# Linear and Multiple Regressions

*Florencia Zúñiga*

*7/9/2019*

In this final part of the specific analysis, we focus on the factors related to drug sensitivity. We aim to discover whether doubling time or the copynumber would be better to predict how much of vorinostat is neccessary to cause a 50% growth inhibition.

*Disclaimers: (1) As we only results in terms of proprotion of values to each other, we will not change the log values of the files used in this exploration. (2) For this exploration to be completed, the biomarkers obtained at the beginning of the specific analysis are used. (3) The code of plots that require the same code of a previous visualization with the exception of small details are not included to avoid redundancy. (4) As the values obtained in the summaries of the linear regression change each time the document is knitted, the values annotated here are an approximation of what expect after running the code several times and observing several values for e.g. Multiple R-squared and the p-value.*

## Table of contents

---

# 1. LOADING DATA

---

**Loading packages**

```
library(readr)
library(rstudioapi)
library(lattice)
library(e1071)
library(ggplot2)
library(scatterplot3d)
library(car)
library(scatterD3)
library(rgl)
library(dplyr)
```

Reading the data

```
# Reading the data
Untreated <- readRDS(paste0(wd,"/data/NCI_TPW_gep_untreated.rds"))
Treated <- readRDS(paste0(wd,"/data/NCI_TPW_gep_treated.rds"))

Metadata = read.table(paste0(wd,"/data/NCI_TPW_metadata.tsv"),
                      header = TRUE, sep ="\t", stringsAsFactors = TRUE)

Sensitivity <- readRDS(paste0(wd,"/data/NegLogGI50.rds"))


Basal <- readRDS(paste0(wd,"/data/CCLE_basalexpression.rds"))
Copynumber <- readRDS(paste0(wd,"/data/CCLE_copynumber.rds"))
Mutations <- readRDS(paste0(wd,"/data/CCLE_mutations.rds"))
```

```
Metadata = read.table(paste0(wd,"/data/NCI_TPW_metadata.tsv"), header = TRUE, sep ="\t", stringsAsFacto

Cellline_annotation = read.table(paste0(wd,"/data/cellline_annotation.tsv"),
                                 header = TRUE, sep ="\t", stringsAsFactors = TRUE)
Drug_annotation = read.table(paste0(wd,"/data/drug_annotation.tsv"),
                             header = TRUE, sep ="\t", stringsAsFactors = TRUE)

# Transforming the data

Treated <- as.data.frame(Treated)
Untreated <- as.data.frame(Untreated)
Sensitivity<- as.data.frame(Sensitivity)
```

Data normalization

```
Untreated_norm <- apply(Untreated, 2, function(x){
  (x - mean(x)) / sd(x)
 })


Treated_norm <- apply(Treated, 2, function(x){
  (x - mean(x)) / sd(x)
 })


FC <- Treated - Untreated
FC_norm <- apply(FC, 2, function(x){
  (x - mean(x)) / sd(x)
 })
```

---

## 2. Preparation of the data: Tables

---

```
#### (1) Table 1: Selection of 100 Biomarkers in copynumber
BM_copynumber = Copynumber[ which(row.names(Copynumber)
                                  %in% rownames(biomarkers_FC_values100)), ]

BM_Copynumber_meancol= colMeans(BM_copynumber)

CN = as.data.frame(BM_Copynumber_meancol)

#### (2) Table 2: All genes in copynumber
CN_meancol= colMeans(Copynumber)

CN_all = as.data.frame(CN_meancol)

#### (3) Table 3: Selection of Doubling time from cellline_annotation
```

```r
#Selecting the desired columns
Doubling_Time <- Cellline_annotation %>%
  select(Cell_Line_Name, Doubling_Time)

#Changig the name of the "name" column to the names of the cell lines

row.names(Doubling_Time) <- Doubling_Time$Cell_Line_Name
Doubling_Time[1] <- NULL

#### (4) Table 4: Drug sensitivity
drug_sensitivity <- Sensitivity[-c(1:14),]

drug_sensitivity = t(drug_sensitivity)
```

---

## 3. SIMPLE LINEAR REGRESSION WITH 100 BIOMARKERS: Drug sensitivity with copynumber

---

Can we predict drug sensitivity using the copynumber data? How much of the variance of the data can be explained using the copynumber data?

### Table with drug sensitivity and copynumber values per cell line

```r
lm_tab2 = transform(merge(CN,DS,by=0,all=TRUE), row.names=Row.names, Row.names=NULL)
lm_tab2 <- na.omit(lm_tab2)
```

Because not all values included in copynumber are in drug sensitivity, the NAs are omitted.

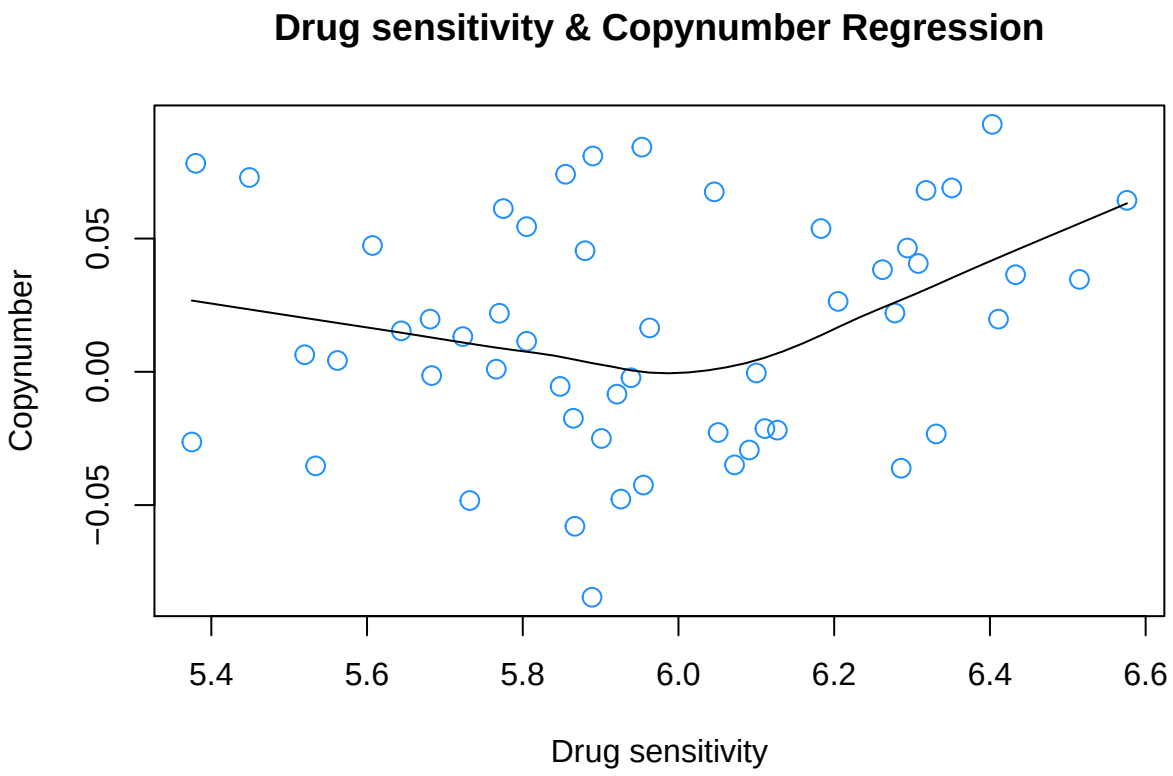
### 3.1 Plots and visualization: Predicting how fit linear regression will be as a model to describe our data

Ploting the data: can a linear relationship be observed? Should we have expected a high value for R-squared?

### Visualizations of copynumber values vs drug sensitivity

#### (1) Scatter Plot

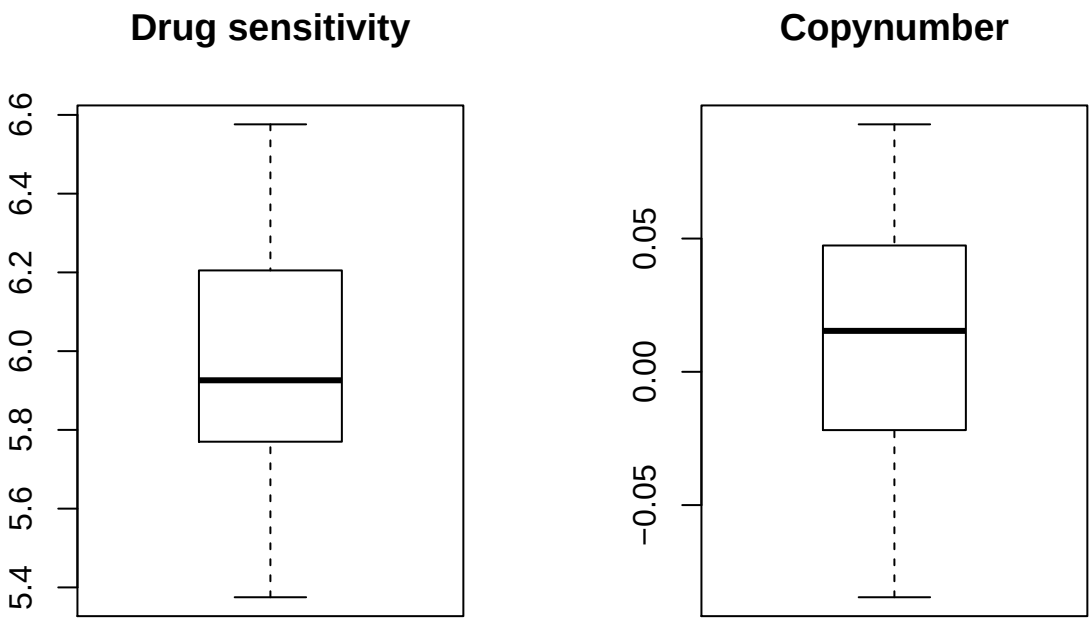
```r
scatter.smooth(lm_tab2$vorinostat,
               lm_tab2$BM_Copynumber_meancol,
               col = "dodgerblue1",
               main = "Drug sensitivity & Copynumber Regression",
               xlab = "Drug sensitivity",
               ylab = "Copynumber",
               cex = 1.3,
               pch = 1)
```




At first look, the points in the plot are so scattered, that a linear relationship seems unlikely. The second problem that one can observe in this graphic, is that the line describing the behaviour is not straight.

**(2) Box plot**

```
par(mfrow=c(1, 2))
boxplot(lm_tab2$vorinostat,
        main="Drug sensitivity")
boxplot(lm_tab2$BM_Copynumber_meancol,
        main="Copynumber")
```

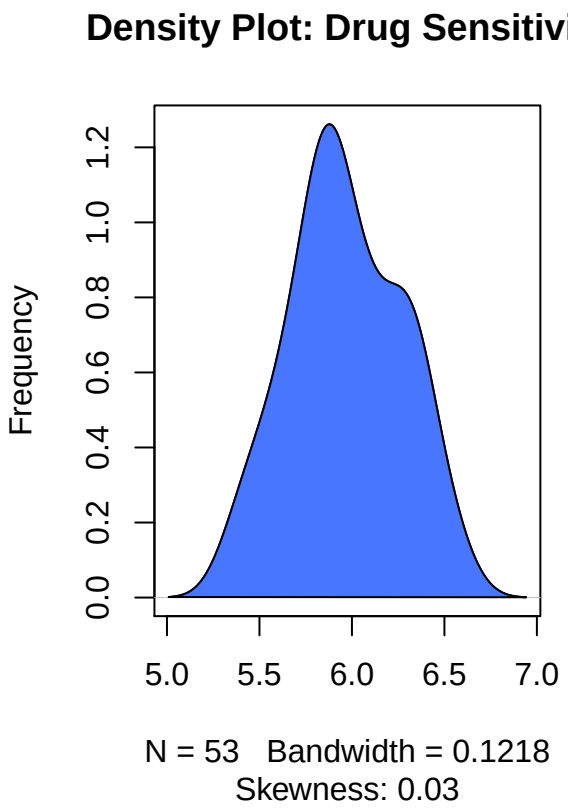


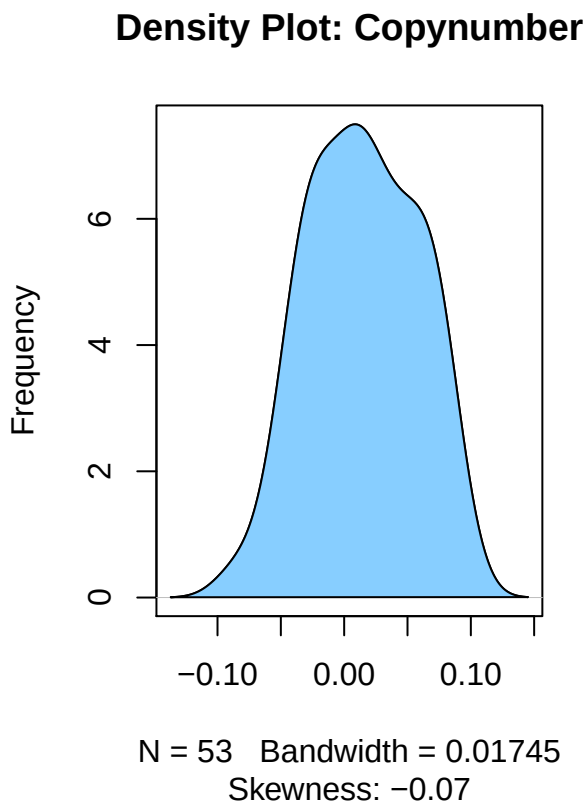

A boxplot can help us visualize the amount of outliers in our data. This is relevant as too many (extreme) outliers can have great impact on the results of our analysis and can change the outcome completely. They can easily affect the slope.

No outliers can be observed. This is good, because there is no extreme data that can affect the slope for the linear regression.

**(3) Density: Should we expect normality for drug sensitivity?**

```
par(mfrow=c(1, 2))

plot(density(lm_tab2$vorinostat),
     main="Density Plot: Drug Sensitivity",
     ylab="Frequency",
     sub=paste("Skewness:", round(e1071::skewness(lm_tab2$vorinostat), 2))
    )

polygon(density(lm_tab2$vorinostat), col="royalblue1")

plot(density(lm_tab2$BM_Copynumber_meancol),
     main="Density Plot: Copynumber",
     ylab="Frequency",
     sub=paste("Skewness:", round(e1071::skewness(lm_tab2$BM_Copynumber_meancol), 2))
    )

polygon(density(lm_tab2$BM_Copynumber_meancol), col="skyblue1")
```




**Skewness of the plot on the left**: 0.03 -> Plot is very slightly skewed to the right.

**Skewness of the plot on the right**: -0.07 -> Plot is very slightly skewed to the left.

The data seems to be well normalized in terms of skewness and in terms of shape.

**(4) Correlation: what is the level level of linear dependence between the two variables?**

```
cor(lm_tab2$vorinostat, lm_tab2$BM_Copynumber_meancol)
```

```
## [1] 0.1587939
```

A good value for correlation lies close to 1 or -1, whilst the value 0 is undesirable. Values closer to 0 indicate that there is a weak relationship. The value here is extremely low, so a linear relationship is not a very good option to describe the data.

Eventhough both the box plots (2) and the density plots (3) results could have been good indicators for a linear relationship, the lack of a fitting straight line on the scatter plot (1) and the low value in the result of the correlation (4) indicate the opposite.

These analysis help us predcit whether a linear regression is or not the best model to describe our data. Taking into consideration all results so far for this part, it is not unreasonable to predict that a linear regression will probably not be the best model to describe the relationships in our data.

---

### 3.2 Linear Regression

#### Linear Regression

```
reg2 <- lm(vorinostat ~ BM_Copynumber_meancol, data = lm_tab2)
```

#### Details about the linear regression: what we need draw some conclusions

```
summary(reg2)
```

```
##
## Call:
## lm(formula = vorinostat ~ BM_Copynumber_meancol, data = lm_tab2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65640 -0.20402 -0.00818  0.22604  0.55500
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.94966    0.04351 136.757   <2e-16 ***
## BM_Copynumber_meancol  1.10884    0.96540   1.149    0.256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2986 on 51 degrees of freedom
## Multiple R-squared:  0.02522,    Adjusted R-squared:  0.006102
## F-statistic: 1.319 on 1 and 51 DF,  p-value: 0.2561
```

**Multiple R-squared**: 0.004101

This indicates that only 0,4101% percent of the variation in the data (drug sensitivity) can be explained by the relationship between drug sensitivity and copynumber. In other words, there is a 0,4101% variance reduction when we take the copynumber into account.

**p-value**: 0.6487

As the p-value for reg2 is significantly larger than 0.05 and R-squared tells us the copynumber only explains 0,4101% of the variation in the data, it is safe to assume that there is no linear relationship between drug sensitivity and copynumer, a.k.a copynumber cannot predict drug sensitivity.

#### More information about the fit (linear ecuation: y = y-intercept + slope * x) :

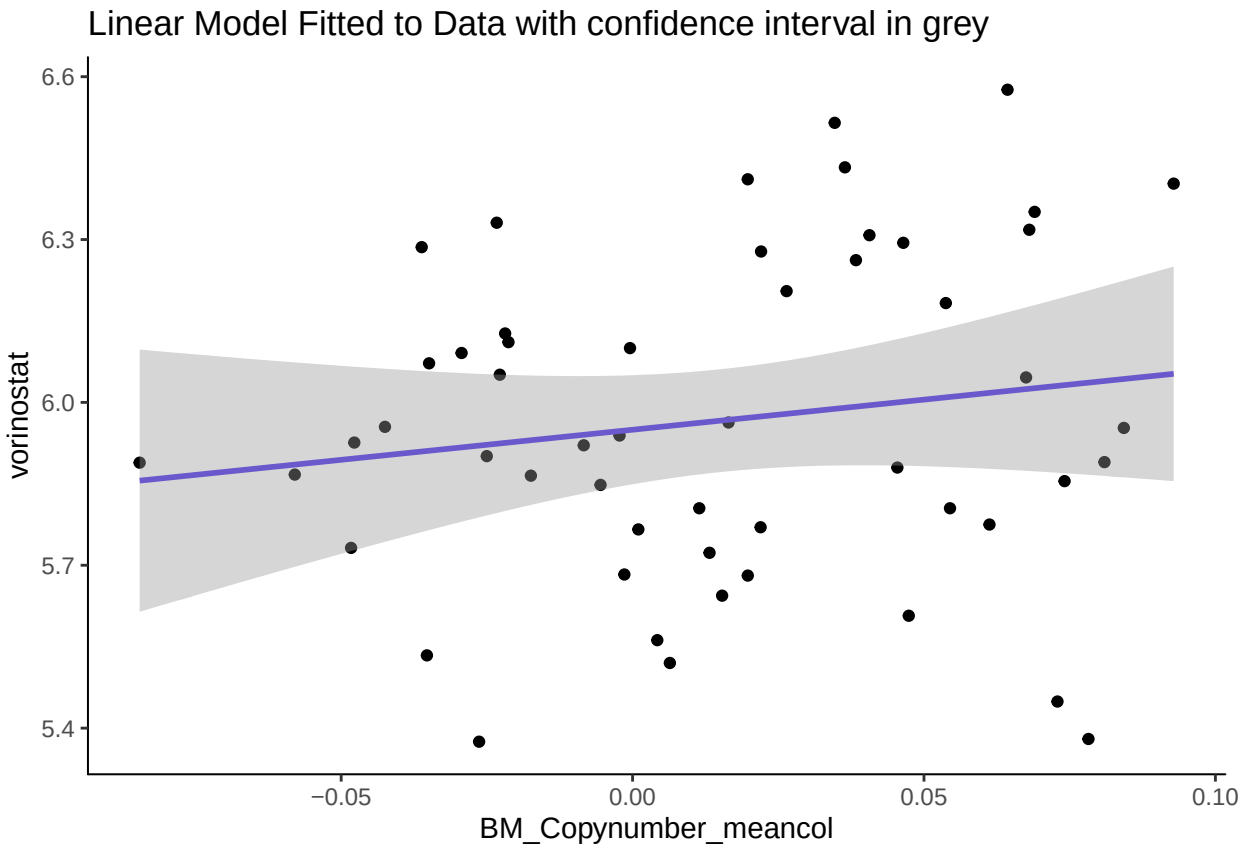
```
confint(reg2)
```

```
##                           2.5 %   97.5 %
## (Intercept)            5.8623184 6.036999
## BM_Copynumber_meancol -0.8292712 3.046958
```

With these results, it is expected that there is a 95% chance that the real value of the y-intercept should lie within 5.86 and 6.03, and that the real value for the slope should lie within -0.82 and 3.04.

We can also visualize the results of the confidence interval

```
ggplot(data = lm_tab2, aes(x = BM_Copynumber_meancol, y = vorinostat)) +
  geom_point() +
  stat_smooth(method = "lm", col = "slateblue3", level = 0.975) +
  theme(panel.background = element_rect(fill = "white"),
        axis.line.x=element_line(),
        axis.line.y=element_line()) +
  ggtitle("Linear Model Fitted to Data with confidence interval in grey")
```




Eventhough many points are inside the grey area, there are many more that are scattered across the graph. The confidence interval is very big and the data dispersed. There is no correlation between the variables.

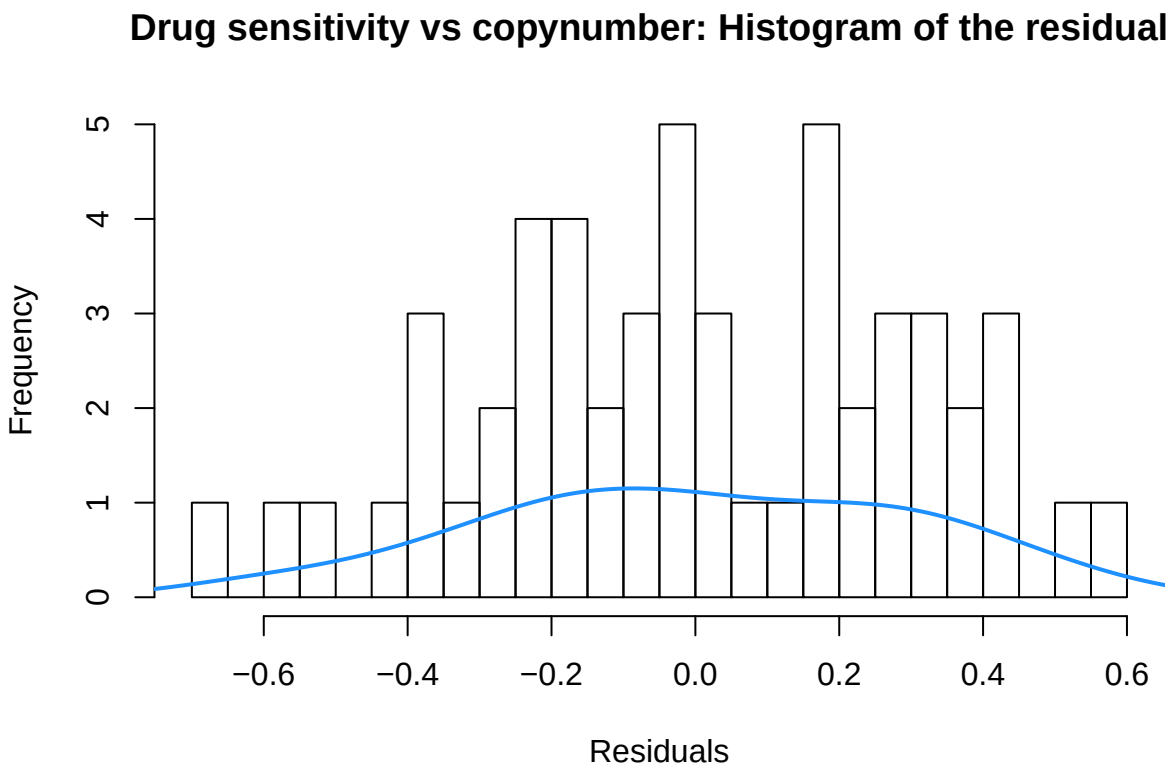
---

### 3.3 Visualization of regression

Here we explore ways to visualize the results of the regression and the normalization of residuals.

**Histogram of the residuls of the linear regression between copynumber values for 100 biomarkers and drug sensitivity**

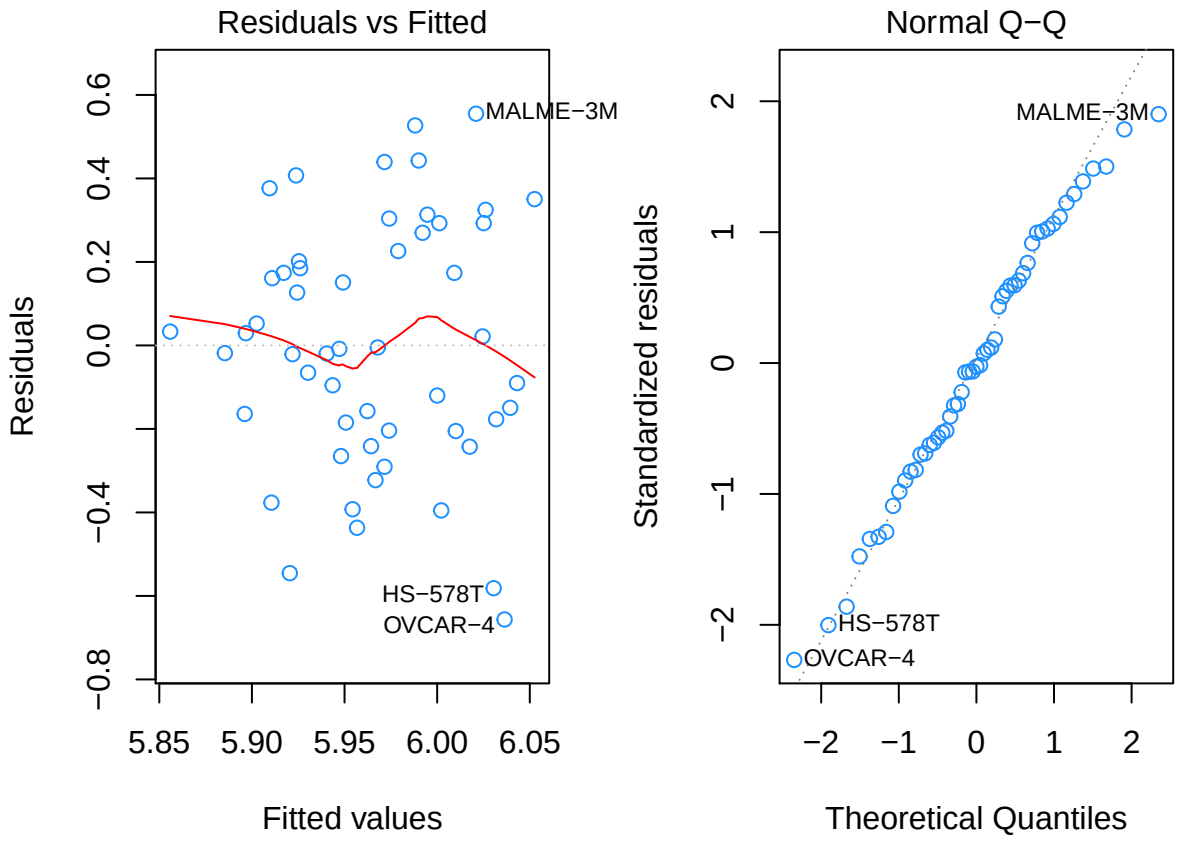
```
hist(reg2$residuals,
     breaks = 20,
     xlab = "Residuals",
     main = "Drug sensitivity vs copynumber: Histogram of the residuals")
lines(density(reg2$residuals), lwd = 2, col = "dodgerblue1")
```




The histogram shows us that the residuals are not well normalized, because the plot does not have the regular shape of a nomarl distribution and the blue line exacerbates this.

**Residual diagnostics: are the various assumptions that underpin linear regression reasonable for our data?**

```
par(mar = c(4, 4, 2, 2), mfrow = c(1, 2))
plot(reg2, which = c(1, 2), col = "dodgerblue1")
```




The red line of the Residual vs Fitted graph is a good tool that lets us visualize just how disperse our data is and that a linear model is not a good fit to describe the data.

Most points of the data in the Q-Q plot seem to meet the red line. However, it is important to take note of the points that are potitioned after the 2nd Quantile, as they are the ones that distance themselves the most from the red line. The residuals are not perfectly normalized.

The names that appear on both plots correspond to those of cell lines, and are there only for reference.

---

## 4. MULTIPLE REGRESSION WITH 100 BIOMARKERS: Drug sensitivity with doubling time and copynumber

---

**Table with drug sensitivity, copynumber and doubling time per cell line**

```
lm_tab_m = transform(merge(CN,
                           lm_tab,by=0,all=TRUE),
                     row.names=Row.names,
                     Row.names=NULL
                    )
```

---

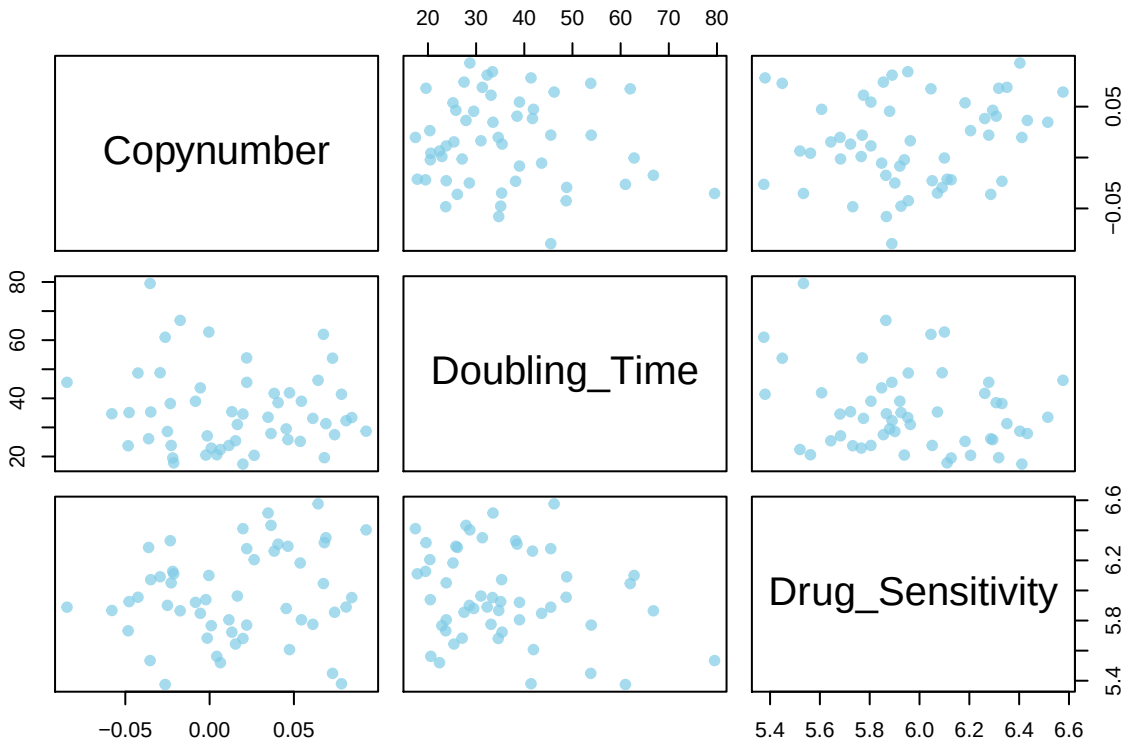
**4.1 Plots and visualization: Predicting how fit linear regression will be as a model to describe our data**

**Ploting the data: can a linear relationship be observed? Should we have expected a high value for R-squared?**

**Visualization of drug sensitivity vs copynumber and doubling time**

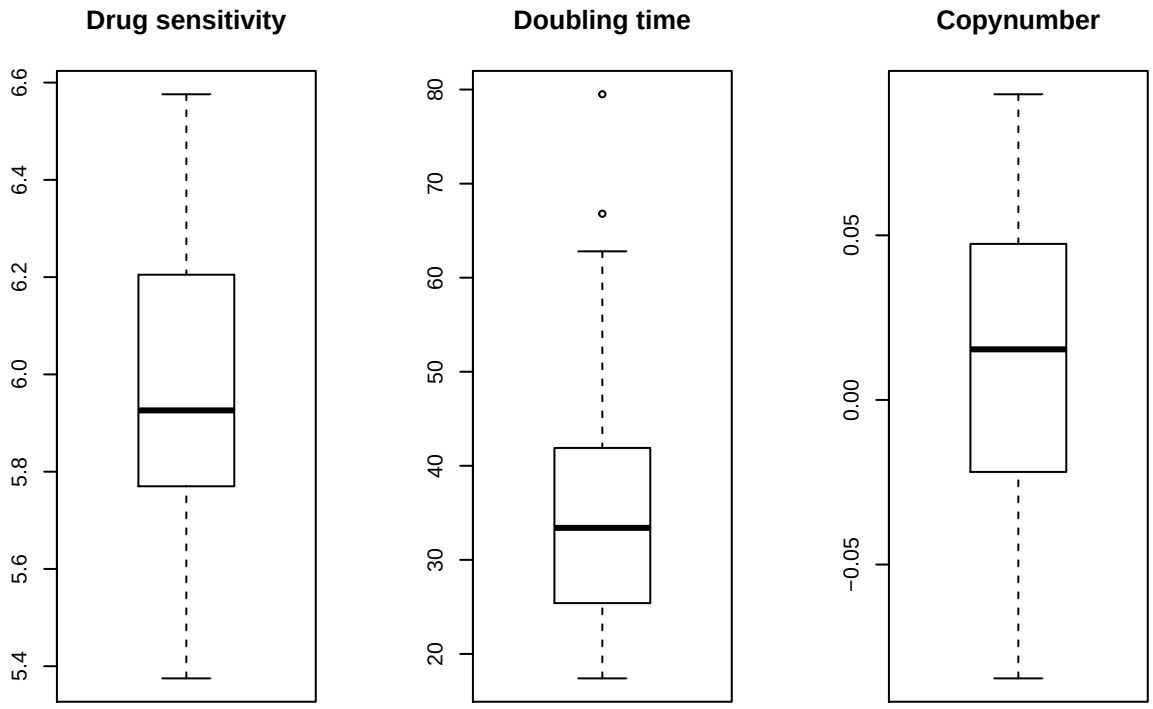
**(1) Plot**

```
plot(lm_tab_m ,
     pch=20 ,
     cex=1.5 ,
     col=rgb(0.5, 0.8, 0.9, 0.7)
    )
```



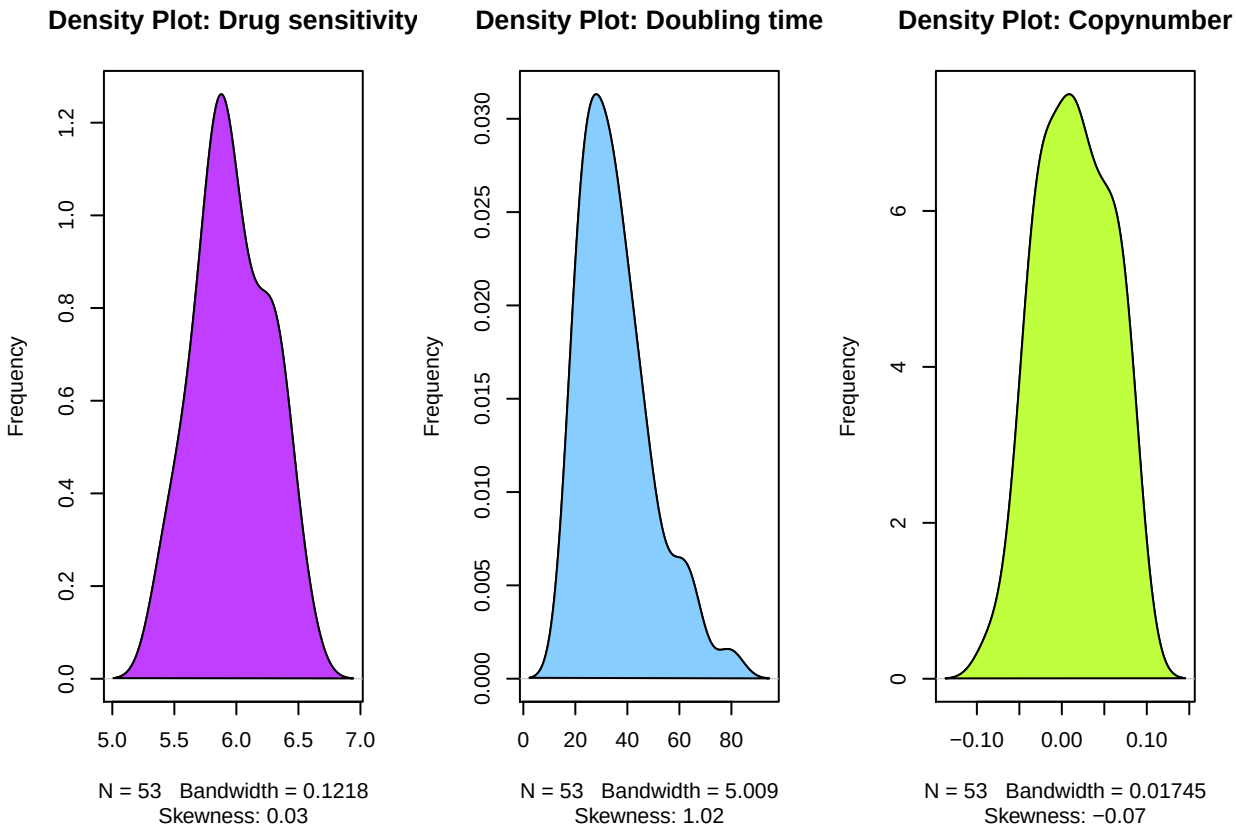

There seems to be no linear relationship, because of how spread the points are.

**(2) Box plot**




There are two outliers in the boxplot of doubling time.

**(3) Density: Should be expect normality for drug sensitivity?**




**Skewness of the plot on the left**: 0.03 -> Plot is very slightly skewed to the right.

**Skewness of the plot on the middle**: 1.02 -> Plot is slightly skewed to the left.

**Skewness of the plot on the right**: -0.42 -> Plot is slightly skewed to the left.

The density plot for doublin time has an unusual shape for normally distributed data and is extremelly skewed in comparison to the other plots.

### (4) Checking for correlation

```
##                   Copynumber Doubling_Time Drug_Sensitivity
## Copynumber         1.00000000   -0.09821746        0.1587939
## Doubling_Time     -0.09821746    1.00000000       -0.2360583
## Drug_Sensitivity   0.15879386   -0.23605826        1.0000000
```

None of the values here indicate a strong linear relationship.

Thanks to the outliers, considerably spread plots and skewed density plots, it is not unreasonable to predict that a multiple regression with these parameters will probably not be the best model to describe the relationships in our data.

---

### 4.2 Multiple Regression

#### Multiple Regression

```
reg_m <- lm(Drug_Sensitivity ~ Copynumber + Doubling_Time, data = lm_tab_m)
```

#### Details about the linear regression: what we need draw some conclusions

```
summary(reg_m)
```

```
##
## Call:
## lm(formula = Drug_Sensitivity ~ Copynumber + Doubling_Time, data = lm_tab_m)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61919 -0.21532 -0.00697  0.23039  0.61326
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.124286   0.115488  53.030   <2e-16 ***
## Copynumber     0.956168   0.954758   1.001    0.321
## Doubling_Time -0.004828   0.002966  -1.628    0.110
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2939 on 50 degrees of freedom
## Multiple R-squared:  0.07429,    Adjusted R-squared:  0.03726
## F-statistic: 2.006 on 2 and 50 DF,  p-value: 0.1452
```

**Multiple R-squared**: 0.05949

This indicates that only 5.949% percent of the variation in the data (drug sensitivity) can be explained by the relationship between drug sensitivity, doubling time and copynumber. In other words, there is a 5.949% variance reduction when we take the both the doubling time and the copynumber into account.

**p-value**: 0.2158

As the p-value for reg_m is significantly larger than 0.05 and R-squared tells us the copynumber only explains 2.355% of the variation in the data, it is safe to assume that there is no linear relationship between drug sensitivity and copynumer, a.k.a copynumber cannot predict drug sensitivity.

**F-statistic**

- F-statistic (multiple regression) : 1.72
  - F-statistic (drug sensitivity vs copynumber): 1.32
    The t-values show that a change in doubling time is more strongly associated with a change in drug sensitivity than a change in copynumber value would be. The coefficients show us aswell that better results are yielded when using doubling time alone to predict drug sensitivity,than when using both doubling time and copynumber.

    **More information about the fit (linear ecuation: y = y-intercept + slope * x) :**

```
confint(reg_m)
```

```
##                      2.5 %     97.5 %
## (Intercept)    5.89232222 6.35625001
## Copynumber    -0.96152010 2.87385667
## Doubling_Time -0.01078476 0.00112821
```

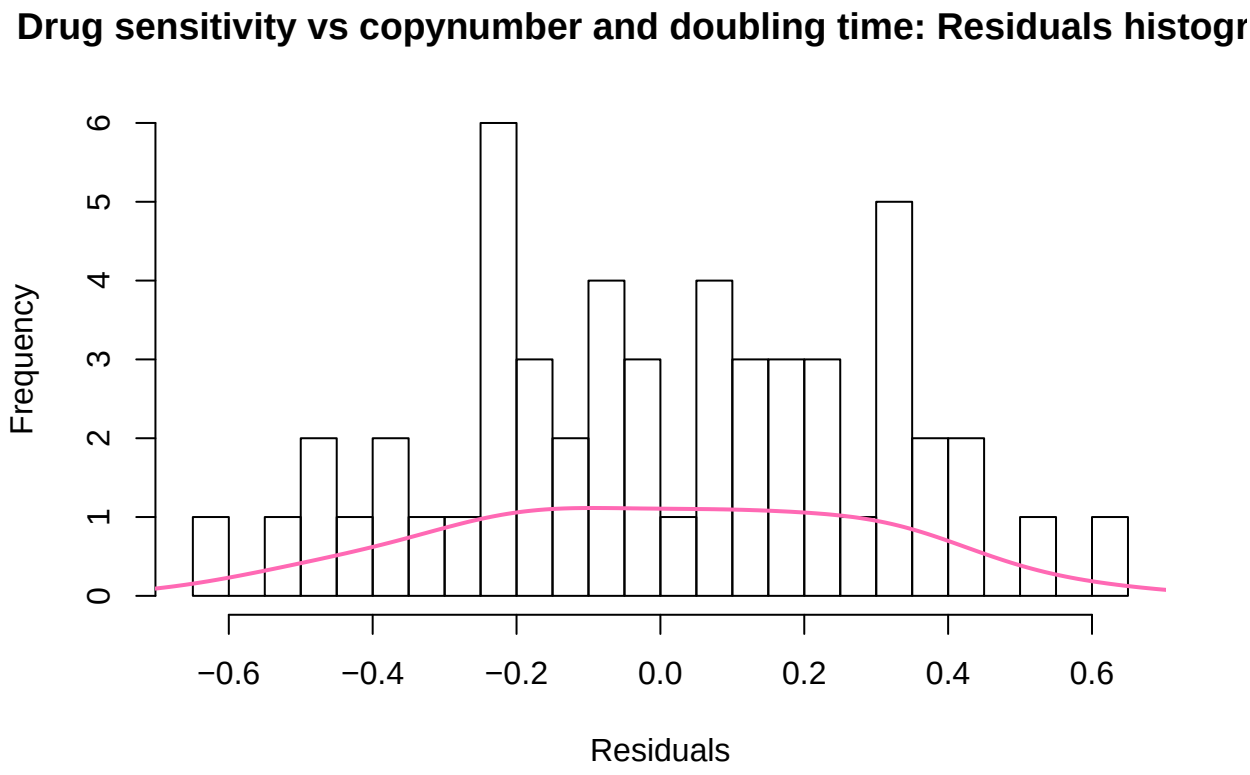
This information can be used to draw a plane of regression with a 95% chance of it being the correct plane for this data.

---

### 4.3 Visualization of regression

Here we one again explore ways to visualize the results of the regression and the normalization of residuals.

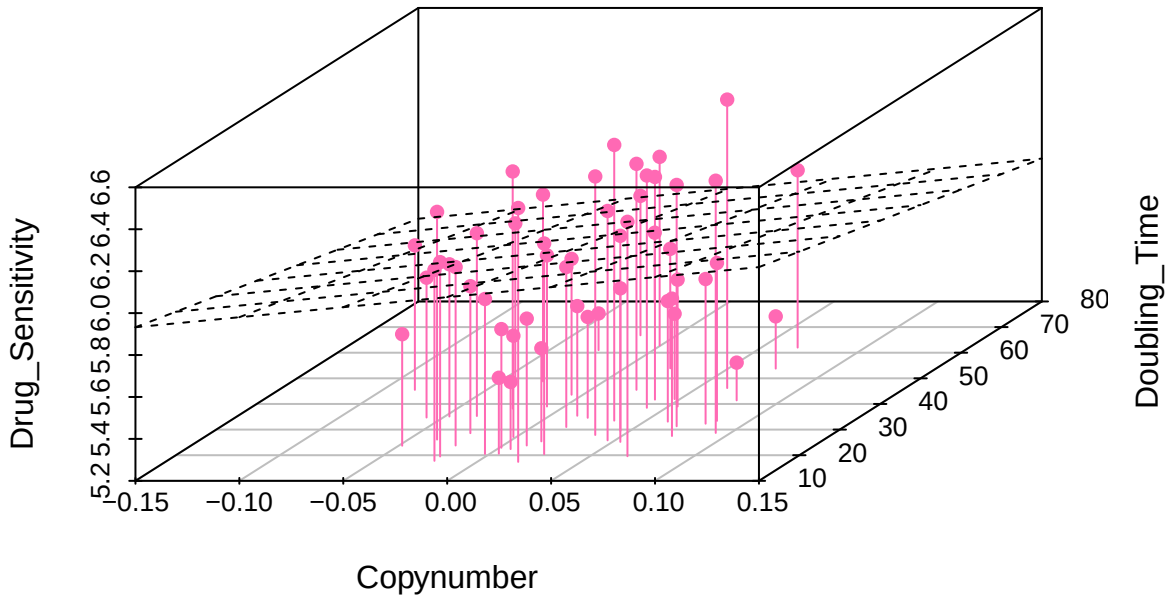
**Histogram of the residuls of the linear regression between copynumber values for 100 biomarkers, doubling time and drug sensitivity**




The data does NOT follow a normal distribution.

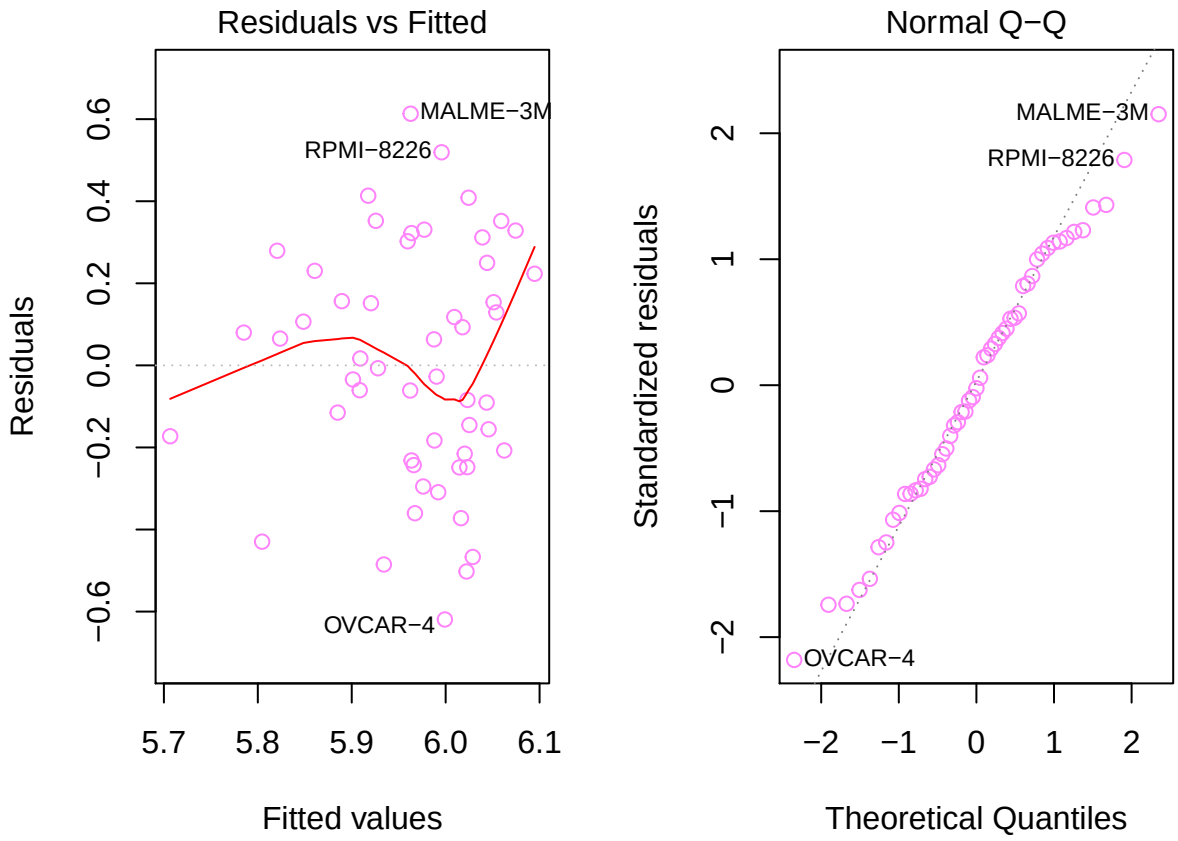
**3D Plot with Regression Plane**

```r
#Creating 3D plot
m_3d <- scatterplot3d(lm_tab_m,
                      type = "h",
                      color = "hotpink",
                      angle=55,
                      pch = 16)
#Adding regression plane
reg_m_3D <- lm(Drug_Sensitivity ~ Copynumber + Doubling_Time, data = lm_tab_m)
m_3d$plane3d(reg_m_3D)
```




This 3D plot allows us to see the plane created for the multivariate regression and the location of data points in relation to to the plane. Most points are either above the plane or below it. This matches our previous results for the multiple regression, from which we conlcuded there is no relevant relationship between the variables.

**Residual diagnostics: are the various assumptions that underpin linear regression reasonable for our data?**




The red line is not straight and the points are very scattered in the graph on the left.

Eventhough, many data points in the graph on the right fit the straight line, there are also many others in both extremes of the line that spread away from it. This plot shows how normalized the data is, and with these results we can conclude that the resudials are very poorly normalized.

---

## 5. SUMMARY: MULTIPLE REGRESSION

---

Here we perform one last large multiple regression to explore how our results might change when we consider the copynumber of a set of predifined genes and when we compare their results for a regression model for drug sensitivity with those of doubling time and copynumber of 100 biomarkers.

Using the table FC_both_sorted from the first part of the specific analysis, which contains the biomarkers in decreasing order of importance we can obtain the most and least relevant biomarkers/genes, aswell as some genes that

We also aim to find the genes with the highest and lowest values for copynumber. To achieve this, we find the mean of the absolute values and then sort the results in decresing order.

---

### 5.1 Genes with the highest and lowest level of expression

**Sorting genes according to their mean value accross cell lines**

```
Copynumber_av= abs(Copynumber)
Copynumber_mean= rowMeans(Copynumber_av)

Copynumber_sav <- sort(Copynumber_mean, decreasing = TRUE)
Copynumber_sav <- as.matrix(Copynumber_sav)
```

**Top 10**

```
CN_top10 = Copynumber_sav[1:10,]
CN_top10
```

```
##      DAZ2       UTY      DAZ1     DDX3Y    TTTY20  FAM197Y5 FAM197Y2P
##  3.645445  3.172730  2.936079  2.818717  2.681708  2.671509  2.671509
##     TSPY1     TSPY3     TSPY4
##  2.671509  2.671509  2.671509
```

**Lowest 10**

```
CN_lowest10 = Copynumber_sav[23306:23316,]
CN_lowest10
```

```
##     MIR545     MIR421      ABCB7   KIAA2022       RLIM       XIST
## 0.09233774 0.09233774 0.09197547 0.09197547 0.09197547 0.09192264
##    CXorf26     MAGEE1    ZDHHC15     MAGEE2     TTC3P1
## 0.09169434 0.09169434 0.09134340 0.09126415 0.09126415
```

---

### 5.2 Table describing the categories to be used for the regression

| | Category |
|---|---|
| DHRS2 | First top 10 biomarker |
| ABAT | Second top 10 biomarker |
| DAZ2 | First top 10 copynumber value |
| UTY | Second top 10 copynumber value |
| DAZ1 | Third top 10 copynumber value |
| TTC3P1 | Lowest lowest copynumber value |
| ZDHHC15 | Third lowest copynumber value |
| Copynumber_100_biomarkers | Mean Copynumber for 100 biomarkers |
| Doubling_Time | Doubling time |

---

### 5.3 Table with all categories for the multiple regression

```
lm_tab_m2 = transform(merge(t(Copynumber[c("DHRS2", "ABAT",
                                           "DAZ2", "UTY",
                                           "TTC3P1", "ZDHHC15" ),
                                         ]
                            ),
                      lm_tab_m,
                      by=0,
                      all=TRUE),
                     row.names=Row.names,
                     Row.names=NULL
                    )
```

---

### 5.4 Multiple Regression

#### Multiple Regression

```
reg_m2 <- lm(Drug_Sensitivity ~ DHRS2 + ABAT +
               DAZ2 + UTY +
               TTC3P1  + ZDHHC15 +
               Copynumber_100_biomarkers + Doubling_Time, data = lm_tab_m2)
```

#### Details about the linear regression: what we need draw some conclusions

```
summary(reg_m2)
```

```
##
## Call:
## lm(formula = Drug_Sensitivity ~ DHRS2 + ABAT + DAZ2 + UTY + TTC3P1 +
##     ZDHHC15 + Copynumber_100_biomarkers + Doubling_Time, data = lm_tab_m2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50356 -0.17552  0.01708  0.18186  0.76789
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 6.095232   0.132572  45.977   <2e-16 ***
## DHRS2                       0.044366   0.131713   0.337    0.738
## ABAT                        0.174081   0.146252   1.190    0.240
## DAZ2                       -0.001103   0.027828  -0.040    0.969
## UTY                        -0.018235   0.024695  -0.738    0.464
## TTC3P1                     -7.625858   9.444832  -0.807    0.424
## ZDHHC15                     7.365561   9.453187   0.779    0.440
## Copynumber_100_biomarkers   1.227682   1.036444   1.185    0.243
## Doubling_Time              -0.005210   0.003392  -1.536    0.132
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.299 on 44 degrees of freedom
## Multiple R-squared:  0.1567, Adjusted R-squared:  0.003325
## F-statistic: 1.022 on 8 and 44 DF,  p-value: 0.4343
```

**Multiple R-squared**: 0.1842

This indicates that 18.42% percent of the variation in the data (drug sensitivity) can be explained by the relationship between drug sensitivity and all the other categories included in this analysis.

**p-value**: 0501

Eventhough our results for multiple R-suared are positively high, because the p-value for reg_m2 is significantly larger than 0.05 it is not safe to reject the H0-hypothesis.

**F-statistic**

- F-statistic (multiple regression 1) : 1.72
  - F-statistic (drug sensitivity vs copynumber): 1.32
    - F-statistic (multiple regression 1) : 0.9481
      The t-values show that both genes with a low copynumber are better predictors for drug-sensitivity than those with a high copynumber. Whether a gene is a biomarker or not does not really correlate with a better t-value.
      We can only conclude that we would need to carry a larger analysis comparing both genes with high mean values for copynumber and with low mean values for copynumber to obtain results that are truthful.

---

## 6. General Conclusions

---

The predictions made with the scatter plots, boxplots, density plots and the correlations between the predicted and predicting variable were generally speaking good, as most of these statistical analyses predicted that building a model using linear/multiple regression would prove to be not ideal. Once we tested for R-squared and p-value, every regression had extremely low R-quared values, and extremely high p-values. This is of course an undesired result.

Considering the complexity of the process of gene expression, is not entirely surprising that we cannot predict drug sensitivity in a satisfactory way just by relying on copynumber and/or doubling time.

Eventhough no linear relationships were found, we did discover that doubling time is a better tool to predict our data for drug sensitivity than copynumber and that a low mean value of copynumber is better correlated to drug sensitivity than a high mean value is.

## 7. Appendix

### 7.1. REMOVING THE OUTLIERS: Simple linear regression for drug sensitivity using doubling time

#### 7.1.1 Simple linear regression: Drug sensitivity with doubling time

Can we predict drug sensitivity using doubling time? How much of the variance of the data can be explained using the doubling time?

**Table with drug sensitivity and doubling time per cell line**

```
lm_tab = transform(merge(DT,DS,by=0,all=TRUE), row.names=Row.names, Row.names=NULL)
```

**Linear Regression**

```
reg1 <- lm(vorinostat ~ Doubling_Time, data = lm_tab)
```

**Details about the linear regression: what we need draw some conclusions**

```
##
## Call:
## lm(formula = vorinostat ~ Doubling_Time, data = lm_tab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57361 -0.22994 -0.03867  0.24528  0.81819
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.144923   0.109253  56.245   <2e-16 ***
## Doubling_Time -0.004621   0.002861  -1.615    0.112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3042 on 59 degrees of freedom
## Multiple R-squared:  0.04235,    Adjusted R-squared:  0.02611
## F-statistic: 2.609 on 1 and 59 DF,  p-value: 0.1116
```

**Multiple R-squared**: 0.04235

**p-value**: 0.1116

### 7.1.2 Boxplot: removing the outliers

```
#Storing the values of the outliers in a vector

outliers_DT <- boxplot(lm_tab$Doubling_Time, plot=FALSE)$out

#Removing the outliers

lm_tab[which(lm_tab$Doubling_Time %in% outliers_DT),]
```

```
##        Doubling_Time vorinostat
## A498            66.8      5.865
## HOP-92          79.5      5.534
```

```
summary(lm_tab$Doubling_Time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.40   25.80   33.10   35.68   41.90   79.50
```

We can see that the outliers correspond to those of the cell lines A498 (renal) and HOP-92 (lung) with respective values of 66.8 and 79.5. YThe mean has a value of only 35.68.

Creating a new table without the outliers

```
lm_tab_out <- lm_tab[-which(lm_tab$Doubling_Time %in% outliers_DT),]
```

### 7.2 Linear Regression removing the outliers

```
reg_out <- lm(vorinostat ~ Doubling_Time, data = lm_tab_out)
```

**Details about the linear regression: what we need draw some conclusions**

```
##
## Call:
## lm(formula = vorinostat ~ Doubling_Time, data = lm_tab_out)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5857 -0.2199 -0.0400  0.2514  0.8151
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.106988   0.122672  49.783   <2e-16 ***
## Doubling_Time -0.003413   0.003370  -1.013    0.316
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3074 on 57 degrees of freedom
## Multiple R-squared:  0.01767,    Adjusted R-squared:  0.0004385
## F-statistic: 1.025 on 1 and 57 DF,  p-value: 0.3155
```

Multiple R-squared has a value of 0.01767. The Multiple R-squared when the outliers are not removed is 0.04235.

The p-value equals 0.3155. The p-value when the outliers are not removed is 0.1116. The result when the outliers are removed is significantly larger.

We can already observe that this small change, the removal of outliers, has a significant impact on our results.

#### 7.2.1 Visualization OF regressions

**Histogram of the residuls of the linear regression with changes in the presence of outliers**

The data in both histograms does not appear normalized and it is hard to say whether there is a case in which the data looks more normally distributed or not.

**Residual diagnostics: are the various assumptions that underpin linear regression reasonable for our data?**

Plots with outliers are on top and those without are on the bottom.

Small changes are observed, with the most relevant one being on the graphs on the left, where the shape of the red line as sightly changed.

### 7.3 Conclusion

There are small differences in the plots when the outliers are removed. The relationship between both variables gets more uncertain when the outliers are removed. As the R-Squared value decreases and p-value increases, it is not safe to say whether removing the outliers is better or not. The resukts of this small exploratory analysis do not allow us to know which option will render more truthfull results. The conclusion we can draw is that outliers have a powerful effect on the end results of a statistical analysis and that they should be considered when making any kind of analysis.