Worksheet 5 Group 1111

Ammara Akhtar, Afia Ibnath, and Reuben Walker

19 May 2024

```
library(tidyverse)
```

Load libraries

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(plotrix)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

Data summary

```
gene_data <- read.csv('genetic_circuit.csv')
str(gene_data)
```

```
## 'data.frame':    567 obs. of  5 variables:
##  $ concentration: int  10 20 40 60 80 100 1000 10 20 40 ...
##  $ seap         : num  19.1 19.3 19 19.5 20.6 ...
##  $ experiment   : chr  "ex1" "ex1" "ex1" "ex1" ...
##  $ figure       : chr  "1f" "1f" "1f" "1f" ...
##  $ cytokine     : chr  "IL-4 and IL-13" "IL-4 and IL-13" "IL-4 and IL-13" "IL-4 and IL-13" ...
```

```
# Summary statistics
summary(gene_data)
```

```
##  concentration        seap           experiment           figure
##  Min.   :  10.0   Min.   :  0.0000   Length:567         Length:567
##  1st Qu.:  20.0   1st Qu.:  0.9836   Class :character   Class :character
##  Median :  60.0   Median :  2.0028   Mode  :character   Mode  :character
##  Mean   : 187.1   Mean   : 25.1850
##  3rd Qu.: 100.0   3rd Qu.: 21.4677
##  Max.   :1000.0   Max.   :212.7027
##    cytokine
##  Length:567
##  Class :character
##  Mode  :character
##
##
##
```

Compute the means and standard errors of the means over the three experiments

```
#Concentration:
#ag1 <- gene_table[, sapply(.SD, function(x) list(mean=mean(x), sd=sd(x))), by=concentration]

ag1 <- aggregate(. ~ concentration, select(gene_data, concentration, seap), function(x) c(mean = mean(x
#aggregate puts the aggregate columns into a results matrix.
#Convert back into df columns:
ag1 <- cbind(ag1[-ncol(ag1)],ag1[[ncol(ag1)]])
ag1
```

```
##   concentration     mean       se
## 1            10 12.65287 1.765546
## 2            20 16.56497 2.400708
## 3            40 21.89383 3.314982
## 4            60 24.21542 3.873906
## 5            80 26.75310 4.243582
## 6           100 27.81042 4.555923
## 7          1000 46.40444 8.115223
```

```
#Set of cytokines:
ag2 <- aggregate(. ~ cytokine, select(gene_data, cytokine, seap), function(x) c(mean = mean(x), se = std
cbind(ag2[-ncol(ag2)],ag2[[ncol(ag2)]])
```

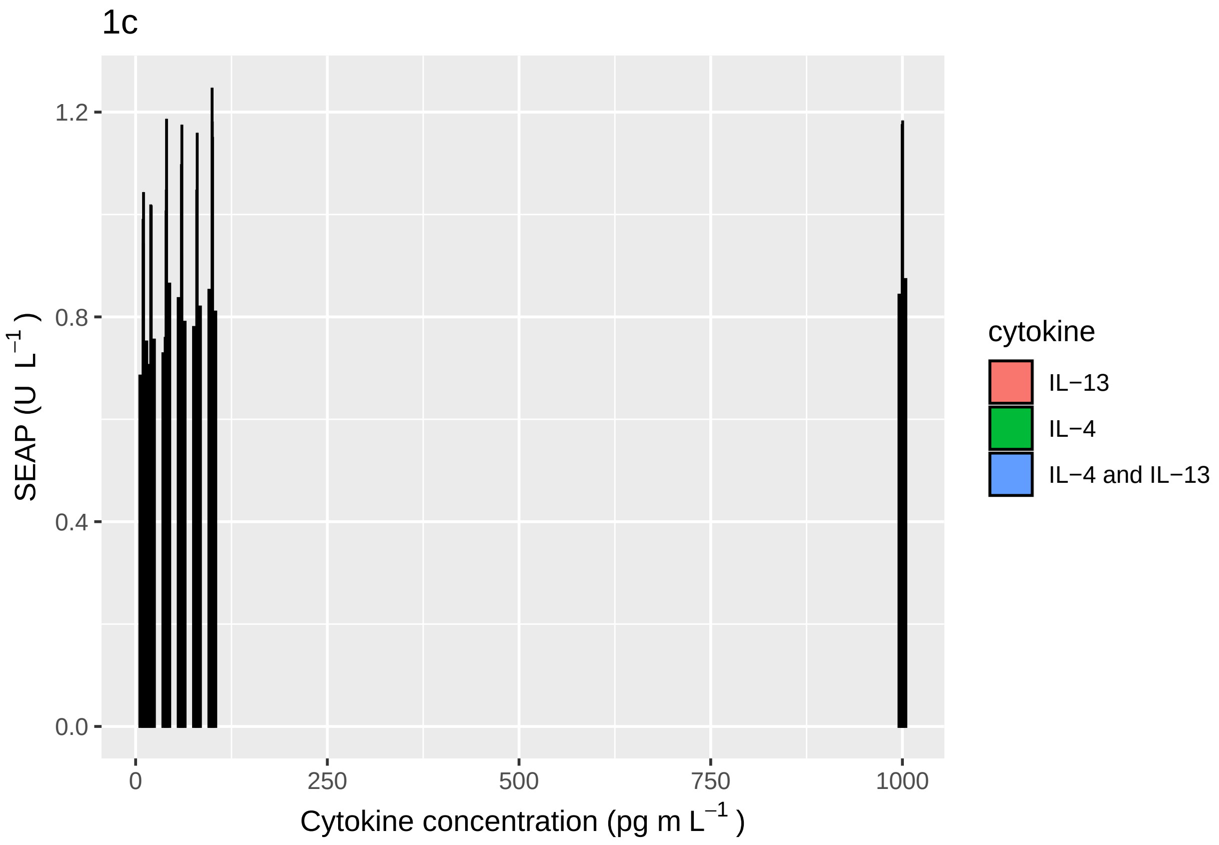
```
##         cytokine     mean       se
## 1          IL-13 22.06974 2.647777
## 2           IL-4 24.02357 2.839754
## 3 IL-4 and IL-13 29.46171 3.436769
```

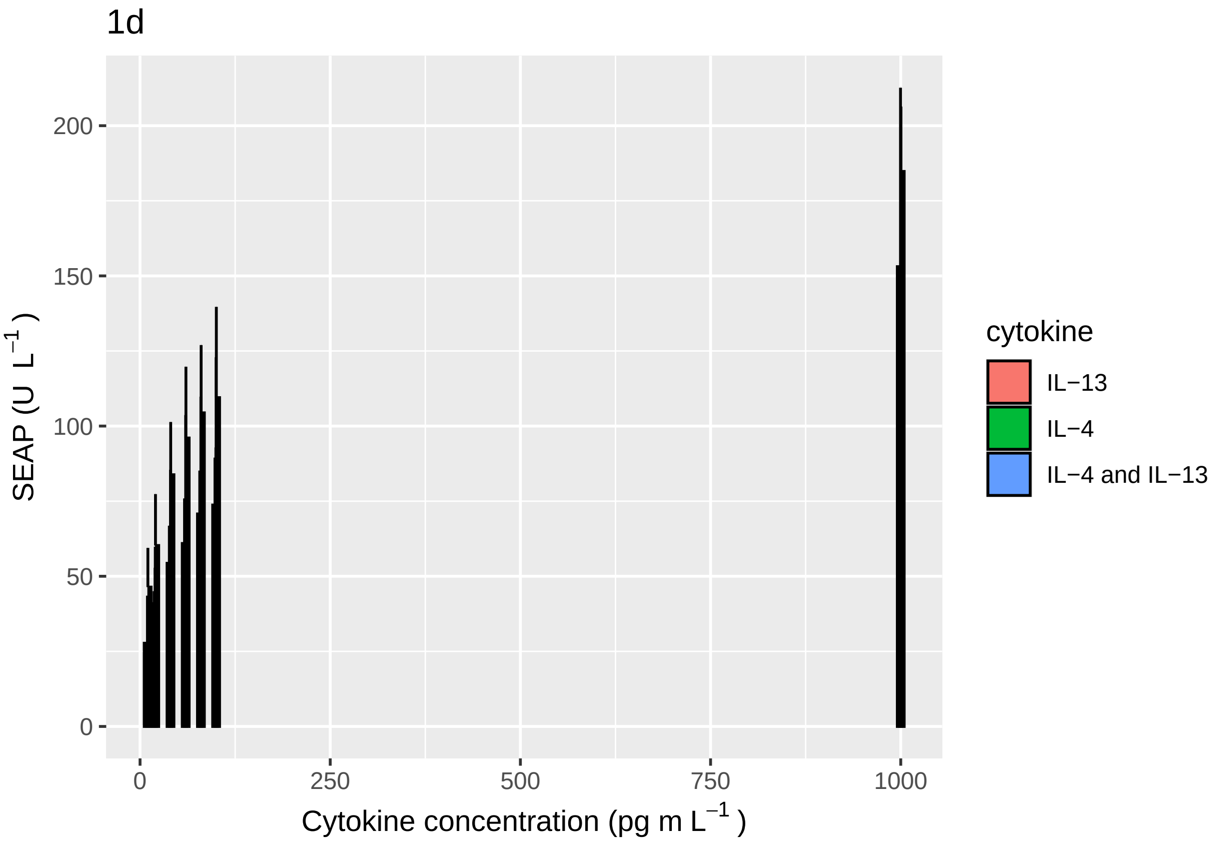
```
#Figure setting:
ag3 <- aggregate(. ~ figure, select(gene_data, figure, seap), function(x) c(mean = mean(x), se = std.er
cbind(ag3[-ncol(ag3)],ag3[[ncol(ag3)]])
```

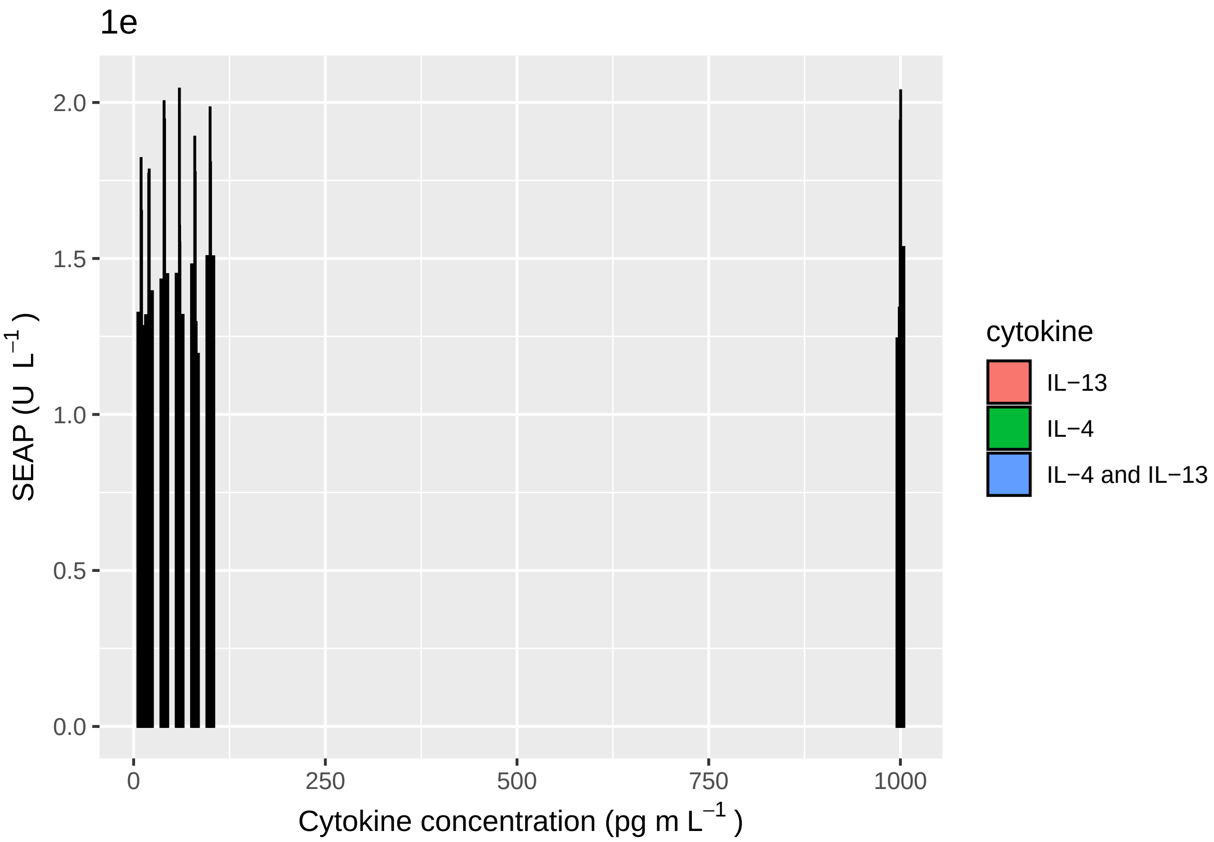
```
##   figure       mean          se
## 1     1c  0.7641888 0.02458086
## 2     1d 80.8623377 3.88219237
## 3     1e  1.3342501 0.03458992
## 4     1f 16.5449609 0.24922959
```

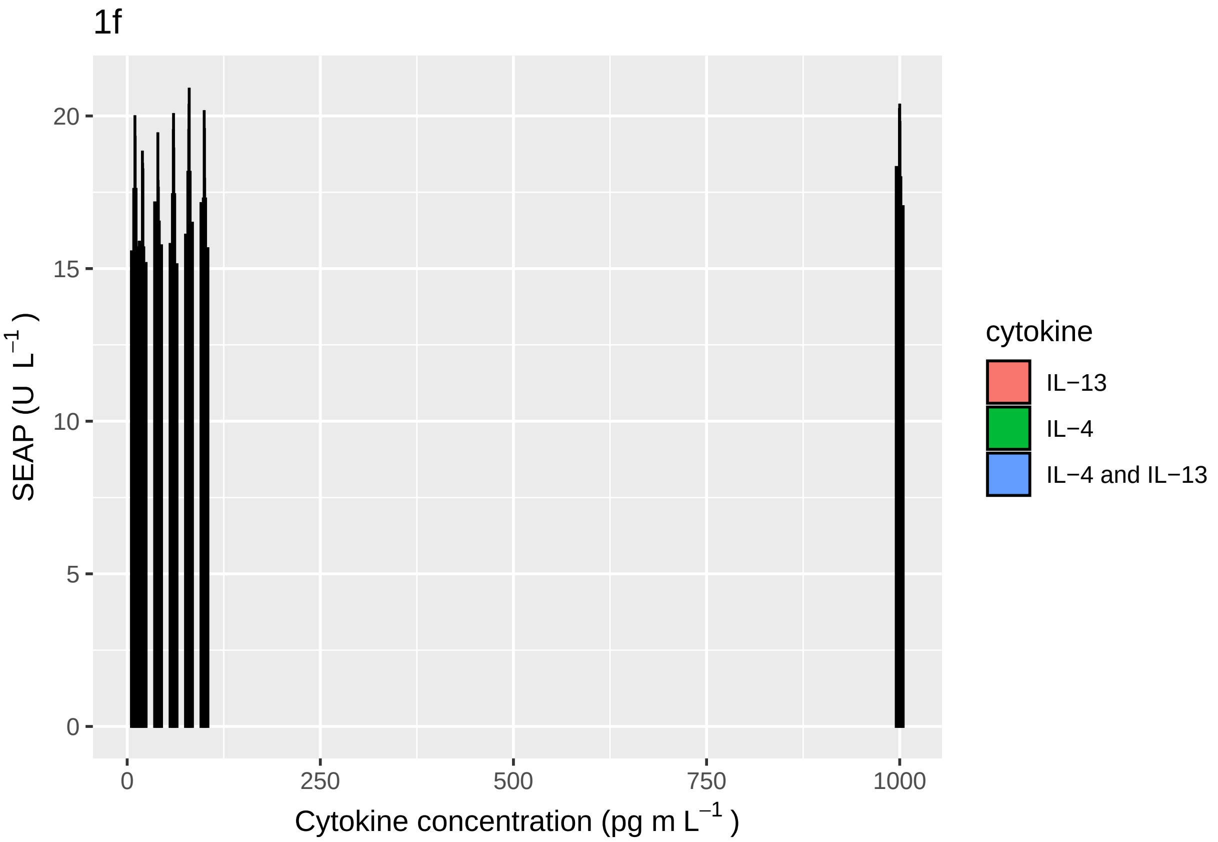
Recreating figures

```
#Figure
#Mean and standard deviation plot grouped by
  #cytokine concentration, and
  #cytokine set
# rectangle
for (group in c("1c","1d","1e","1f")) {

  ag_c <- aggregate(. ~ cytokine+concentration, select(subset(gene_data, gene_data$figure == group), cy
  #aggregate puts the aggregate columns into a results matrix.
  #Convert back into df columns:
  ag_c <- cbind(ag_c[-ncol(ag_c)],ag_c[[ncol(ag_c)]])


  print(
    ggplot(ag_c, aes(x=concentration, y=mean, fill=cytokine)) +
    geom_bar(position=position_dodge(), stat="identity", colour='black') +
    #geom_col(position=position_dodge(), stat="identity", colour='black') +
    geom_errorbar(aes(ymin=mean, ymax=mean+sd), width=.2,position=position_dodge(.9)) +
    labs(x=bquote("Cytokine concentration (pg m"~L^-1~")"), y=bquote("SEAP (U "~L^-1~")")) +
    ggtitle(group)
  )
}
```


1c
SEAP (U L^{-1})
1.2
0.8
0.4
0.0
0
250
500
750
1000
Cytokine concentration (pg m L^{-1})
cytokine
🟥 IL−13
🟩 IL−4
🟦 IL−4 and IL−13


1d
200
150
100
50
0
SEAP (U L^{-1})
0
250
500
750
1000
Cytokine concentration (pg m L^{-1})
cytokine
IL−13
IL−4
IL−4 and IL−13


1e
SEAP (U L^{-1})
2.0
1.5
1.0
0.5
0.0
0
250
500
750
1000
Cytokine concentration (pg m L^{-1})
cytokine
🟥 IL−13
🟩 IL−4
🟦 IL−4 and IL−13



Once the axis breaks are removed from the paper's figures, it becomes apparent that our data has large non-normal disparities in cytokine concentration as well as large differences in the circuit response between groups and any model will need to account for them.

Question 2

```
data <- read.csv("genetic_circuit.csv")
#data
print(head(data))
```

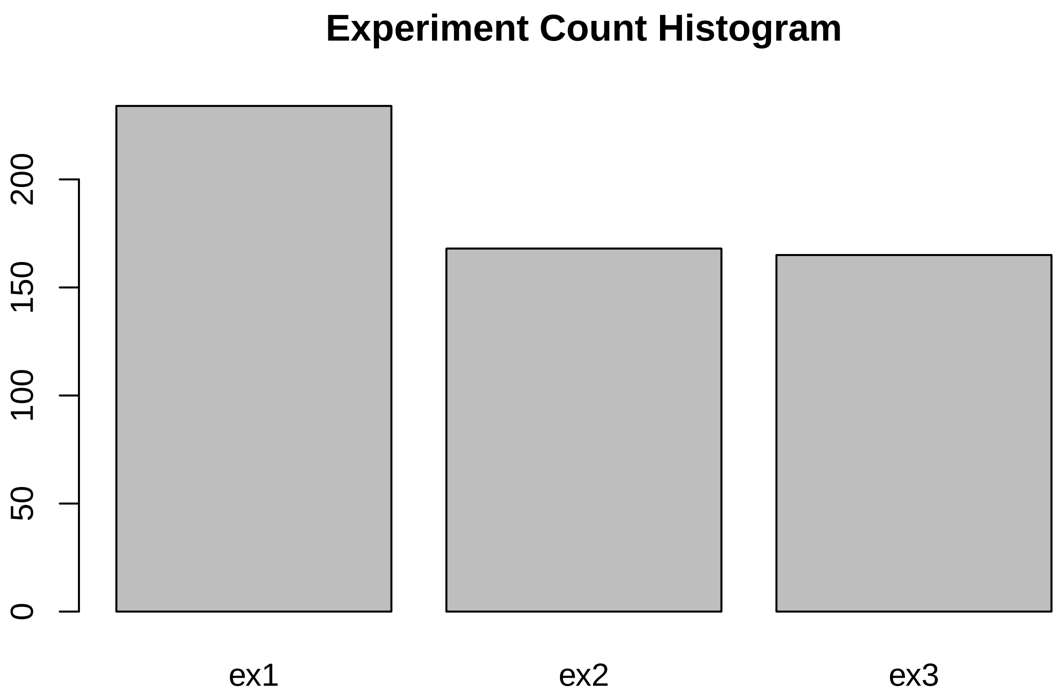
```
##   concentration   seap experiment figure       cytokine
## 1            10 19.131        ex1     1f IL-4 and IL-13
## 2            20 19.266        ex1     1f IL-4 and IL-13
## 3            40 19.009        ex1     1f IL-4 and IL-13
## 4            60 19.506        ex1     1f IL-4 and IL-13
## 5            80 20.631        ex1     1f IL-4 and IL-13
## 6           100 16.851        ex1     1f IL-4 and IL-13
```

```
print(summary(data))
```

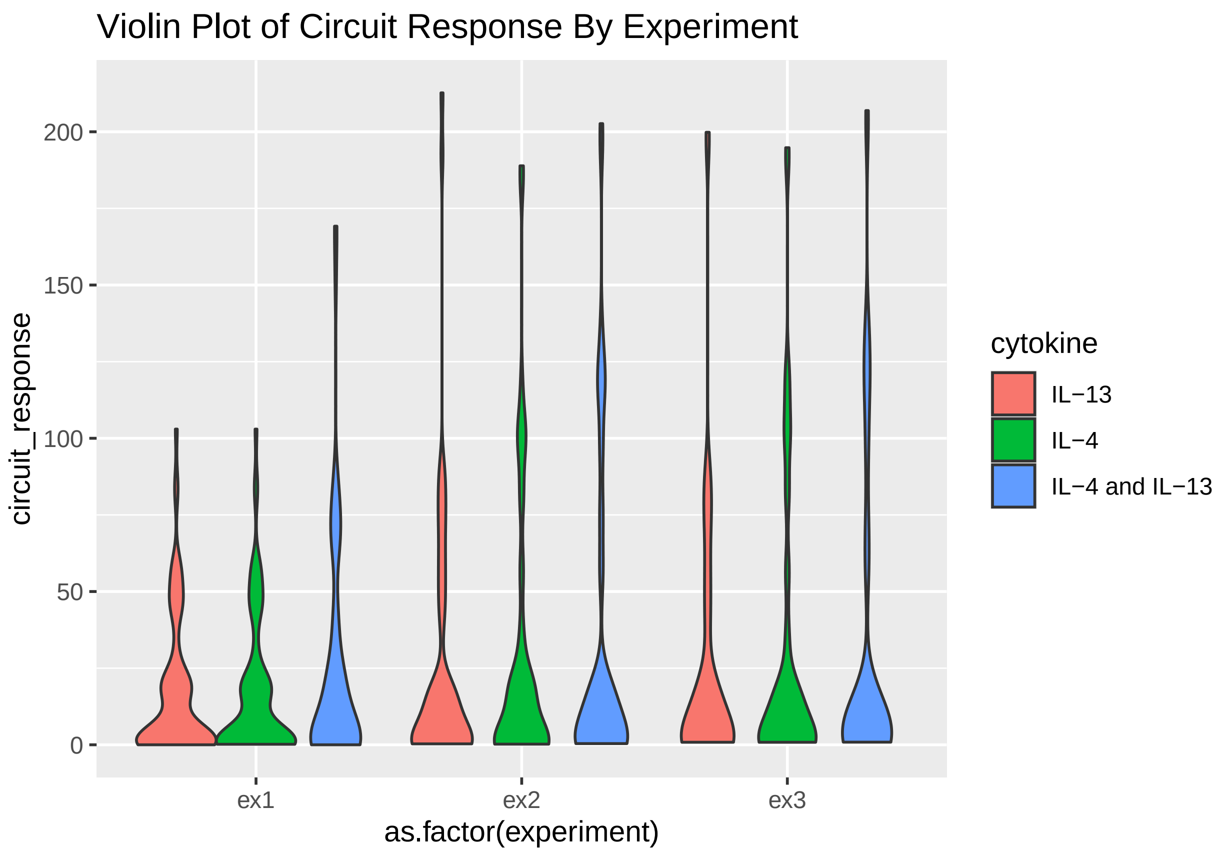
```
##  concentration          seap            experiment           figure
##  Min.   :  10.0   Min.   :  0.0000   Length:567         Length:567
##  1st Qu.:  20.0   1st Qu.:  0.9836   Class :character   Class :character
##  Median :  60.0   Median :  2.0028   Mode  :character   Mode  :character
##  Mean   : 187.1   Mean   : 25.1850
##  3rd Qu.: 100.0   3rd Qu.: 21.4677
##  Max.   :1000.0   Max.   :212.7027
##    cytokine
##  Length:567
##  Class :character
##  Mode  :character
##
##
##
```

```
colnames(data) <- c("concentration", "circuit_response", "experiment", "figure", "cytokine")
##data <- data %>%
#  mutate(C_normalized = (concentration - min(concentration)) / (max(concentration) - min(concentration
#          R_normalized = (circuit_response - min(circuit_response)) / (max(circuit_response) - min(circ
```

```
#what do the different experiments look like?
barplot(table(data$experiment), main="Experiment Count Histogram")
```



```
ggplot(data, aes(x=as.factor(experiment), y=circuit_response, fill=cytokine)) +
  geom_violin() +
  #stat_summary(fun.y=median, geom="point", size=2, color="red")+
  labs(title="Violin Plot of Circuit Response By Experiment")# (Red:Median)")
```

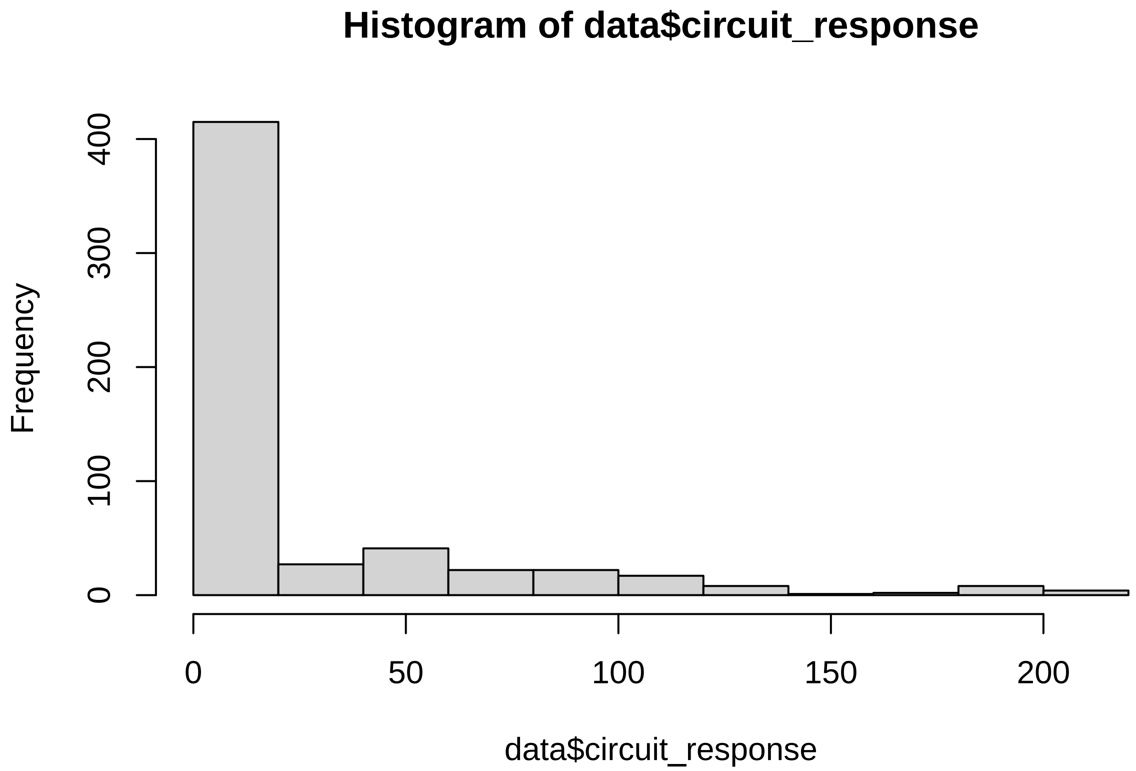


```
  #geom_bar(stat="identity")
  #geom_errorbar(aes(ymin=mean, ymax=mean+sd), width=.2,position=position_dodge(.9)) +
  #labs(x=bquote("Cytokine concentration (pg m"~L^-1~")"), y=bquote("SEAP (U "~L^-1~")"))
```

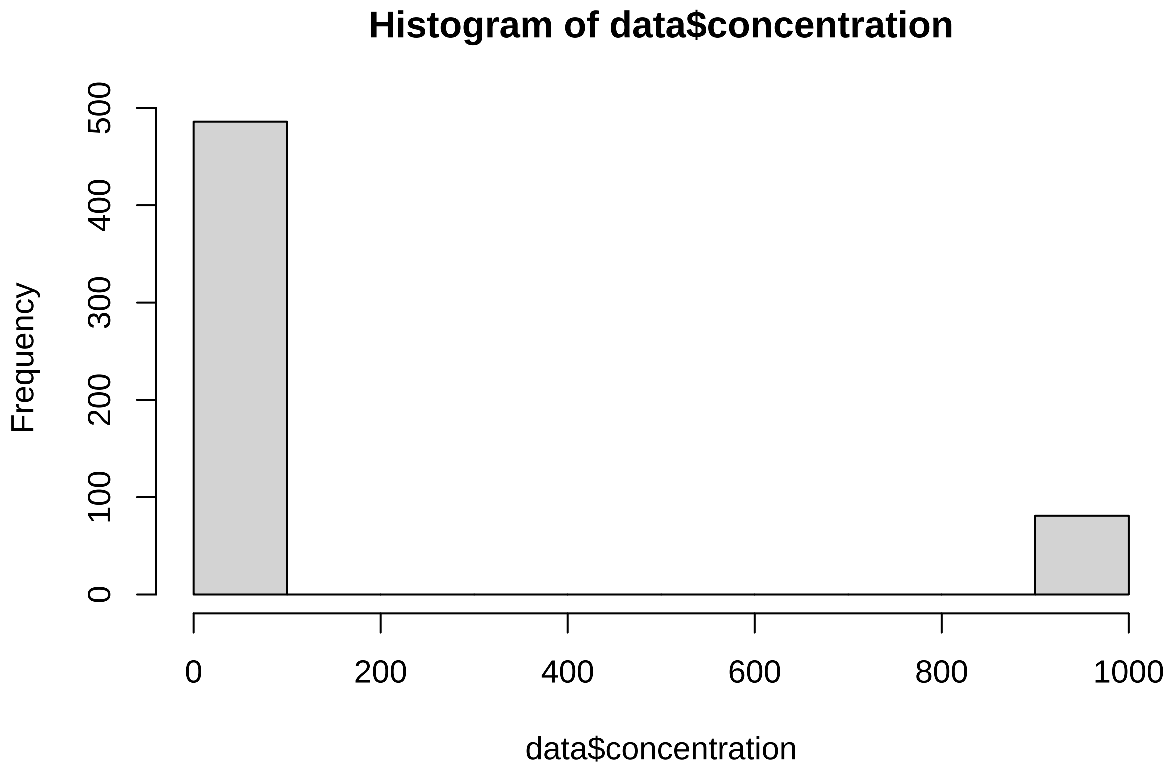
Since the distribution of the experiments differ, we should include the experiment number as a factor in the multivariate linear regression. Cytokine value is a little more difficult to identify visually. Revisiting figure 1d it looks like the observed differences might differ for each cytokine but that the observed pattern is the same.

Let's take a quick look at the distribution of our continuous variables:

```
hist(data$circuit_response)
```

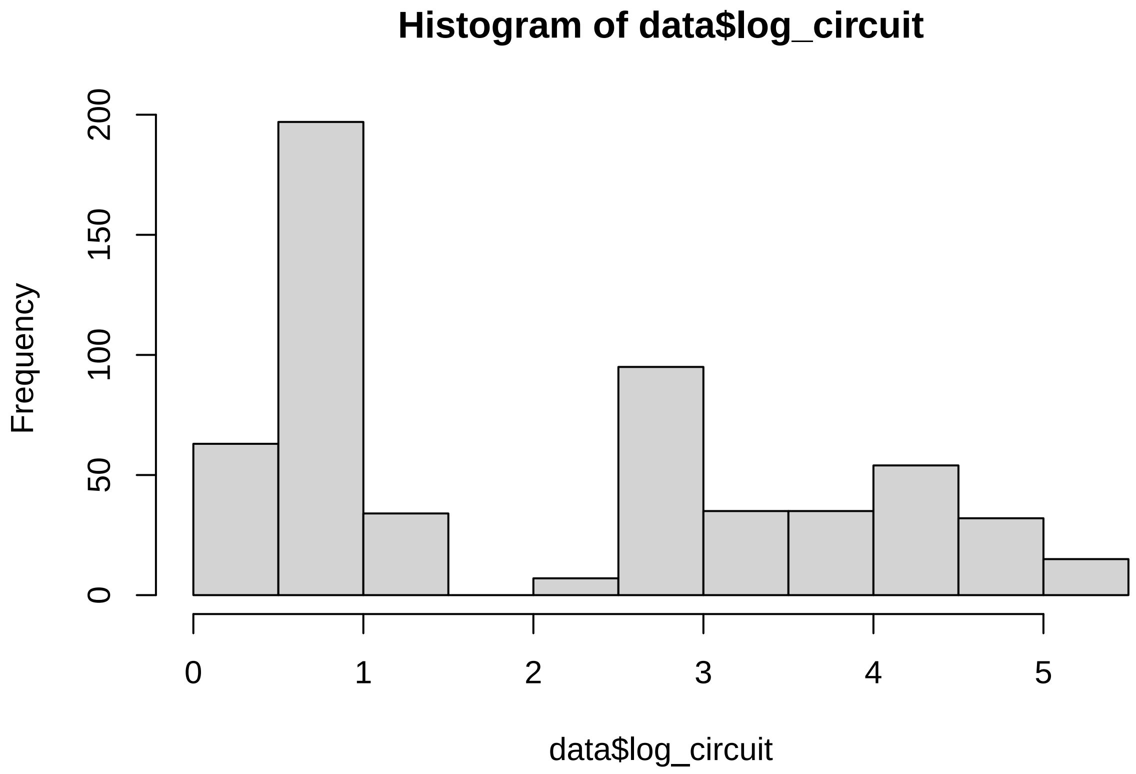


```
hist(data$concentration)
```

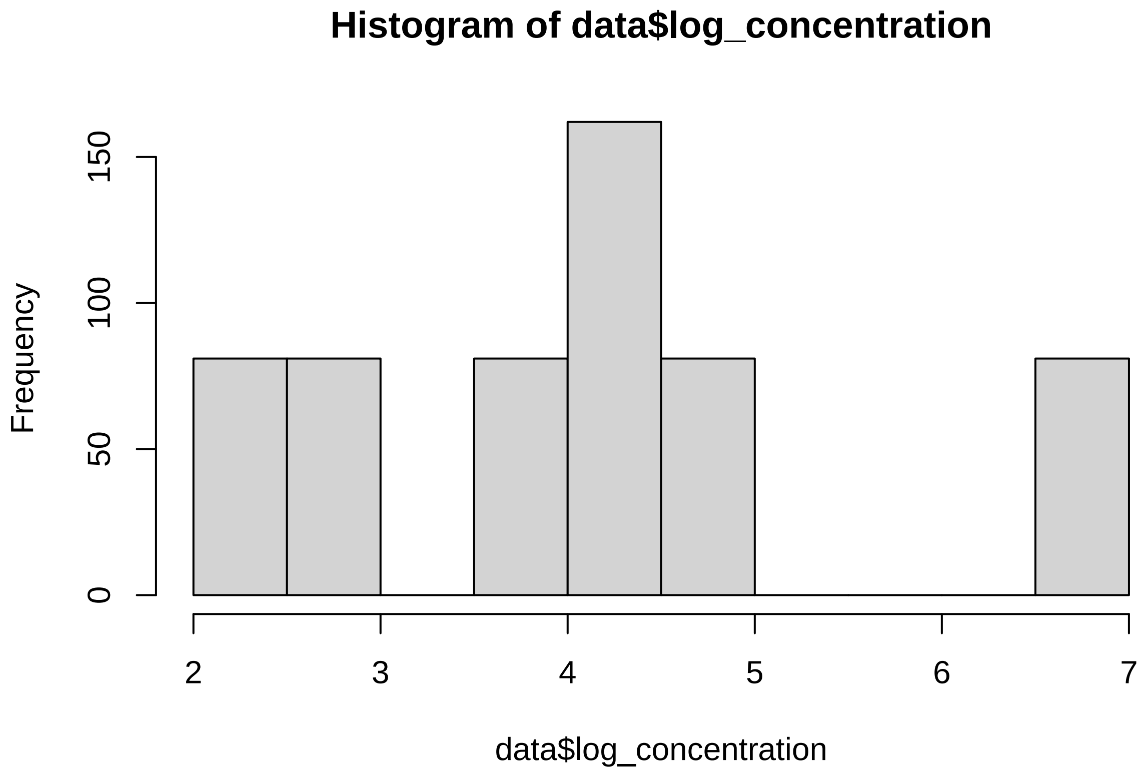


Both continuous variables are highly skewed. Would log transforming them be an option?

```
data$log_circuit <- log(data$circuit_response +1)# +1 due to zero values
data$log_concentration <- log(data$concentration)
hist(data$log_circuit)
```



```
hist(data$log_concentration)
```



Visually that doesn't seem to fix everything, but it's certainly better than before.

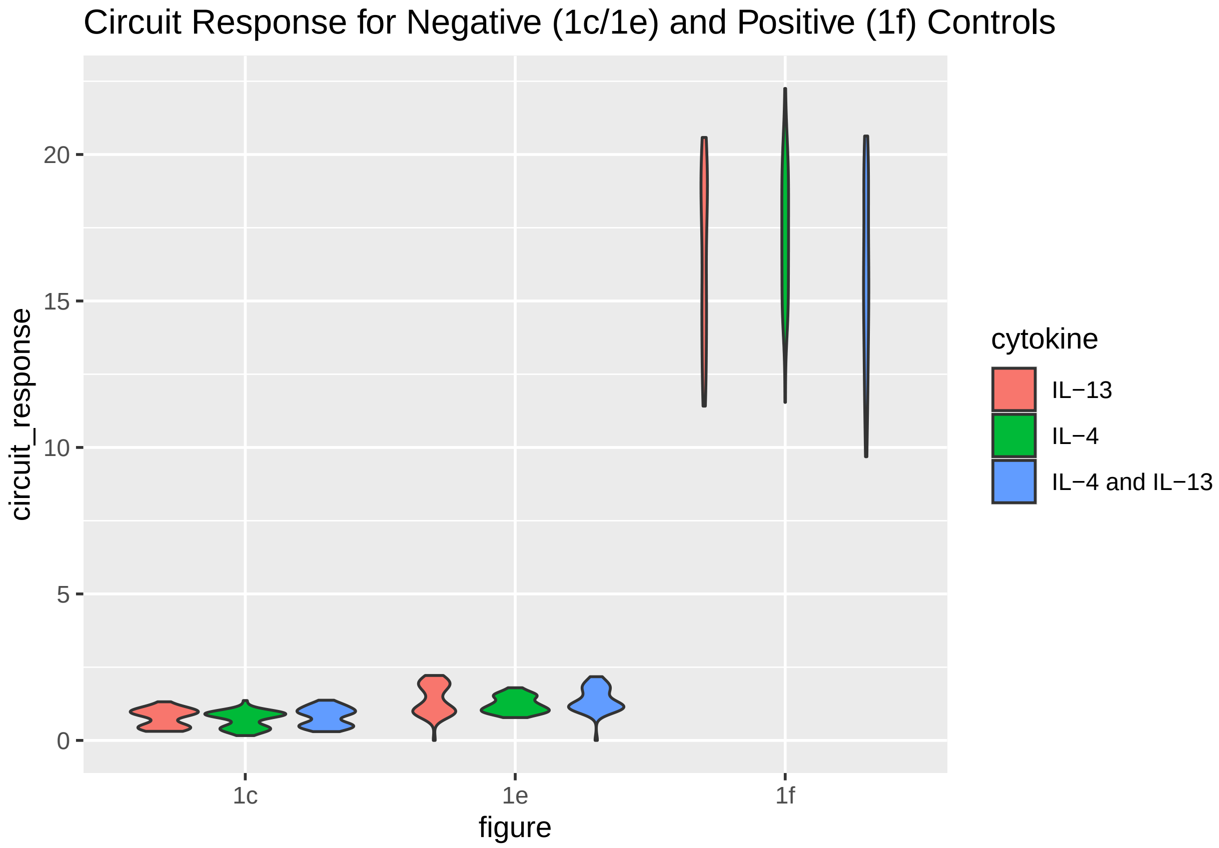
We'll try a multivariate linear regression with an interaction for concentration and experimental group with the genetic circuit (our experimental group) as the reference group with a constant term for experiment number. We will perform a post-hoc comparison of means to see if our positive and negative control groups behaved as we would have expected.

```
data$figure <- as.factor(data$figure)
data$experiment <- as.factor(data$experiment)
data$figure <- relevel(data$figure, ref = "1d")
#model <- lm(circuit_response~concentration*figure + experiment, data = data)
model <- lm(log_circuit~log_concentration*figure + experiment, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = log_circuit ~ log_concentration * figure + experiment,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.19870 -0.14672  0.02694  0.15701  0.55252
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.790817   0.068030  41.023  < 2e-16 ***
## log_concentration            0.344261   0.015362  22.410  < 2e-16 ***
## figure1c                    -2.318196   0.094619 -24.500  < 2e-16 ***
## figure1e                    -2.020105   0.094619 -21.350  < 2e-16 ***
## figure1f                    -0.054349   0.098499  -0.552  0.58133
## experimentex2                0.005841   0.025541   0.229  0.81919
## experimentex3                0.067822   0.025672   2.642  0.00848 **
## log_concentration:figure1c -0.329930   0.021725 -15.186  < 2e-16 ***
## log_concentration:figure1e -0.334814   0.021725 -15.411  < 2e-16 ***
## log_concentration:figure1f -0.322056   0.022621 -14.237  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2521 on 557 degrees of freedom
## Multiple R-squared:  0.9747, Adjusted R-squared:  0.9743
## F-statistic:  2383 on 9 and 557 DF,  p-value: < 2.2e-16
```

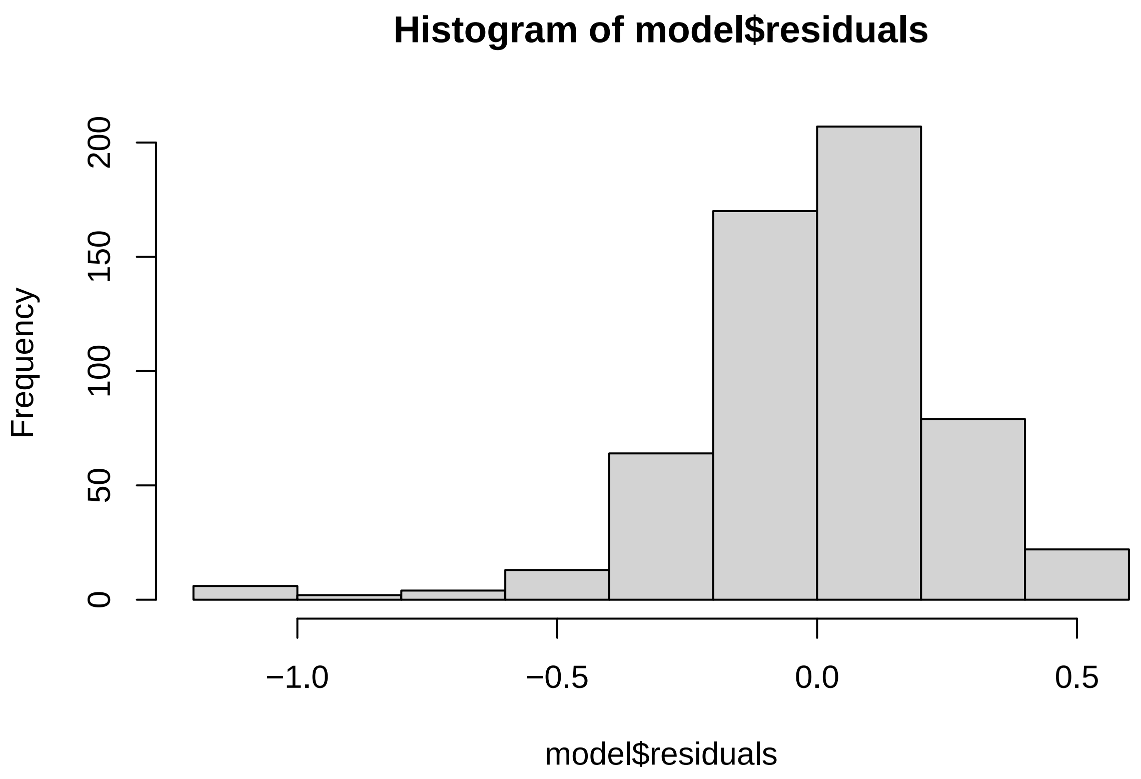
There is a significant effect for concentration within the experimental group in our model. There are significant differences in the constant terms for the negative controls, showing differences for our experimental group with the genetic circuit. The difference between the positive control and the genetic circuit was non-significant. Did the positive control behave differently than the negative controls as hoped?

```r
data_controls <- subset(data, data$figure != "1d")
ggplot(data_controls, aes(x=figure, y=circuit_response, fill=cytokine)) +
  geom_violin() +
  #stat_summary(fun.y=median, geom="point", size=2, color="red")+
  labs(title="Circuit Response for Negative (1c/1e) and Positive (1f) Controls")# (Red:Median)")
```

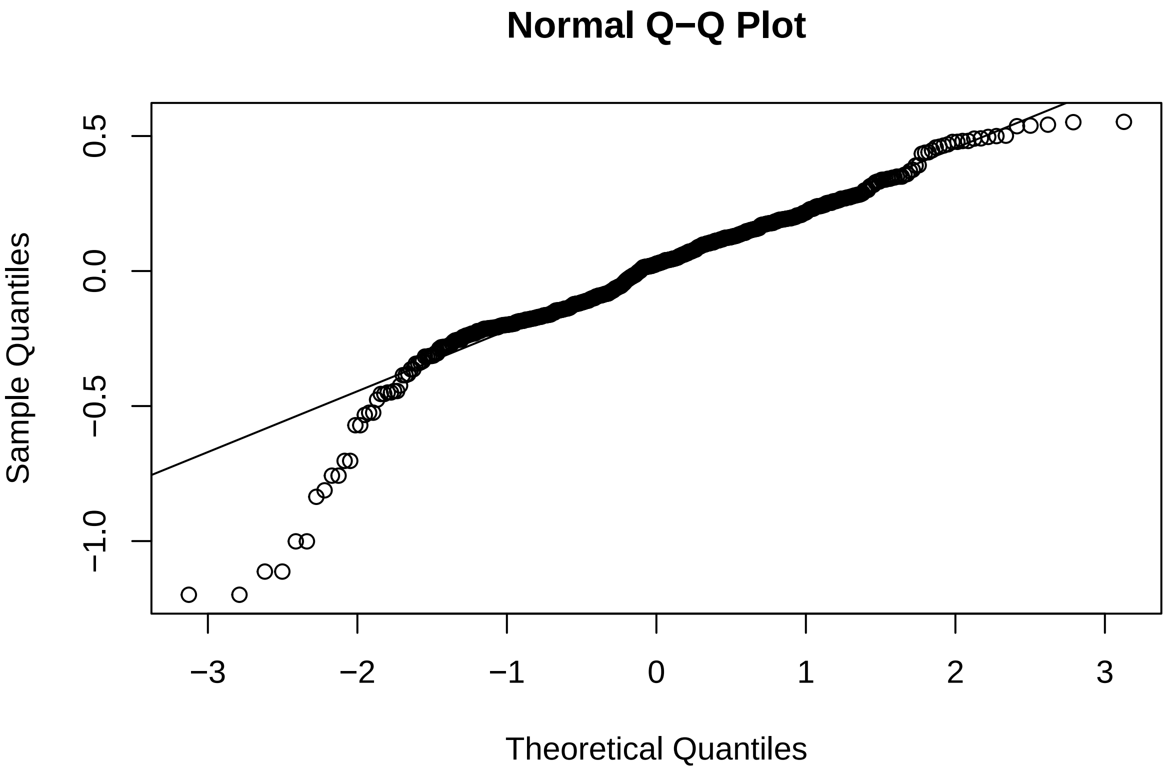


The positive and negative controls behaved as expected. Let's look at the residuals.

```r
hist(model$residuals)
```



```r
qqnorm(model$residuals)
qqline(model$residuals)
```



The qqplot looks fairly acceptable once we log-transform our continuous variables. There is a large negative tail that isn't being modeled well.

Statistical Methods: In order to measure the effect of concentration on circuit response for an experimental, a positive control, and two negative control groups over three different experiment settings, we modeled the data with a multivariate linear regression. As the dependent variable circuit response was skewed and the independent variable concentration was not uniformly distributed, we regressed the log-transformed circuit response against an interaction term between log-transformed concentration and experimental group (figures 1c/1d/1e/1f) with a constant adjustment for experiment number (1, 2, or 3).

Summary: There was a significant effect for log concentration in the experimental (genetic circuit) group. The coefficient for log-concentration of 0.34 indicates that for every one-unit increase in concentration, the circuit response in the experimental group increases by a factor of 1.4 ($e^($0.34)).