Gene Expression IMRaD Report

Vivienne Niejalke

5/27/23

Introduction

The dataset contains 87 observations of an experiment testing the effects of a new treatment on gene expression. This data was provided by Dr Karl Berator, a research assistant at the Institute of -omics. The key variables were:

- treatment: the new treatment was compared with a saline placebo,
- concentration: an integer between 0 and 10, measured in $\mu g/ml$,
- gene line: there were eight specific genes, and
- cell type: each gene line was classified as Wild-type or Cell-type 101.

The aim was to investigate the effect of the growth factor on gene expression. This was evaluated by designing a predictive model for gene expression.

Methods

The data was analysed using the packages `lme4` (Bates et al. 2015), `lmerTest` (Kuznetsova, Brockhoff, and Christensen 2017) and `ggplot2` (Wickham 2016) in R (R Core Team 2022) in RStudio[1].

The data was manually preprocessed to combine the data from for each gene line into one dataset.

Firstly to better understand the data, an exploratory data analysis (EDA) was conducted to visualize the data using box plots and plots.

Then, a linear random mixed effects model with a random intercept term was fitted to the data, with concentration, treatment and cell type as predictors and gene line as the random intercept term. The joint interaction term is significant due to the p-value less than 0.05.

[1] https://www.rstudio.com

A linear random mixed effects model with a random slope was also fitted, with concentration, treatment and cell type as predictors and concentration per gene line as the random slope term.

The factors of the two models were assessed using ranova for random effects and anova for fixed effects. The models were compared using the AIC (Akaike Information Criterion) to select the best model.

Results

Figure 1 of the box plots show the gene expression of different gene lines and cell types. It shows that the treatment has a greater gene expression, however greater variability in the results.

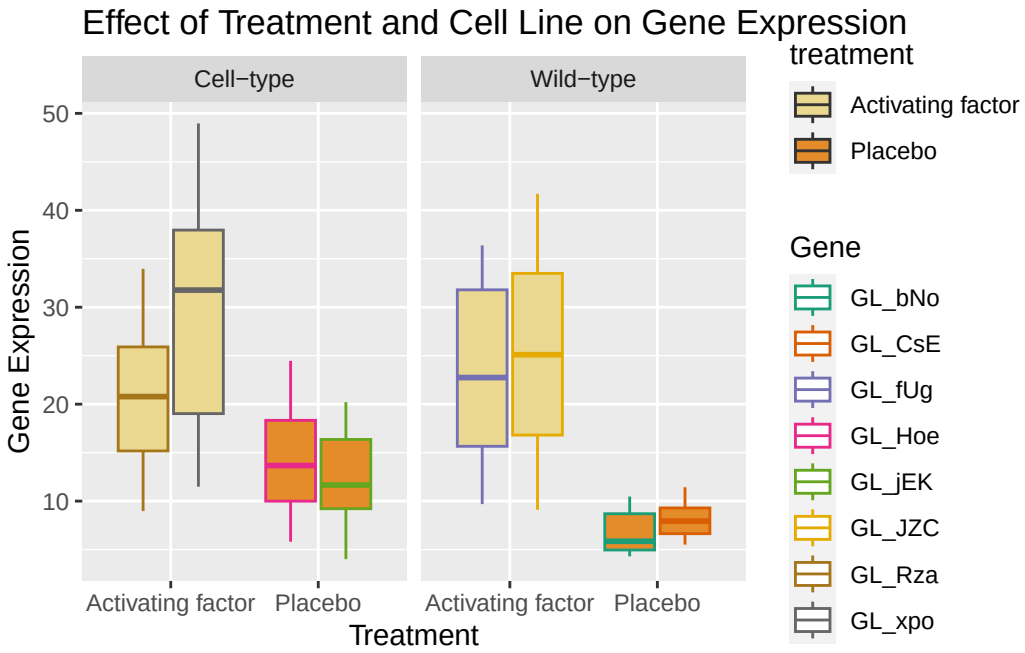

Figure 1: Box Plot of the Effect of Treatment and Cell Line on Gene Expression

The 'ranova' and 'anova' tests indicated that the models are suitable and all factors are relevent, especially due to the significance of the high order interaction terms. The table below indicates that 'model_2', the random slope mixed effects model with interaction terms is the better predictive model.

```
         df      AIC
model_1 10 369.1289
model_2 12 309.8771
model_3  7 318.7105
```

The selected linear mixed effects model with a random slope is plotted below as the chosen predictive model.This shows that the gene expression is not only greater for genes with the treatment, but also increases as the concentration increases.

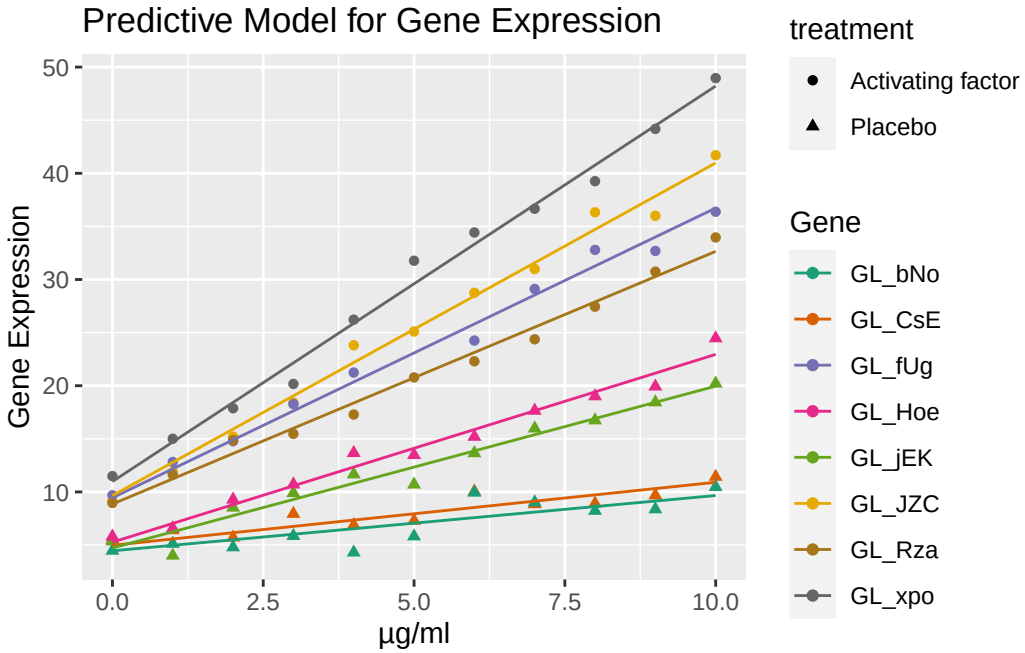

Figure 2: Linear Mixed Effects Random Model

Discussion

From the provided plots, it is illustrated that the treatment increased the growth factor of the gene expression for both cell types. Concentration, treatment and cell type are considered significant factors for the predictive model.

The AIC results indicate that a linear mixed effects model with a random slope is better than a linear mixed effects model with only a random intercept. This means that the correlated concentration and gene line random effect is important to consider.

References

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H. B. Christensen. 2017. "lmerTest Package: Tests in Linear Mixed Effects Models." *Journal of Statistical Software* 82 (13): 1–26. https://doi.org/10.18637/jss.v082.i13.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.