IMRaD Report

Shufan wang

June 8, 2024

1 Introduction

In this study, we aimed to investigate how different treatments and growth factors influenced and whether the new treatment had a significant influence on gene expression in cells. We used the gene dataset offered by Dr. Karl, which contains 88 observations and 5 variables. Among the variables, 'conc' (growth factor) and 'gene_expression' are numerical, while the remaining variables are categorical. Table 1 listed all the variables in the dataset.

Variable	Description
cell_line	Type of cells: Wild-type or Cell-type 101
treatment	Treatment used on cells: Activating factor 42 or Placebo
name	Name of cells. There are 8 different names in the dataset
conc	Concentration of growth factors
gene_expression	The response variables in this study

Table 1: The variables in the gene dataset.

2 Method

2.1 Data Cleaning, exploration and Pre-processing

After exploring the raw data, we found no missing values. However, the levels of the categorical variables were inconsistently formatted in terms of letter case. Dr. Karl confirmed these discrepancies were typographical errors and provided the correct formats for each categorical variable level. Consequently, we corrected the mistyped values to their appropriate formats.

After cleaning the data, we performed exploratory data analysis. We filtered the observations based on the same cell line and created scatter plots of gene expressions versus the concentration of growth factors. Different colours were used to distinguish observations with different treatments.

we normalized variables 'gene_expression' and 'conc' with means of 0 and variances of 1, to prevent the model from being strongly influenced by observations with large gene expression values and to keep the values of 'conc' as the same scale of the gene expression.

2.2 Mixed-effect models

We built 5 mixed-effect linear regression models with the gene data. The forms of the models were given in Table 2.

Models ‘me1’, ‘me2’ and ‘me3’ had random intercepts for variable ‘name’, while models ‘me4’ and ‘me5’ allowed the intercept and coefficient of the variables ‘conc’ randomly influenced by variable ‘name’.

Model ‘me1’ included ‘conc’ and ‘cell_line’ as predictors. In model ‘me2’, we added ‘treatment’ as a predictor. Model ‘me3’ further included the interaction between ‘treatment’ and ‘conc’. For models ‘me4’ and ‘me5’, both the intercept and the coefficient for ‘conc’ were allowed to vary randomly by ’name’. The difference in model "me5" is the inclusion of ‘cell_line’ as a predictor.

Model	Forms
me1	gene_expression ∼ conc + cell_line + (1 | name)
me2	gene_expression ∼ treatment + conc + cell_line + (1 | name)
me3	gene_expression ∼ treatment * conc + cell_line + (1 | name)
me4	gene_expression ∼ treatment * conc + (1 + conc | name)
me5	gene_expression ∼ treatment * conc + cell_line + (1 + conc| name)

Table 2: Mix effect models tried for the gene data.

We selected the model with the smallest AIC as our final model. We also performed the Likelihood ratio test with a significance level of 0.05 to check whether the model is significantly better than other simpler models.

3 Results

3.1 Exploratory Data Analysis

Figure 1 illustrated the relationships between gene expressions and the concentration of growth factors across different treatments, cell lines and names of cells. For each cell, gene expressions appear to have a strong linear relationship with the concentration of growth factors. However, the slopes and intercepts of the linear relationships varied among different cells. Moreover, by comparing Subfigure A and Subfigure B in Figure 1, we observed that the impact of treatments on gene expressions also varied across different cells.

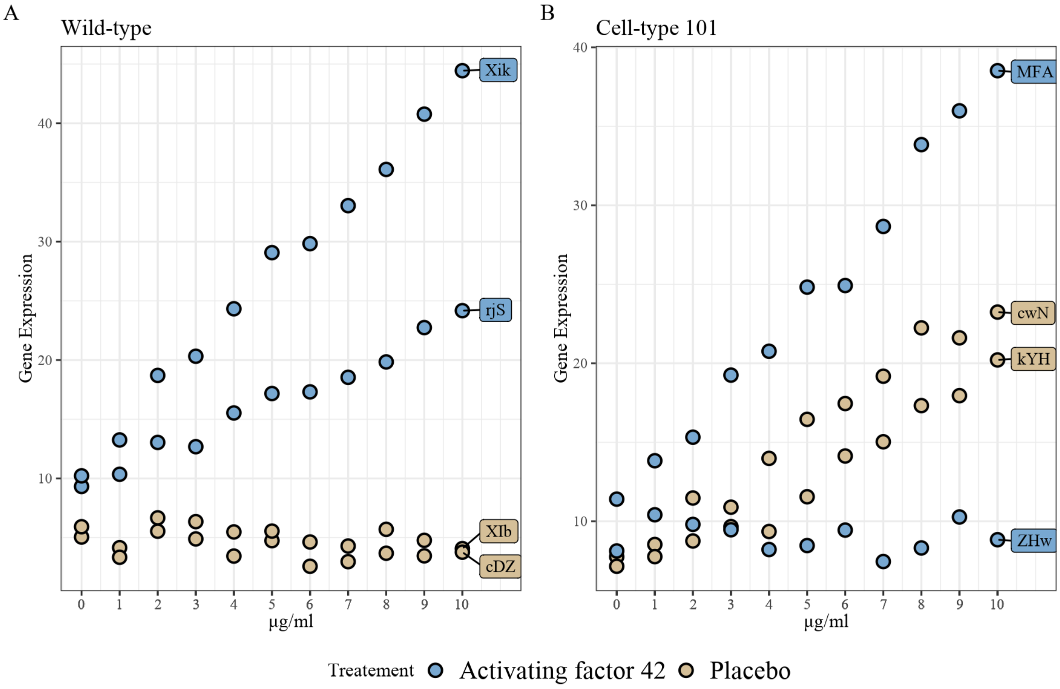

Figure 1: The relationships between gene expression and growth factors in each cell. They were strongly linearly correlated but the parameters varied with the name.

3.2 Model Performance and Selection

The model performance and the Likelihood ratio test results were displayed in Table 3, 'me4' achieved the best model performance, with AIC of -78.843. All models, except for 'me5' and 'me1', had p-values less than 0.05, indicating that they were significantly better than the previous model in the table. Therefore, we had no strong evidence that 'me5' performed better than 'me4' and 'me4' was significantly better than the remaining models. Model 'me4' was selected as the final model.

model	AIC	p value
me1	146.399	NA
me2	144.185	4.008489e-02
me3	128.007	2.011757e-05
me4	-78.843	2.448102e-47
me5	-78.411	2.105159e-01

Table 3: AIC and Likelihood ratio test results. Model 'me4' achieved the smallest AIC and all models were significantly better than their previous models, except for models 'me1' and 'me5'.

Figure 2 illustrated the linear submodels for different names of cells in our final model. In this final model, a linear regression was fitted to each set of observations with the same name. The data points for each name are closely and evenly distributed around their corresponding linear regression line, indicating that the random effect term effectively captured the influence of 'name' on the intercepts and slopes.

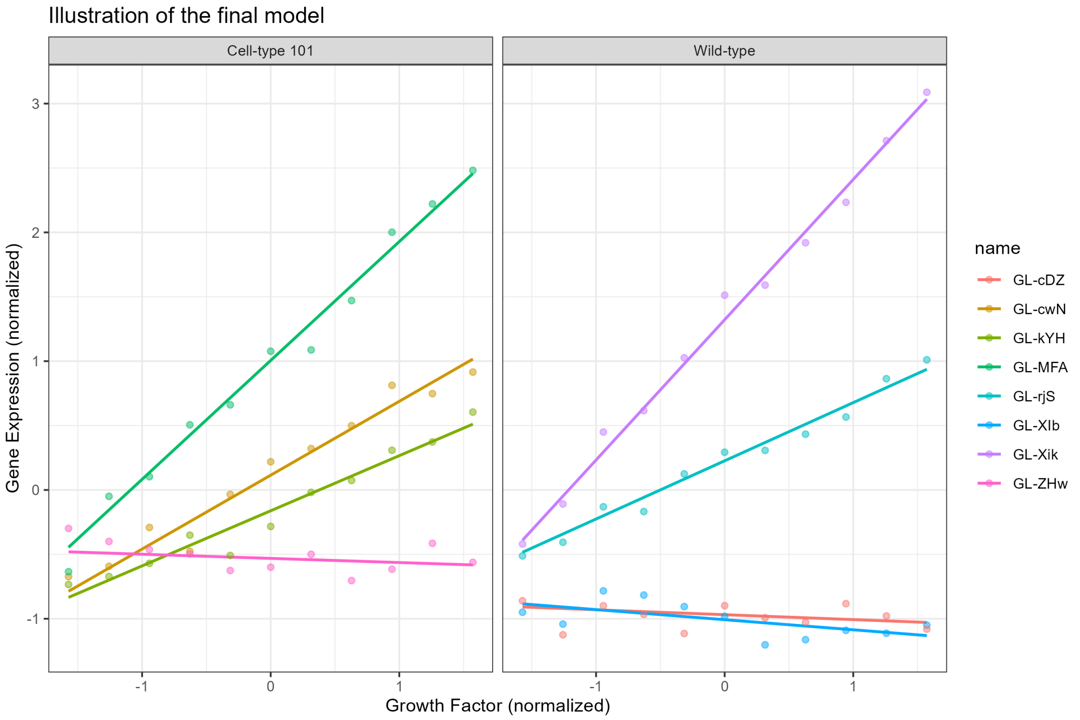

Figure 2: Final model. x-axis: normalized ‘conc’; y-axis: normalized ‘gene_expressions’. The final model fitted the data well.

Table 4 displayed the parameters of the linear models illustrated in Figure 2. According to it, all the values in columns ‘Placebo’ and ‘Placebo:Growth factor’ were negative, indicating treatment ‘Placebo’ and ‘Placebo:Growth factor’ had negative influences on the model’s prediction. In contrast, except for cell ‘GL-ZHw’, the predictor ‘conc’ had positive values. This means the predictor had a positive influence on the model’s predictions, indicating that the model would increase the values of prediction of gene expression when the concentration of growth factors increases. The cell is also the only one that had a negative intercept, but since we are modelling with normalized data a negative intercept does not mean the model is wrong.

		Coefficients		
Name	Intercept	Placebo	conc	Placebo:conc
GL-cDZ	0.043	−1.011	0.349	−0.387
GL-cwN	1.126	−1.011	0.960	−0.387
GL-kYH	0.850	−1.011	0.815	−0.387
GL-MFA	1.006	−1.011	0.923	−0.387
GL-rjS	0.226	−1.011	0.452	−0.387
GL-XIb	0.004	−1.011	0.308	−0.387
GL-Xik	1.322	−1.011	1.089	−0.387
GL-ZHw	−0.531	−1.011	−0.032	−0.387

Table 4: The coefficients of models for every name of cells (rounded to 3 decimals).

4 Discussion

This study investigated the influence of the new treatment ‘Activating factor 42’ on cells’ gene expressions. To address this, we cleaned the data, did exploratory data analysis and built mix-effect linear models for the cleaned and normalized data. Our final model took ‘treatment’, ‘conc’ and the interaction term between them as predictors. The intercept and coefficient of ‘treatment’ were influenced by the name. Both the model and exploratory data analysis results revealed that the relationships between gene expressions and the concentration of growth factors with different treatments were linear but strongly influenced by the cells themselves.

Moreover, we displayed the intercepts and coefficients with the random effect of name in Table 4. According to it, predictor "Placebo" and ‘Placebo:Growth factor’ had negative coefficients for all the cells. Because the final model took the treatment ‘Activating factor 42’ as the reference level, these results indicated that the new treatment positively influenced the values of gene expression for every cell.

The main limitation of this study is that we did not test whether every predictor in the final model had significant influences on the predictions. We can not directly test it because our random effect term included the predictors, making them become random variables rather than parameters. However, we performed Likelihood ratio tests between models to approximate it. According to the test, we had no evidence that "me5" is significantly better than "me4", so "cell line" might have had no significant influence on the predictions. In contrast, "me2" is significantly better than "me1", indicating "treatment" is a significant predictor to some extent. Therefore, we can conclude that the new treatment had a significant influence on gene expression to some extent.

5 Appendix

The project is implemented in R language. The used packages and versions are: "lme4 1.1-35.3" [1], "tidyverse 2.0.0"[4], "caret6.0-94"[3], "gt 0.10.1"[2]. All the code, figures and tables are stored in the Github repository: https://github.com/WangShufan0061/2024-5-29-Karl-collaboration-project.git

References

[1] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[2] Richard Iannone, Joe Cheng, Barret Schloerke, Ellis Hughes, Alexandra Lauer, and JooYoung Seo. *gt: Easily Create Presentation-Ready Display Tables*, 2024. R package version 0.10.1.

[3] Kuhn and Max. Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26, 2008.

[4] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.