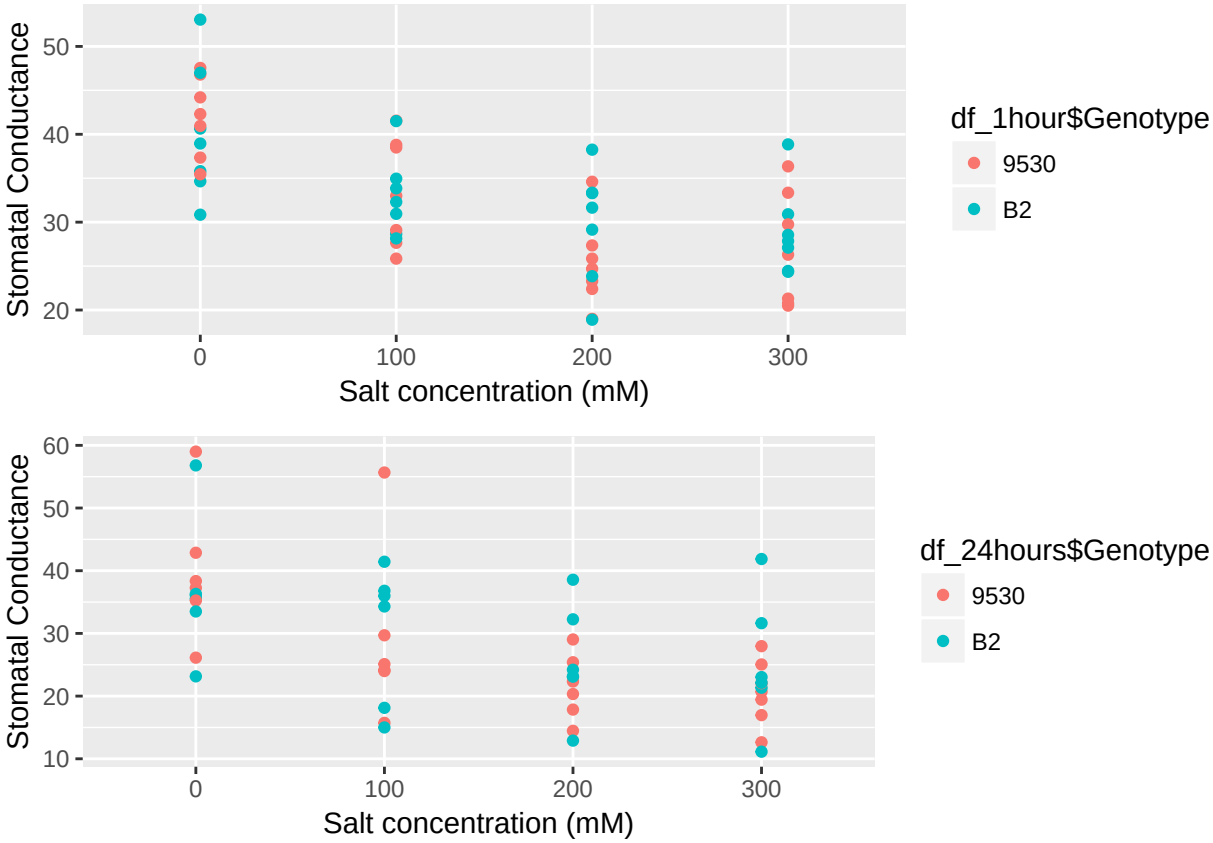
Time and Salt treatment effect

Now we will play a closer look to each time treatment:

Using the `qplot` function from the ggplot2 package we can take a look at the data points when plants were subjected to 1 hour and 24 hours of Salt treatment:

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```



From both plots we can see that there is a declining trend for the stomatal conductance as Salt increases (a common mechanism in most plants: closing stomata to avoid water loss by reducing the transpiration rate) but in neither case seems to be a difference between genotypes (or an interaction between the Genotype and Salt variables, in other terms).

Statistical Analysis and Hypothesis testing

Considering the variable type we are studying (a quantitative one) we shall evaluate the assumptions made while calculating linear models before modelling the data to determine if there are any statistical significant differences between treatments.

First we will calculate the full model with interactions but no random effects with the `lm()` function:

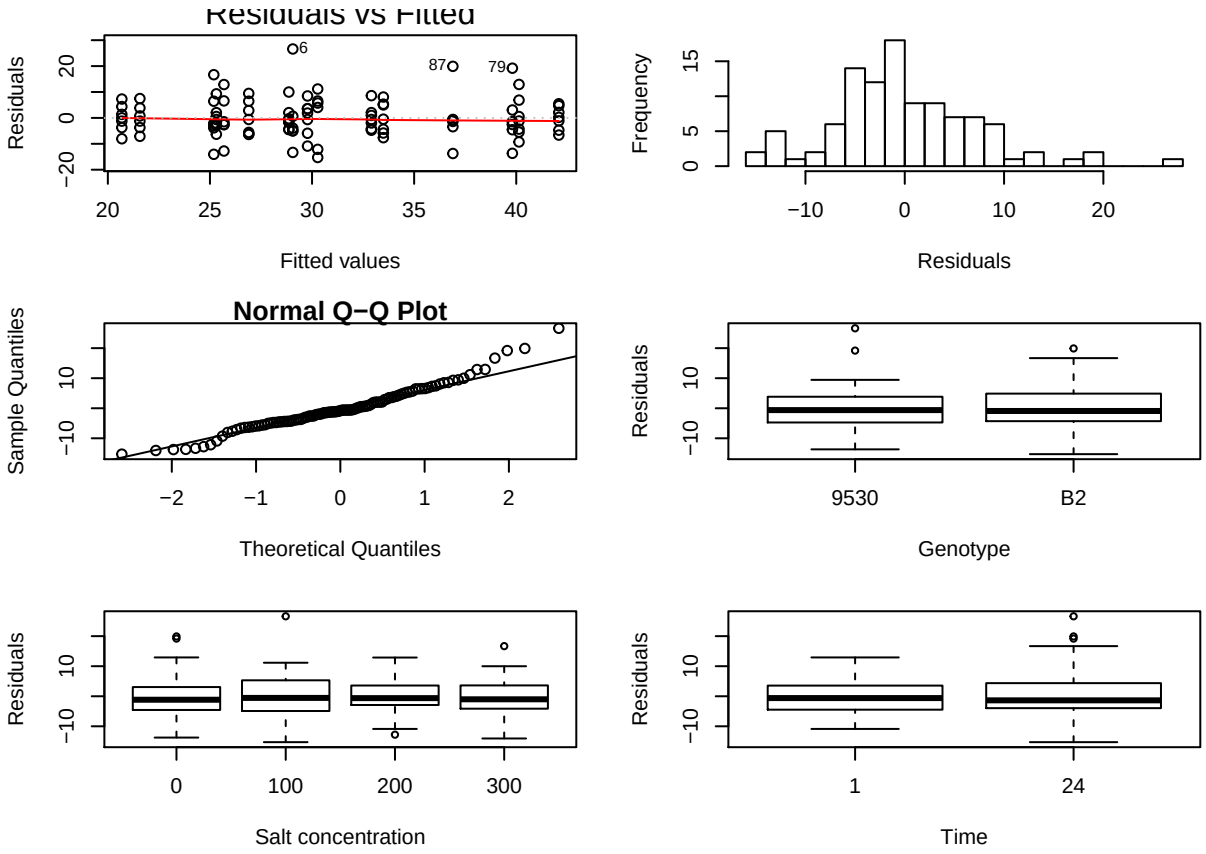
```
m0 <- lm(gs ~ Genotype * Salt * Time, data = df)
```

We then check the assumptions and validate the model:

```
op <- par(mfrow = c(3, 2), mar = c(5, 4, 1, 2))
#Homogeneity of variances
plot(m0, add.smooth = TRUE, which = 1)
#Normality of residuals
e <- resid(m0)
hist(e, xlab = "Residuals", main = "", breaks = 16)
qqnorm(e)
qqline(e)

#Independent Observations
```

```
plot(df$Genotype, e, xlab = "Genotype",
     ylab = "Residuals")
plot(df$Salt, e, xlab = "Salt concentration",
     ylab = "Residuals")
plot(df$Time, e, xlab = "Time",
     ylab = "Residuals")
```




```
par(op)
par(mar = c(0, 0, 0, 0))
```

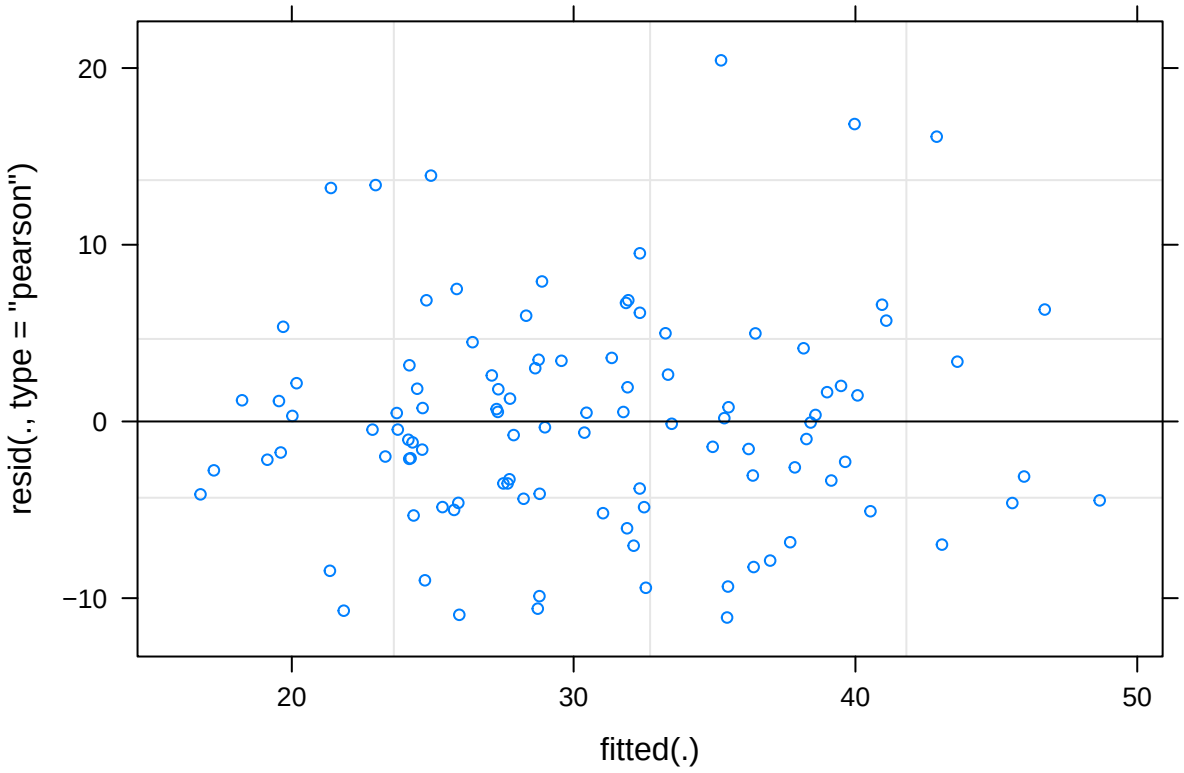
From the first plot at the upper left (residuals vs fitted) we see that errors have a random distribution around the 0 line, that the variances for te treatments seem equal and that our assumption of a linear relationship seems reasonable. The second plot to the upper right and the QQPlot tells us that although there is certain normality in the data, the values depart from it to the higher and lower quantiles, considering the linear regression is robust to the lack of normality, we will choose to proceed despite this (For a discussion on how to interpret QQplots, this discussion results very useful).

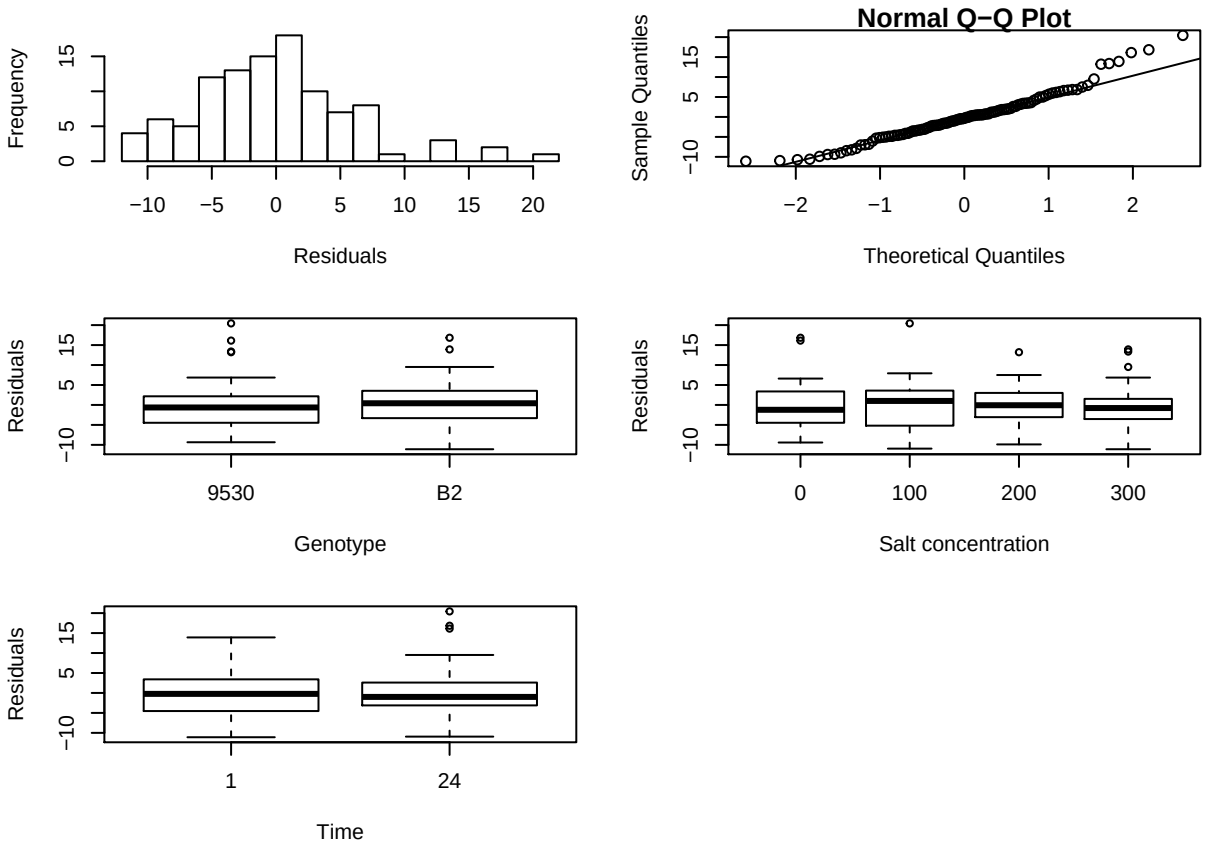
Finally, from the three last plots we see that the variances for each treatment are very similar, in other words, no heteroscedasticity is detected.

After this, we wanted to see how a model with the Set variable as a random effects one behaved, we import the library **lme4** and use the `lmer()` function:

```
library(lme4)
m1 <- lmer(gs ~ Genotype * Salt * Time + (1 | Set), data = df, REML = FALSE)
```

As before, we validate the model:


resid(., type = "pearson")
20
10
0
−10
20
30
40
50
fitted(.)



It can be seen that the assumptions made by fitting a linear model (independence between errors, homoscedasticity and normality of the error distribution) look better with the later model. We'll also check the AIC criterion for both models with the `AIC()` function (it determines how well the data supports each model, taking into account both the sum of squares and the number of parameters estimated):

```
##    df      AIC
## m0 17 749.3499
## m1 18 731.9468
```

We also see that the relative quality measured by this criterion is higher for the model that considers random effects, even though it estimates more parameters.

Now that we have a model we can finally start considering the statistical significance of the results observed to answer our intiial questions: Does the stomatal conductance change with increasing salt concentration? Does this change differ between genotypes? Does it change depending of the salt treatment time?

Tipically, we'd use `summary(model)` to assess the p-values for each coefficient and then calculate comparisons between groups. Since we are dealing with a mixed effects models, we will use the Likelihood Ratio Test. What it basically does, is compare the model *without* the factor we're interested with the model *including* this factor (for a more detailed explanation, see this link, an article written by Bodo Winter, from the University of California, 2013).

First, we'll test the triple interaction:

```
m2 <- lmer(gs ~ Genotype * Salt + Time + (1 | Set), data = df)
anova(m2, m1)
```

```
## Data: df
## Models:
```

```
## m2: gs ~ Genotype * Salt + Time + (1 | Set)
## m1: gs ~ Genotype * Salt * Time + (1 | Set)
##    Df    AIC    BIC  logLik deviance Chisq Chi Df Pr(>Chisq)
## m2 11 718.76 747.95 -348.38   696.76
## m1 18 731.95 779.72 -347.97   695.95 0.811      7     0.9973
```

We see that time, salt and genotype are not interdependent (p-value >>> 0.05), so we proceed with the interaction between salt and genotype:

```
## Data: df
## Models:
## m3: gs ~ Genotype + Salt + Time + (1 | Set)
## m2: gs ~ Genotype * Salt + Time + (1 | Set)
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m3  8 717.36 738.60 -350.68   701.36
## m2 11 718.76 747.95 -348.38   696.76 4.6056      3     0.2031
```

As it is not significant either, we proceed with each of the variables separately:

```
## Data: df
## Models:
## m4: gs ~ Genotype + Salt + (1 | Set)
## m3: gs ~ Genotype + Salt + Time + (1 | Set)
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m4  7 725.02 743.59 -355.51   711.02
## m3  8 717.36 738.60 -350.68   701.36 9.6522      1   0.001891 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Data: df
## Models:
## m5: gs ~ Genotype + Time + (1 | Set)
## m3: gs ~ Genotype + Salt + Time + (1 | Set)
##    Df    AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m5  5 772.03 785.3 -381.02   762.03
## m3  8 717.36 738.6 -350.68   701.36 60.668      3  4.231e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Data: df
## Models:
## m6: gs ~ Salt + Time + (1 | Set)
## m3: gs ~ Genotype + Salt + Time + (1 | Set)
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m6  7 716.65 735.22 -351.32   702.65
## m3  8 717.36 738.60 -350.68   701.36 1.2822      1     0.2575
```

From this output we conclude that time and salt had significant effects over the stomatal conductance but the genotype of the plants didn't.

So, the final model used will be the number 6 which considers salt, time and random effects associated to the set from which each plant came from. To detect the differences between the levels for the salt variable, we use the `glht()` function from the multcomp package that allows us to perform multiple comparisons with the Tukey method.

```
library("multcomp")
comparisons = glht(m6,linfct=mcp(Salt="Tukey"))
summary(comparisons)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lmer(formula = gs ~ Salt + Time + (1 | Set), data = df)
##
## Linear Hypotheses:
##                   Estimate Std. Error z value Pr(>|z|)
## 100 - 0 == 0     -8.26859    1.83433  -4.508  < 0.001 ***
## 200 - 0 == 0    -14.10064    1.83433  -7.687  < 0.001 ***
## 300 - 0 == 0    -14.02032    1.82410  -7.686  < 0.001 ***
## 200 - 100 == 0   -5.83205    1.83433  -3.179  0.00801 **
## 300 - 100 == 0   -5.75173    1.82410  -3.153  0.00863 **
## 300 - 200 == 0    0.08032    1.82410   0.044  0.99997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

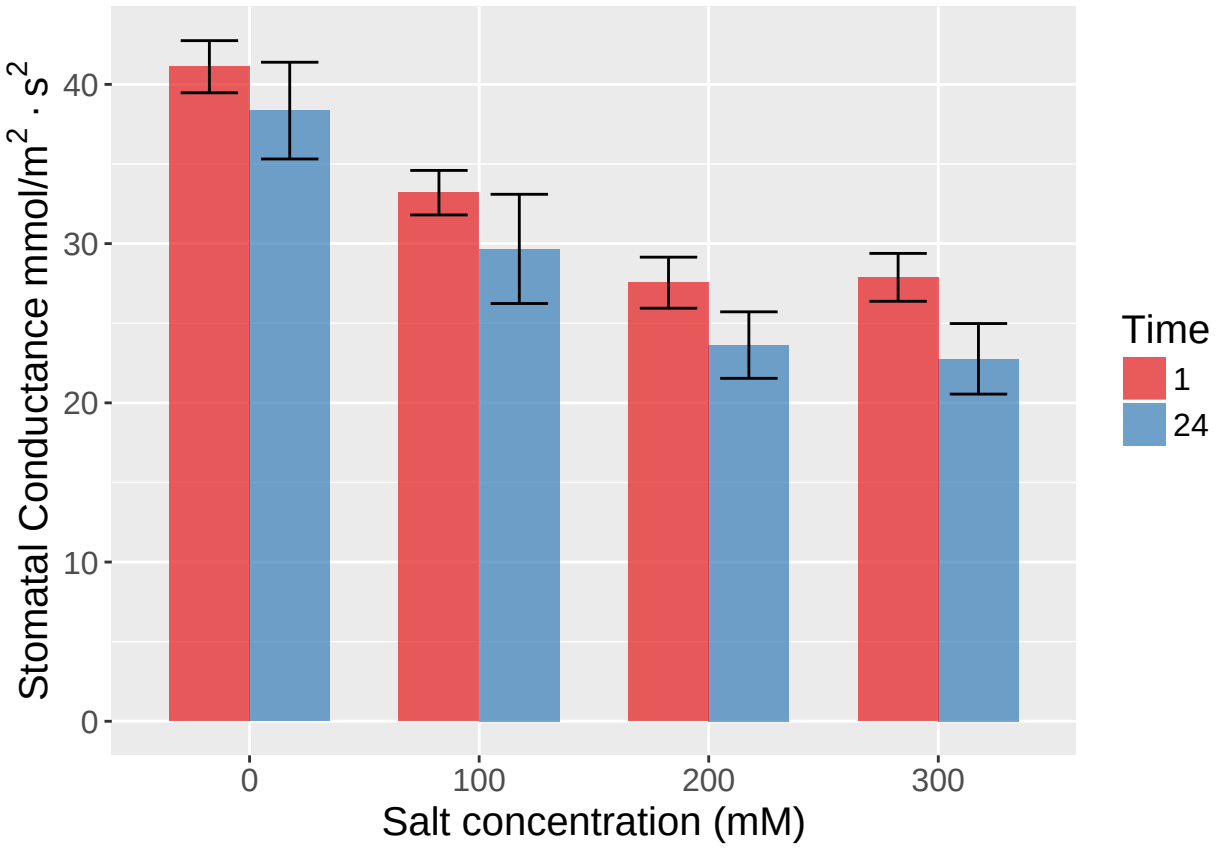
Finally, this tells us that the stomatal conductance under 100 mM of NaCl was significantly different from both the control (0 mM) and the 200-300 mM treatments, and that these two resulted equal to each other and diferent from the control as well. As for the time, we don't need to perform comparisons since there are only two levels on that factor, we know for sure that they're significantly different between them.

Final data visualization

We'll wrap up this analysis with two plots. The first one, showing a barplot of the response variable on the y axis and the salt concetration on the x axis, with Time as a grouping color. The error bars indicate the 95 % confidence intervals for the mean, the code used to generate it is shown below:

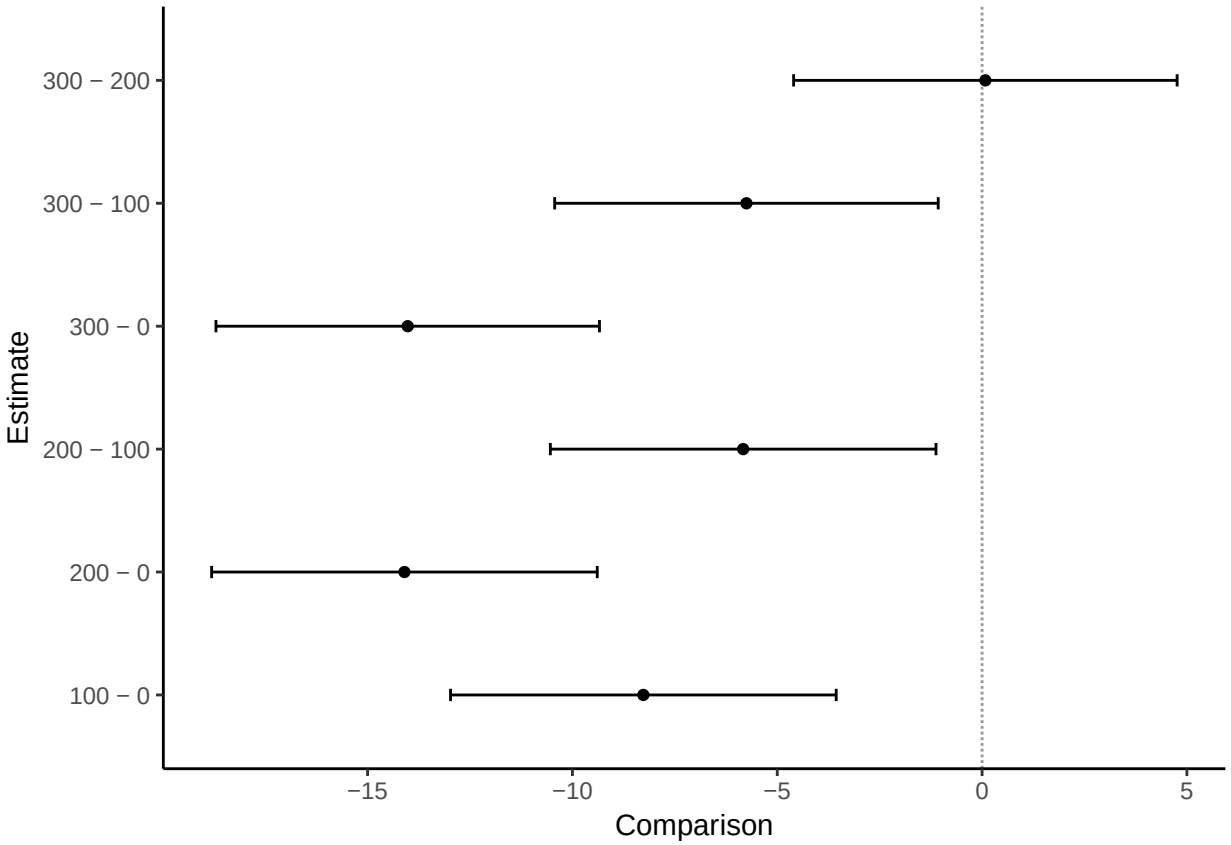
```
summary <- df %>% #Necessary summary data frame for barplot
  group_by(Salt, Time) %>%
  summarize(Mean = mean(gs),
            sem = se(gs),
            UppLim = Mean + sem,
            LimInf = Mean - sem)

ggplot(summary, aes(x = Salt, y = Mean, fill= Time))+
  geom_bar(stat = "identity", width = 0.7, position = "dodge", alpha = 0.7)+
  geom_errorbar(aes(x = Salt, ymin = UppLim, ymax = LimInf),
                width = 0.50, position = position_dodge(width = .7))+
  labs(x = "Salt concentration (mM)", y = expression("Stomatal Conductance"~"mmol/m"^2~"·"~"s"^2))+
  scale_fill_brewer(palette="Set1")+
  theme(text = element_text(size = 15))
```



And then, the confidence intervals for the differences between each salt treatment:

```
conf_int = confint(comparisons)
conf_int %>%
  tidy %>%
  ggplot(aes(lhs, y=estimate, ymin=conf.low, ymax=conf.high)) +
    geom_hline(yintercept=0, linetype="11", colour="grey60") +
    geom_errorbar(width=0.1) +
    geom_point() +
    coord_flip() +
    theme_classic()+
    labs(x = "Estimate", y = "Comparison")
```



We conclude from our study that increasing salt concentration in an hydroponic medium causes a reduction in the stomatal conductance in *Sorghum bicolor*, i.e. closing of stomata. This was expected since a saltier medium reduces the plant's capacity to absorb water from the soil, so one strategy is to close stomata in order to reduce the transpiration.