Date: 30 May 2024

Predictive analysis of gene expression

Written by Dhwani Patel

Introduction

An experiment was conducted to capture the gene expressions for two different cell lines with a new treatment of *activating factor 42*. The gene expressions were also recorded with saline acting as the placebo. Gene expressions help understand the process of information translation into function and therefore it is important to identify if the new treatment effects the levels of gene expressions and if there is a difference between the two cell types.

The dataset contains three variables: cell line (two levels), treatment (two levels) and concentration. There are a total of 88 samples collected from the experiment (44 for each cell line). An identifier has been given to each observation group from the sample.

In this report we first conduct exploratory data analysis to investigate the relationships between concentration and gene expressions for each cell type, with and without treatment, followed by a mixed effects predictive model for the gene expression.

Methods

We cleaned the dataset to keep consistent labels for each level under the categories of cell line, treatment and name using the function *str_to_title* in R. The package used for this was the *stringr*. This ensured that R did not discriminate between labels due to case sensitivity. Initial exploratory analysis was done to visualise the dependencies using ggplots in R. We grouped the data firstly by cell lines and then by the identifiers of the group in the dataset. Each group was then classified as either using the new treatment activating factor 42 or placebo from the data. Plots of concentration versus gene expressions were created individually for each cell line for their respective groups. The growth rate of each group under the influence of the new treatment was calculated for each group using the formula

$$\frac{geneE_i - geneE_{i-1}}{c_i - c_{i-1}}, for\ 1 \leq i \leq 10$$

where $geneE_i$ is the value of the gene expression when concentration for concentration c_i.

The data involved repeated measurements of the gene expressions at different concentration levels. It contains clusters of measurements as identified by the different groups identified. These properties of the dataset suggested the use of a mixed effects predictive model where more than one source of random variability (Love, 2024) is allowed. Using random intercepts in this model supported the variability of the outcome for each group. The predictors for the gene expression were concentration, treatment and cell line. The model and results were obtained using the libraries *lme4* and *sjPlot*.

Results

Table 1 shows the different groups or clusters present in the data. There are 8 groups that are repeatedly measured for gene expressions at 11 different concentration levels. The name column stores the groupings of the clusters.

Table 1: Categorical variables

cell_line	treatment	name
Wild-Type	Placebo	Gl-Cdz
Wild-Type	Placebo	Gl-Xib
Wild-Type	Activating Factor 42	Gl-Rjs
Wild-Type	Activating Factor 42	Gl-Xik
Cell-Type 101	Placebo	Gl-Cwn
Cell-Type 101	Placebo	Gl-Kyh
Cell-Type 101	Activating Factor 42	Gl-Mfa
Cell-Type 101	Activating Factor 42	Gl-Zhw

Figure 1 shows the variation in trends of the gene expressions between all different groups as concentration is increased from 0 to 10 $\mu g/ml$. In cell line called *Wild Type*, the groups Gl-Xik and Gl-Rjs with the treatment of the activating factor 42 had higher gene expression values than groups Gl-Xib and Gl-Cdz that had the placebo. As the concentration increased, the cells with placebo maintained the same gene expression whereas the activating factor 42 allowed the gene expression value to grow. In the cell line *Cell Type 101*, the gene expression for the group Gl-Mfa was affected more by the new treatment than the group Gl-Zhw. The groups Gl-Cwn and Gl-Kyh for this cell line that had the placebo had a growing gene expression value as concentration was increased.

Date: 30 May 2024

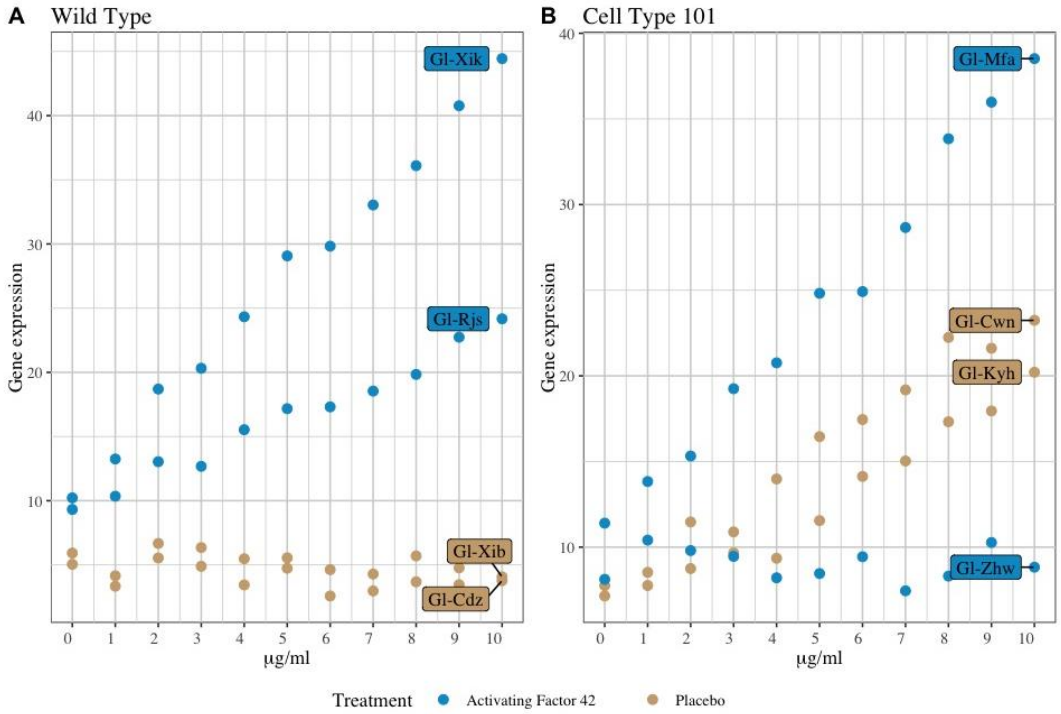

Figure 1: Relationship between concentration of cells and gene expression for each group under the effects of the new treatment activating factor 42 and placebo.

It was clear from Figure 1 that at least one group from each cell line was receptive to the new treatment. To quantify the growth of the gene expression as concentration increases with the new treatment, Table 2 calculates the rate of change of gene expression for groups Gl-Mfa, Gl-Rjs, Gl-Xik and Gl-Zhw for $c_i, 1 \leq i \leq 10$. The growth rate being close to 0 for Gl-Zhw shows that the treatment was not affected by the concentration.

Table 2: Growth rate of gene expression with treatment activating factor 42.

Concentration	Gl_Mfa	Gl_Rjs	Gl_Xik	Gl_Zhw
1	5.71	1.03	3.03	-0.99
2	1.49	2.69	5.45	-0.61
3	3.93	-0.36	1.62	-0.34
4	1.51	2.85	4.01	-1.25
5	4.06	1.64	4.74	0.25
6	0.10	0.14	0.76	0.98
7	3.74	1.23	3.21	-1.99
8	5.18	1.30	3.06	0.86
9	2.14	2.90	4.67	1.96
10	2.54	1.43	3.67	-1.44

The mixed affects model was run on the data with the concentration, treatment and cell line as predictors for the gene expression. As we saw in the figure, there is variability in the trajectory of the gene expression with and without treatment between both cell lines. Hence, we set the group names as the random effects. Each group appears once, and therefore random slopes were not required as the variance of the effect of treatment vs placebo for each group would not be possible to conclude for. Table 3 shows the results of the predictive model (using standard maximum likelihood estimation).

Table 3: Results from the mixed model with random effect for intercepts (formula: gene_expression ~ conc + treatment + cell_line)

	gene_expression		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	13.98	6.61 – 21.34	**<0.001**
conc	1.27	0.98 – 1.56	**<0.001**
treatment [Placebo]	-9.87	-18.20 – -1.53	**0.021**
cell line [Wild-Type]	-2.18	-10.51 – 6.15	0.604
Random Effects			
σ^2	19.04		
$\tau_{00\ name}$	33.36		
ICC	0.64		
N_{name}	8		
Observations	88		
Marginal R^2 / Conditional R^2	0.446 / 0.799		

In Table 3, we have the coefficients (estimates), confidence intervals (CI) and p-values (p) for each predictor and the intercept. We also have the variance, marginal and conditional R^2 values for the model. The results tell us that at conc = 0, treatment = Activating Factor 42 and cell_line = Cell-Type 101, the intercept is at 13.98 (with 95% confidence interval between 6.61 and 21.34). The p-value is less than 0.001. The effect of the concentration (conc) is significant with coefficient of 1.27 (with 95% confidence interval between 0.98 and 1.56). The effect of treatment activating factor 42 is also significant on the gene expression which can be concluded from the predictor treatment [Placebo] coefficient estimate being negative and the p-value being less than 0.05. There is minimal difference between the cell lines and therefore are not as significant on the gene expressions at the other two as can be observed from the higher p-value.

Figure 2 is a residual plot that shows the differences between the measured gene expression value and the predicted value. The plot is a representation of the accuracy of the prediction for the data points. We can see that points can be both above and below the prediction equation with the maximum frequency between 0

and ± 2. The plot is not evenly distributed vertically and have a little bit of a pattern, which means that this predictive model is not the most accurate and that work can be done (such as using a nonlinear model) to improve the results.

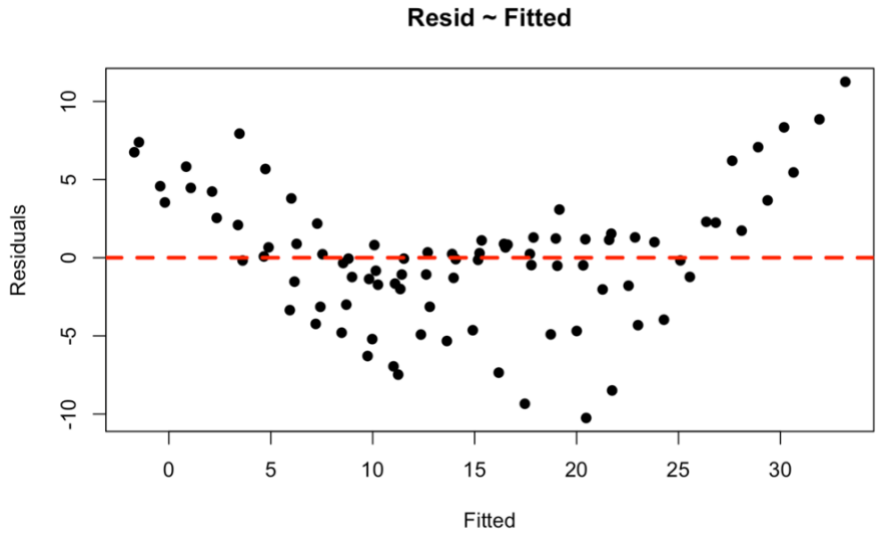

Figure 2: Model residual

Discussion

We looked at the data for gene expression measurements for two different cell lines under the influence of a new treatment at multiple concentration levels. From the initial analysis, we saw that increasing concentration increased gene expressions for most groups after being treated by the activating factor 42. The predictive model used in this analysis was a mixed affects model with random intercepts, which can be improved in future work. The results from the model however also suggested that the significant predictors were concentration and treatment and hence, we can conclude that the treatment can affect the growth factor of the gene expressions.

Bibliography

Clark, M. (no date) *Mixed models with R, Mixed Models*. Available at: https://m-clark.github.io/mixed-models-with-R/random_intercepts.html#exercises-for-starting-out (Accessed: 30 May 2024).

Love, K. (2024) *Understanding random effects in mixed models, The Analysis Factor*. Available at: https://www.theanalysisfactor.com/understanding-random-effects-in-mixed-models/ (Accessed: 30 May 2024).

R Core Team (2024) *The R project for statistical computing, R: A Language and Environment for Statistical Computing*. Available at: https://www.r-project.org/ (Accessed: 30 May 2024).