IMRaD Report

William Pincombe

Introduction

We are analysing data provided by Karl Berator. The data consists of the results of 8 trials where the gene expression was measured for 11 integer levels of concentration of growth factor between 0 and 10. Four trials each were carried out using the AF42 treatment or the placebo and similarly four trials were done on each of two different cell lines, Wild-Type (WT) and CT101. Two trials were carried out on each combination of treatment and cell line. Every trial was carried out on a different gene line.

There are five variables in the data:

- gene expression
- concentration of growth factor
- treatment: a factor variable with levels AF42 and placebo.
- cell line: a factor variable with levels WT and CT101.
- gene line: a factor variable with 8 levels: (CsE, bNo, JZC, fUg, jEK, Hoe, Rza and xpo).

We want to study the effect of treatment on the relationship between the concentration of growth factor and the gene expression.

Methods

We cleaned the data for analysis by saving it in a .csv file in a long format, with a column representing each of the five variables. There was one observation with a missing level of gene expression, at concentration 5 with cell line WT, treatment AF42 and gene line fUg. Since gene expression is our response variable, we excluded this observation from our data, leaving 87 observations for the analysis.

We are interested in the relationship of three predictor variables - concentration, treatment and cell line - with the response variable gene expression. However, we also must control for different gene lines used in each trial of the experiment. Therefore, we will use a mixed-effects model, with concentration, treatment and cell line as our fixed effects and gene line as a random effect.

tuning of model

We used the R statistical programming language (R Core Team 2023) for our analysis, using the tidyverse package (Wickham et al. 2019) to clean the data and the lme4 package (Bates et al. 2015) to fit the mixed-effects model.

Results

Firstly, we plot the observed values, shown in Figure 1. The treatment AF42 appears to have a substantial impact on both the intercept and slope of the relationship between concentration and gene expression, for both cell lines. The relationship also appears to be different between the cell lines, with both the treatment and placebo observing higher gene expression for the same concentration in cell line CT101, as compared to the Wild Type.

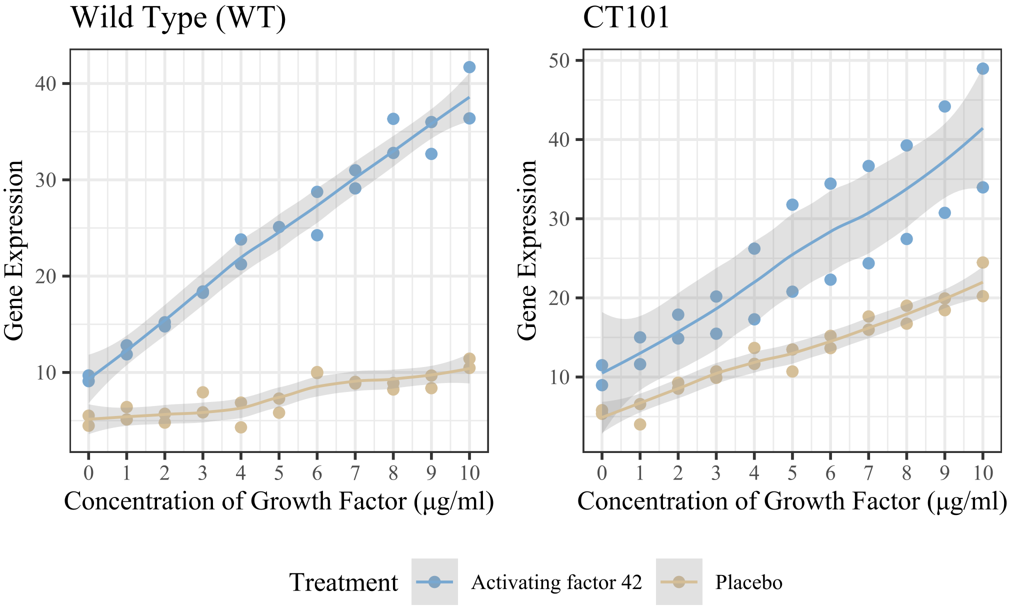

Figure 1: Scatter plot of gene expression for different concentrations of growth factor. The different cell lines are shown in side-by-side plots, and the treatment is indicated by the colour.

As a result of the relationships observed in Figure 1, we fit a model including concentration, treatment and cell line as predictors for gene expression, with interaction terms between concentration and the two categorical variables to allow the slope to vary. However, we know that each series of observations was done on a different cell line. This could have an effect on the gene expression, so we should include it in the model. However, we are not interested in this effect, we only want to remove bias from our estimates of the other coefficients, so we include it as a random effect in a Mixed-Effects Model. The fitted coefficients for the fixed effects in this model are given by Table 1.

Table 1: Summary of estimated coefficients in Mixed-Effects Model

	Estimate	Std. Error	df	t value	Pr(>|t|)
Intercept	9.8967	2.1084	5.6863	4.6939	0.0039
Concentration	3.2933	0.1053	76.0033	31.2840	0.0000
Treatment (placebo)	-4.8800	2.4348	5.6886	-2.0043	0.0945
Cell Line (WT)	-0.3213	2.4348	5.6886	-0.1320	0.8995
Concentration * Treatment (placebo)	-1.8892	0.1216	76.0033	-15.5420	0.0000
Concentration * Cell Line (WT)	-0.6052	0.1216	76.0033	-4.9785	0.0000

The results in Table 1 suggest that the there is a strong positive relationship between concentration and gene expression. However, the effect of treatment and cell line is more ambiguous, with the slope terms returning very low standard error and high test statistics while the intercept terms have p-values greater than the standard rule of thumb of 0.05. To test the overall effect of each factor variable, we fit models excluding these variables, and then compare to the complete model.

Table 2: Model metrics for the complete mixed-effects model, and with each of the fixed effect factor variables treatment and cell line removed.

Model	AIC	BIC
Complete	387.6244	407.3516
Without Cell Line	408.4662	423.2616
Without Treatment	507.5614	522.3568

The AIC and BIC of the different models is shown in Table 2. We can see here that the complete model, with both treatment and cell line, has the lowest, hence most optimal, AIC and BIC. The removal of treatment has a much larger negative effect on both metrics than the removal of cell line.

We are also interested in whether the random effects are necessary. Firstly, we can extract the values of the intercept under each level of gene line from the complete model. We subtract the overall, average, intercept for the model to show the difference in intercept for each random effect. This is shown in Table 3.

Table 3: Intercept under each level of gene line, the random effect, in the complete model. The difference to the overall intercept is also given.

	Intercept	Difference
bNo	8.197826	-1.6988968
CsE	9.281852	-0.6148705
fUg	10.077388	0.1806658
Hoe	11.965958	2.0692356
jEK	10.141254	0.2445317
JZC	12.029824	2.1331015
Rza	4.393999	-5.5027231
xpo	13.085678	3.1889558

We can test the significance of the random effect using the `ranova` command from the `lmerTest` package, which performs a likelihood ratio test. This fits a model without the random effect, and compares the log-likelihood. The results of this test are shown in Table 4, where we find that the reduced model has higher AIC and lower log likelihood. The likelihood ratio test rejects the null hypothesis that there is no difference between the log likelihoods of the models, so we take the better complete model.

Table 4: Results of a likelihood ratio test on removing the intercept random effect for gene line (GL) from the mixed-effects model.

	npar	logLik	AIC	LRT	Df	Pr(>Chisq)
Complete Model	8	-185.8122	387.6244	NA	NA	NA
Random Effect Removed	7	-224.6211	463.2421	77.61776	1	1.250408e-18

To check how well the final model fits the data, we can plot the fitted relationships above the scatter plot of the data from Figure 1, as shown in Figure 2.

Discussion

discussion of results in context of research question

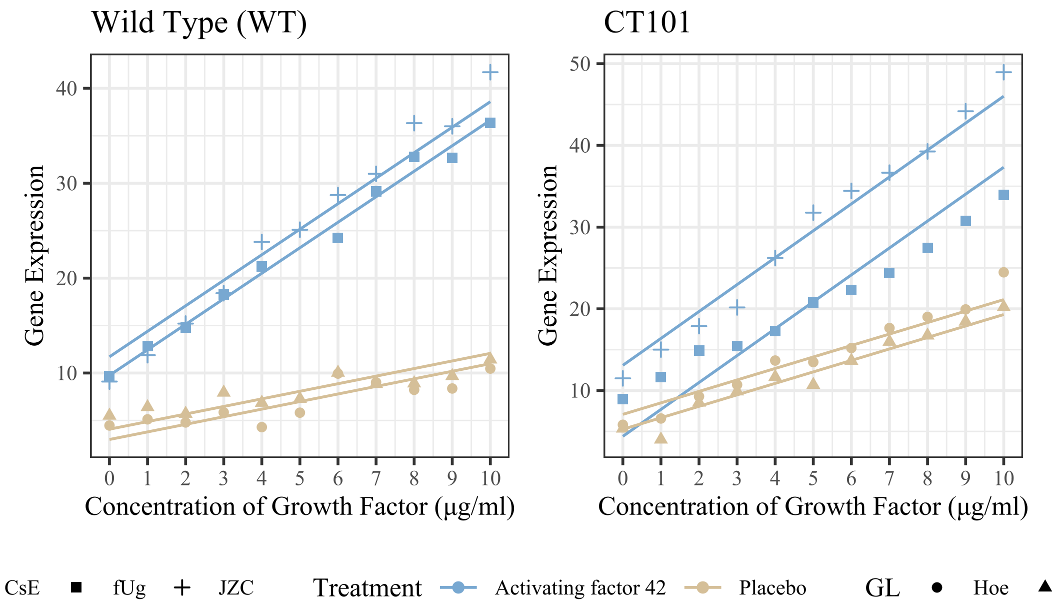

Figure 2: Scatter plot of the data with the fitted values from the complete mixed-effects model.

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.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.