# IMRaD Report

Taryn H

2023-05-23

## Introduction

An experiment was conducted where the gene expression was measured for a collection of individuals for various concentrations, treatments, cell lines and gene lines. The key variables were:

- Concentration: an integer value between 0 and 10
- Treatment: Patients either received placebo or Activating Factor 42
- Cell line: The patient's cell line was either Wild Type or Cell Type 101
- Gene line: The gene lines were GL-bNo, GL-CsE, GL-fUg, GL-Hoe, GL-jEK, GL-JZC, GL-Rza and GL-xpo.

The goal of the research was to predict gene expression using the concentration, treatment and cell line the patients received and gene lines they had.

## Method

- The data was provided in an excel spreadsheet with multiple tabs containing the data. From each tab, the data was collated into one spreadsheet and loaded into R using the read_csv() command.
- Using EDA and analysis plots, it was identified that there was a data value with a gene expression less than 0. This data point was removed to ensure it didn't effect the trend and relationships found in the data.
- Prior to modelling the data, a side by side scatter plot, separated by cell line was produced. Concentration was on the x axis and gene expression on the y axis. The plot was coloured by treatment.
- To fit a model to the data, a random intercept model was chosen and modelled using the lmer function from the lme4 package (Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015)). Two models were considered, a model with no interaction terms and a model with a 3 way interaction. The random effect in both models was gene line.
- To evaluate the best model, the ranova() function from the lmerTest (Kuznetsova A, Brockhoff PB, Christensen RHB (2017)) package and the anova() function were utilized. To evaluate the two models, the p-values from the anova analysis were considered and predictors with a p-value less than 0.05 were considered significant.
- To further evaluate the validness of the model, residual vs fitted plots were also created using the plot() function.
- Based off the p-values and the residual vs fitted plots, the best model was the 3 way interaction model.
- The data was plotted using a scatter plot of concentration and gene expression. The predicted model was then plotted onto the created scatter plot using geom_line() and the predict() function.
- This plot was then recreated using facet_wrap() which split the graph into 8 individual graphs sorted by gene line. Each small plot modelled gene expression vs concentration and was coloured by gene line.

## Results

Before beginning the analysis to develop a model to predict gene expression, it was important to first visualize the data to look for obvious patterns between gene expression and the various predictors.

Figure 1 was produced which demonstrates the relationships within the data. It can be seen that for each cell line, there are 2 clear trends. For the cell line, Wild Type, the values of gene expression which also received the treatment, Activating Factor 42, had a larger gene expression for greater values of concentrations where patients who received the placebo treatment had less variation in gene expression as concentration increased. For patients of the cell line, Cell-type 101, the values of gene expression for patients who received the placebo were generally lower than those who received the Activating Factor 42 treatment.

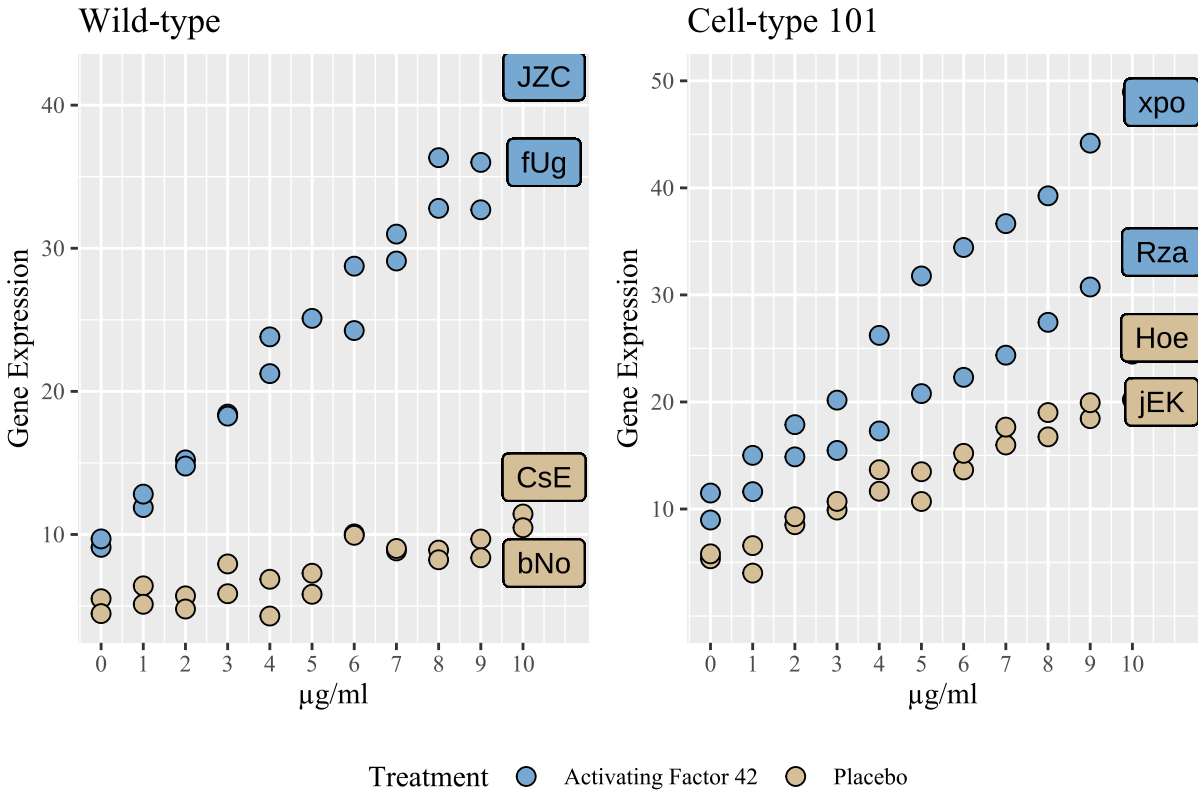


Figure 1: Scatter plot of gene expression vs concentration, split by cell line and coloured by treatment.

To attempt to accurately model the relationship between gene expression and the predictors concentration, cell line, treatment and gene line, two random intercept models were considered. The first was a random intercept model with gene line as the random effect and predictors concentration, cell line and treatment. Gene line was chosen to be the random effect as it was a "variable" in the experiment which couldn't be recreated or reproduced.

A plot of the residuals vs fitted plot was obtained for this model which clearly demonstrated that this model was not suitable as there was significant variation in the residuals, they did not have random scatter.


Figure 2: Scatter plot of Residual vs Fitted plot for model 1. The residuals show a clear distinct pattern and are not scattered randomly and demonstrate the model's poor fit to the data.

It was noted that, using the ranova function from the lmerTest package, the random effect was statistically significant with a p-value less than 0.05.

The next model considered was a 3 way interaction model between concentration, cell line and treatment. Gene line was included as a random effect. Again, the ranova shows that the random effect was statistically significant. Utilizing an anova analysis, the three way interaction term was shown to be statistically significant and thus this model was chosen as the final model. No smaller models can be considered given that the three way interaction term is significant, no other interaction terms can be removed to simplify the model. Below the coefficients of the final model have been displayed for the fixed effects and the random effects.

Table 1: Coefficients of Random Effects for each gene line

| gene line | Coefficents |
|---|---|
| GL-bNo | -0.5448884 |
| GL-CsE | 0.5448884 |
| GL-fUg | -0.9801050 |

| | |
|---|---|
| GL-Hoe | 0.9171917 |
| GL-jEK | -0.9171917 |
| GL-JZC | 0.9801050 |
| GL-Rza | -4.3688927 |
| GL-xpo | 4.3688927 |

Table 2: Coefficent of Fixed Effect Terms in Model 2

| Term in Model | Coefficient Value |
|---|---|
| Intercept | 9.91750000 |
| concentration | 3.05140909 |
| Cell_LineWild Type | -0.36156344 |
| treatmentPlacebo | -4.92159091 |
| concentration:Cell_LineWildType | -0.12145455 |
| concentration:treatmentPlacebo | -1.40550000 |
| CellLineWild Type: treatmentPlacebo | 0.08179071 |
| concentration:Cell_LineWildType:treatmentPlacebo | -0.96740909 |

To visualize this model, values of gene expression were predicted from the chosen model using the predict() function and figure 3 was produced which compares the predicted values of gene expression to the measured values.

The model has predicted the value of gene expression for each gene line. It can be seen that overall, the model does a good job of predicting the values of gene expression and follows the general trend of gene expression for each gene line.

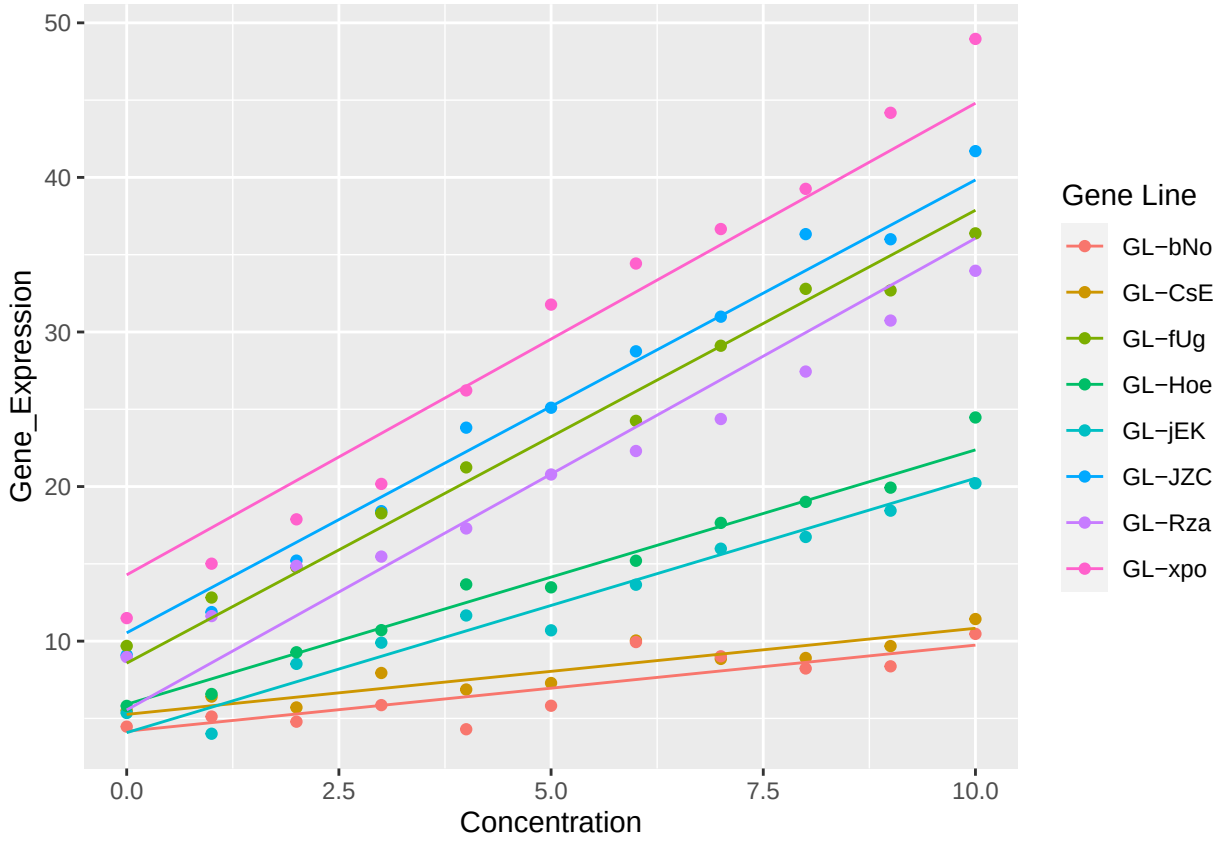


Figure 3: Scatter plot of chosen model compared to the measured values. The straight lines show the predicted gene expression values for each gene line and the points show the measured values. The graph has been coloured by gene line.

To further assess the fit of the model, figure 3 has been re-displayed as 9 individual plots to allow for a clearer view of the fit for each gene line.

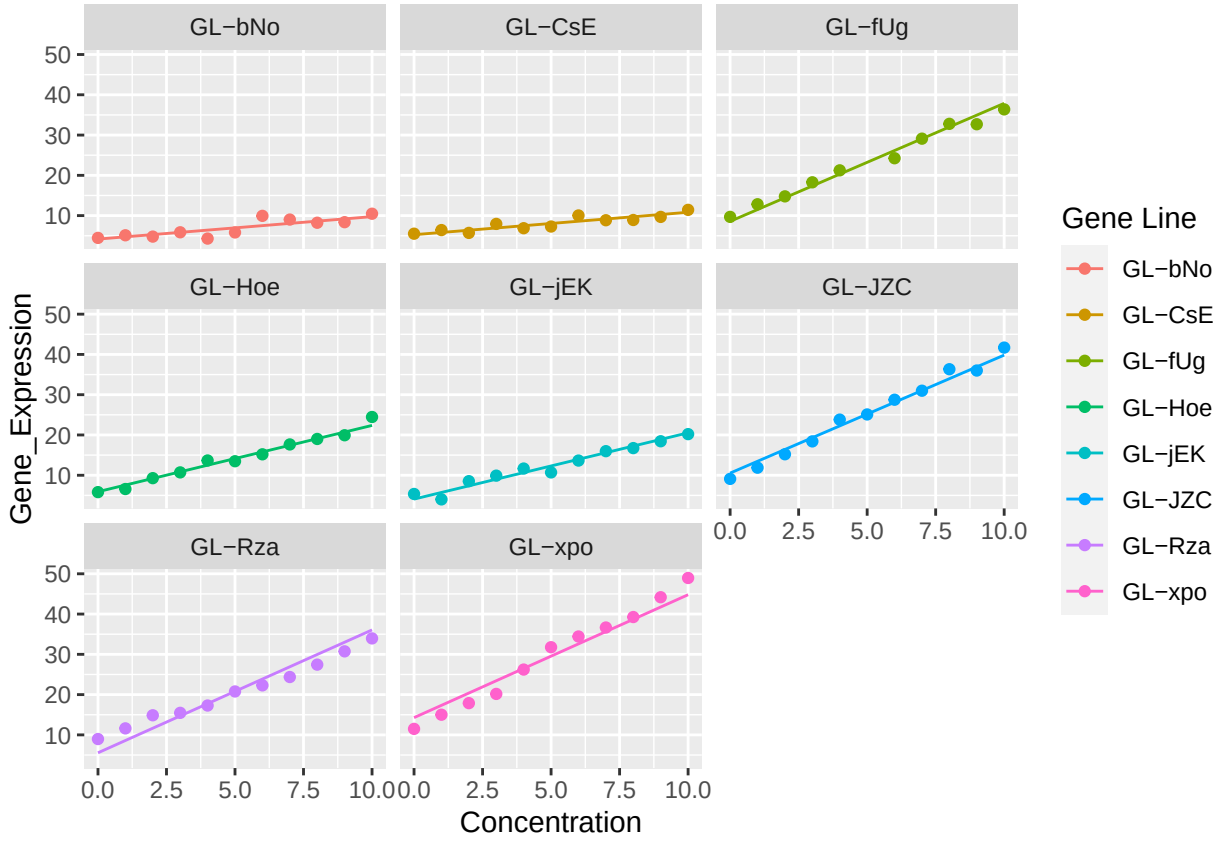


Figure 4: Scatter plot of gene expression vs concentration seperated by gene line. The predicted values from the chosen model have been been displayed as straight lines

## Discussion

From figure 4, it can be seen that the chosen model, closely approximates the data for most of the gene lines. The trend of gene expression for the genes GL-Rza and GL-xpo look as though they have more curvature suggesting that the random intercept model may not be the best fit for those two genes. The values of gene expression for all other gene lines appear to be randomly scattered above and below the prediction line supporting the choice of model.

It can be seen from Figure 3, the rate of increase in gene expression as concentration increases varies significantly. Gene line GL-bNo has a much smaller increase in gene expression when compared to gene GL-xpo. It appears that the four genes GL-xpo, GL-Rza, GL-fUg and GL-Rza generally have large values of gene expression compared to the other four genes. It can also be noted that the genes GL-xpo, GL-Rza, GL-fUg and GL-Rza also appear to have a greater increase in gene expression as concentration increases when compared to the other genes.

Overall, there is a clear relationship between gene expression and the predictors concentration, cell line, treatment and cell line. This model is one way of modelling the relationships demonstrated in the data. Other models could be considered in attempt to determine the ultimate fit.