Predicting gene expression from growth factor, cell line and treatment

Alexandra Stephenson

5/25/23

Introduction

The data set consists of eight gene lines, each with one of two cell lines (wild-type or cell-type 101), one of two treatments (the placebo treatment or activating factor 42), and eleven different concentrations of growth factor (recorded in mg/ml). Thus, for each pair of cell line and treatment, there are two gene lines, each with eleven concentrations of growth factor (from 0 to 10). Only one data point is missing, that of concentration 5 mg/ml for gene line GL-fUg (with cell line wild-type and treatment activating factor 42).

This report investigates the impact of growth factor concentration, treatment, and cell line on gene expression, as well as the effect of gene line.

Method

The data was cleaned and analysed using the R language (R Core Team 2022), and the packages , knitr (Xie 2023), lme4 (Bates et al. 2015), lmerTest (Kuznetsova, Brockhoff, and Christensen 2017), , performance (Lüdecke et al. 2021), readr (Wickham, Hester, and Bryan 2023), sjPlot (Lüdecke 2023), and tidyverse (Wickham et al. 2019). Any data points recorded as -99 were taken to indicate no data was recorded, or `NA`.

Exploratory data analysis was then conducted on the data, including plotting gene expression versus concentration, gene expression versus cell line, and gene expression versus treatment.

Several mixed effects models were then fit, and compared using Akaike's Information Criterion (AIC), R^2 values and root mean squared error.

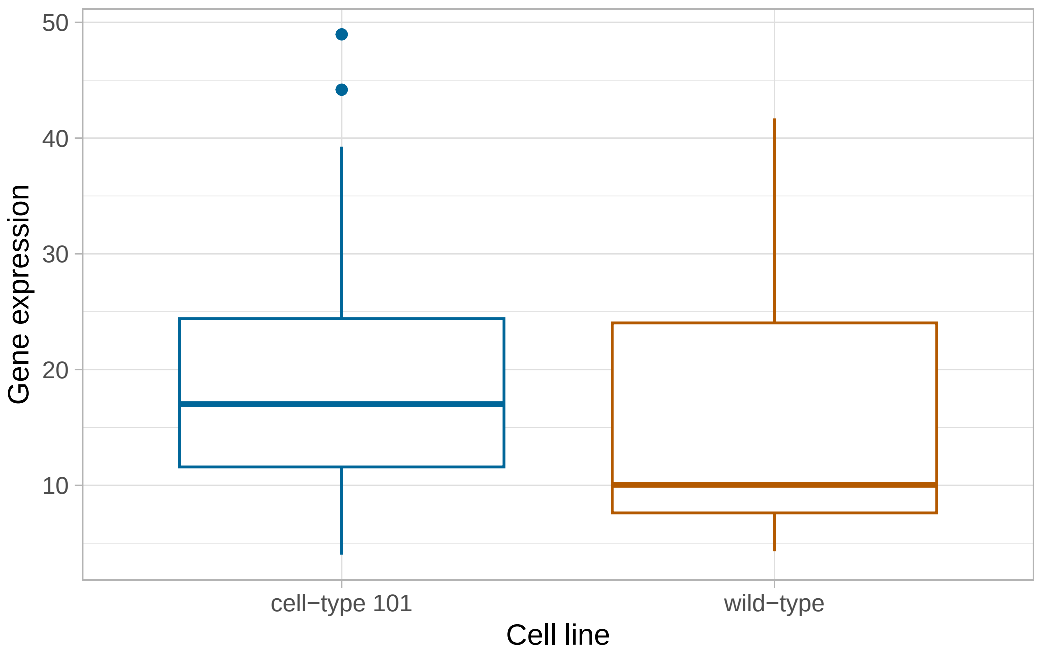

Figure 1: A boxplot of gene expression for each cell line (wild-type and cell-type 101).

Results

A boxplot of gene expression, grouped by cell line, is shown in Figure 1. From this boxplot, it can be seen that there does not appear to be a significant difference between gene expression for wild-type and gene expression for cell-type 101. This suggests that cell line may not be a predictor of gene expression, or at least, not on its own.

Figure 2 shows a boxplot of gene expression for each treatment type (placebo or activating factor 42). From this boxplot, it can be seen that there does appear to be a difference between gene expression for placebo and gene expression for activating factor 42. This suggests that treatment is a predictor of gene expression.

The data is plotted in Figure 3, with gene expression on the y axis and concentration on the x axis, with the data points coloured by cell line and treatment. From this plot, it can be seen that there does appear to be a relationship between concentration and gene expression, which suggests that concentration is a predictor of gene expression. From Figure 3, it can be seen that there appear to be differences between the pairs (cell-type 101, placebo) and (wild-type, placebo) and the other two pairs of cell line and treatment. However, there does not appear to be a difference between (cell-type 101, activating factor 42) and (wild-type, activating factor 42). This suggests that for the placebo treatment, cell line has an impact on gene expression, but for the activating factor, cell line may not have an impact on gene expression. Thus, cell

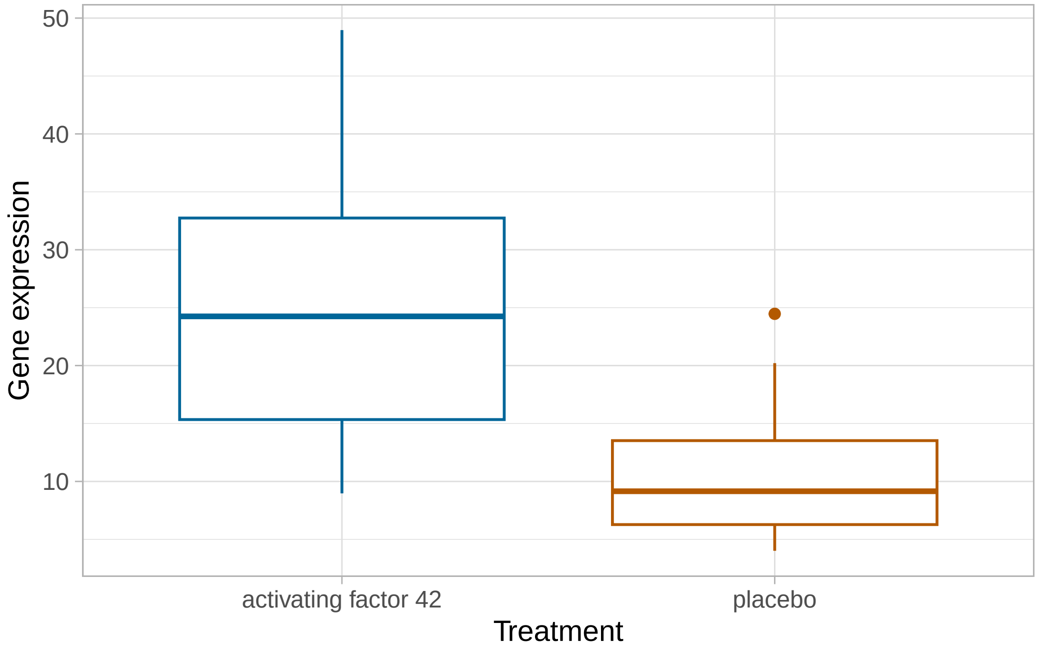

Figure 2: A boxplot of gene expression for each treatment (placebo and activating factor 42).

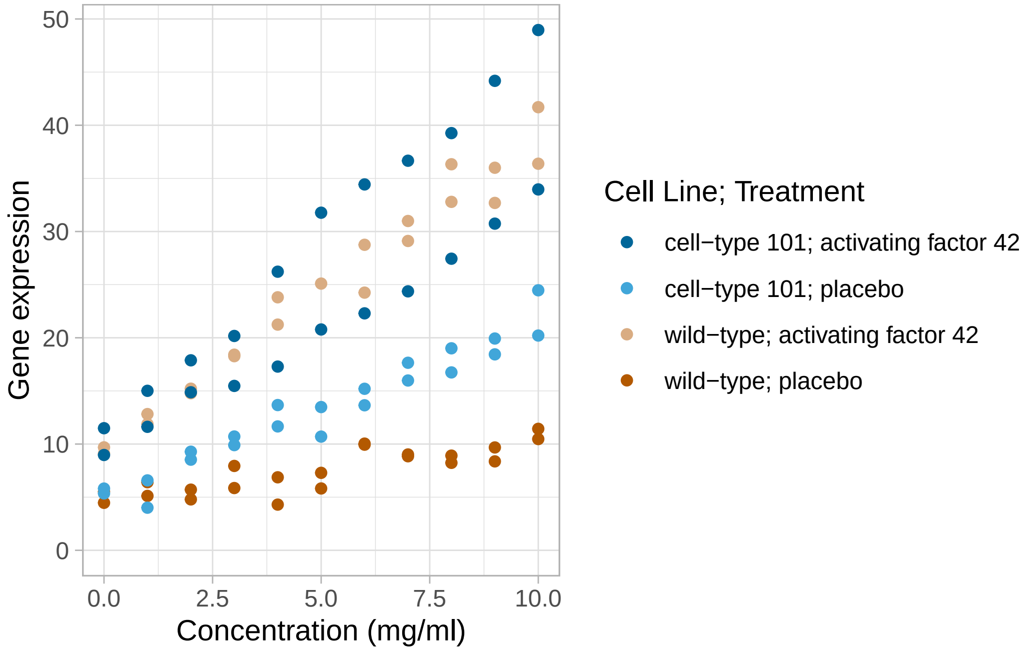

Figure 3: A plot of gene expression as a function of concentration, coloured by cell line (wild-type or cell-type 101) and treatment (placebo or activating factor 42).

line may be a predictor for gene expression.

Given that the gene expression for each cell line and treatment was measured for different concentrations of growth factor for the same gene line, then this must be taken into account in fitting models on the data.

Fixed effect model

A linear model can be fit using the `step` function to select the best model based on AIC, where the full scope is gene expression as a function of concentration, treatment, and cell line, with interaction terms between all three predictors. Using AIC, the function selects the full model as the best model.

The statistical significance of the terms in the selected model can be found using an ANOVA test. Table 1 shows the results of the ANOVA test conducted upon the selected fixed effect model, from which it can be seen that all of the terms are statistically significant.

Table 1: An ANOVA table showing the statistical significance of each predictor in the fixed effects model.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
concentration	1	3684.06853	3684.06853	427.186096	0.000000
treatment	1	4484.95153	4484.95153	520.052469	0.000000
CL	1	244.22439	244.22439	28.319034	0.000001
concentration:treatment	1	785.20064	785.20064	91.047925	0.000000
concentration:CL	1	80.56786	80.56786	9.342244	0.003054
treatment:CL	1	125.32971	125.32971	14.532605	0.000272
concentration:treatment:CL	1	51.47342	51.47342	5.968599	0.016794
Residuals	79	681.29889	8.62404		

Mixed effects models

The fixed effect model does not consider the impact of gene line. To consider this as a factor, a model can be fitted for gene expression as a function of concentration, with gene line as a random effect. The residuals plot for this model is shown in Figure 4. From this figure, it can be seen that there is still variance not explained by concentration alone, so this model will not be considered further.

The next model considered fits gene expression as a function of concentration and treatment (with interaction terms), as well as gene line as a random effect. The residuals plot for this model is shown in Figure 5.

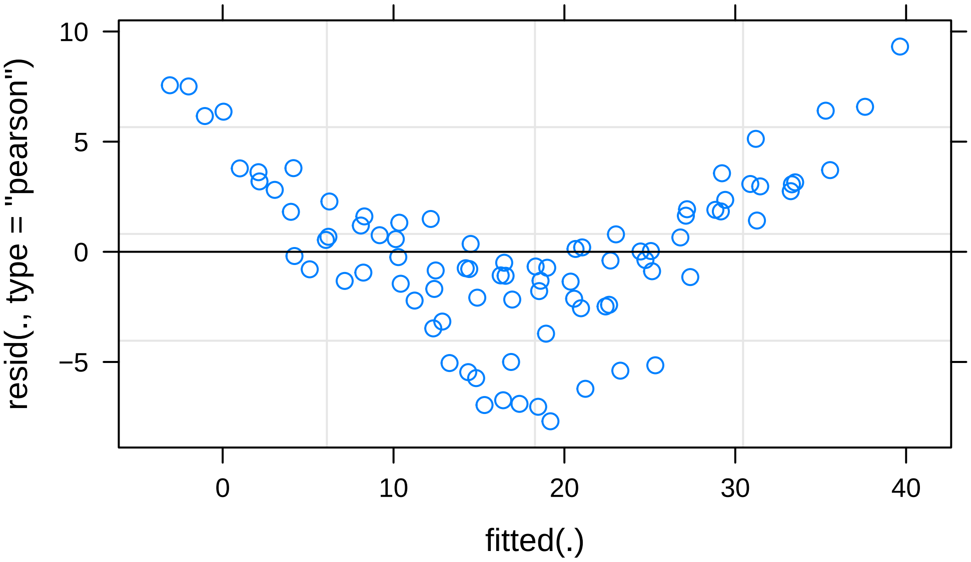

Figure 4: The residuals plot for the model of gene expression as a function of concentration, with gene line as a random effect.

Two models are fitted that include cell line as a predictor. One with interaction terms between concentration and treatment only, and one with interaction terms between concentration, treatment and cell line. The residuals plots for these models are shown in Figure 6 and Figure 7, respectively.

The residuals plots in Figure 5, Figure 6 and Figure 7 all show that there does not seem to be any residual variance for these models.

Table 2: An ANOVA table showing the statistical significance of each fixed effect predictor in the mixed effects model with concentration, treatment and cell line predictors, and an interaction term between concentration and treatment, as well as a gene line random effects.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
concentration	3684.06853	3684.06853	1	77.004313	865.879883	0.000000
treatment	16.76539	16.76539	1	5.909502	3.940431	0.095093
CL	8.57315	8.57315	1	5.004305	2.014977	0.214931
concentration:treatment	785.20064	785.20064	1	77.004313	184.548532	0.000000

Table 2 shows the statistical significance of each term in the mixed effects model with con-

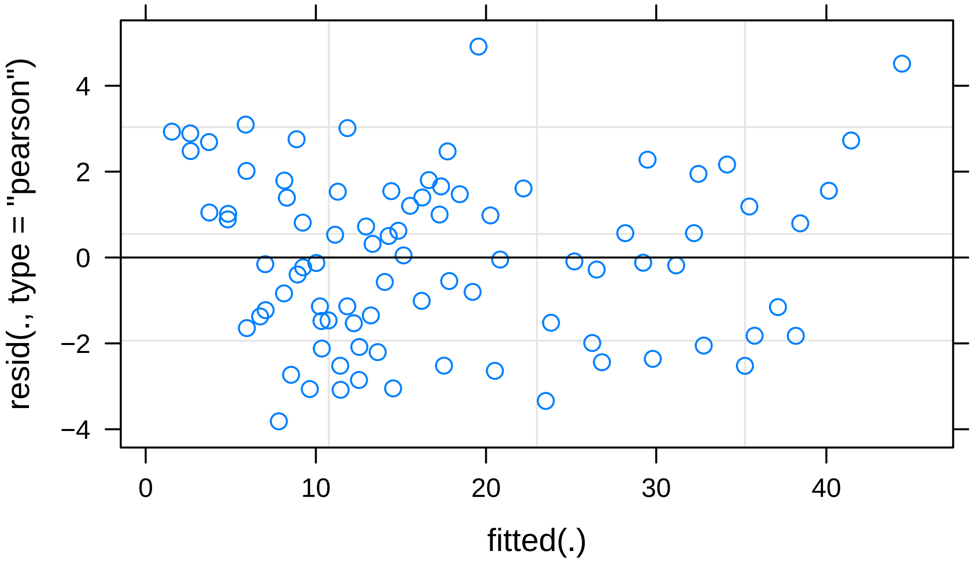

Figure 5: The residuals plot for the model of gene expression as a function of concentration and treatment (with interaction terms), with gene line as a random effect.

centration and treatment (with the interaction term) and cell line (without interactions with this predictor), as well as the gene line random effects. From this table, it can be seen that the cell line predictor is not statistically significant, so this term should be removed from the model. Removing this term results in the mixed effects models with concentration and treatment predictors (with the interaction term) and the gene line random effects. The statistical significance of each term in this model are shown in Table 3, from which it can be seen that all fixed effect terms in this model should be kept.

Table 3: An ANOVA table showing the statistical significance of each fixed effect predictor in the mixed effects model with concentration and treatment predictors, and an interaction term, as well as a gene line random effects.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
concentration	3684.06853	3684.06853	1	77.003443	865.874974	0.000000
treatment	14.51612	14.51612	1	6.926695	3.411757	0.107671
concentration:treatment	785.20064	785.20064	1	77.003443	184.547486	0.000000

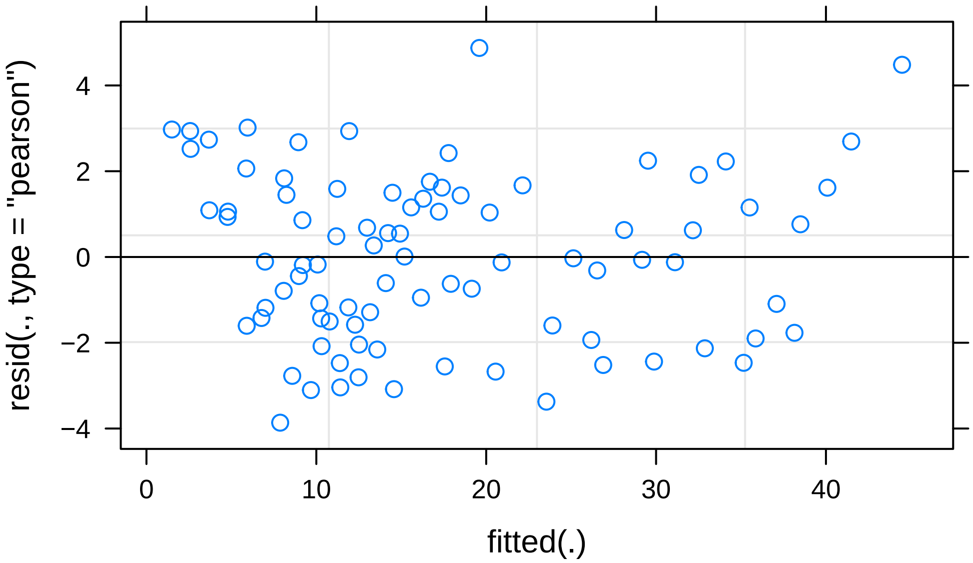

Figure 6: The residuals plot for the model of gene expression as a function of concentration and treatment (with interaction terms) and cell line (without any interaction terms), with gene line as a random effect.

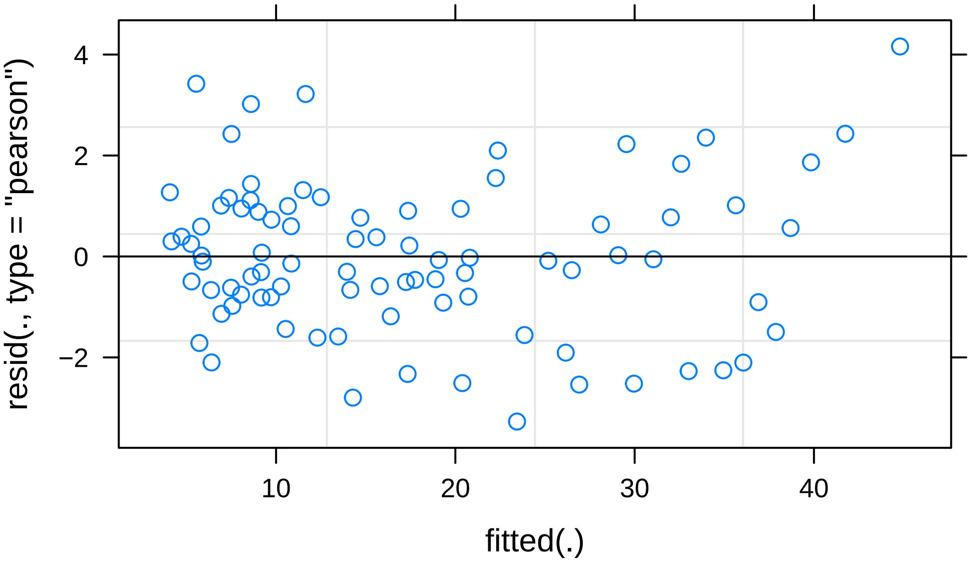

Figure 7: The residuals plot for the model of gene expression as a function of concentration, treatment and cell line (with interaction terms between all three predictors), with gene line as a random effect.

Table 4: An ANOVA table showing the statistical significance of each fixed effect predictor in the mixed effects model with concentration, treatment and cell line predictors, and interaction terms between all of the predictors, as well as a gene line random effects.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
concentration	3684.06853	3684.06853	1	75.001751	1412.759236	0.000000
treatment	10.65457	10.65457	1	4.441105	4.085794	0.106327
CL	0.04599	0.04599	1	4.441105	0.017637	0.900152
concentration:treatment	785.20064	785.20064	1	75.001751	301.107172	0.000000
concentration:CL	80.56786	80.56786	1	75.001751	30.896000	0.000000
treatment:CL	0.00075	0.00075	1	4.441105	0.000287	0.987221
concentration:treatment:CL	51.47342	51.47342	1	75.001751	19.738924	0.000030

Table 4 shows the statistical significance of each term in the model with concentration, treatment and cell line as predictors, along with interaction terms between all predictors, and gene line random effects. From this table, it can be seen that the interaction term between concentration, treatment and cell line is statistically significant, so this term should be kept. Because this term should be kept, then all of the other fixed effect terms should also be kept.

Table 5: An ANOVA-like table showing the statistical significance of the random effect term in the mixed effects model with concentration and treatment (and the interaction term between these predictors) as fixed effects, showing that the gene line random effect term is statistically significant.

	npar	logLik	AIC	LRT	Df	Pr(>Chisq)
	6	-198.2331	408.4662			
(1 | GL)	5	-237.8918	485.7837	79.31749	1	5.288934e-19

Table 6: An ANOVA-like table showing the statistical significance of the random effect term in the mixed effects model with concentration, treatment and cell line (and the interaction terms between these predictors) as fixed effects, showing that the gene line random effect term is statistically significant.

	npar	logLik	AIC	LRT	Df	Pr(>Chisq)
	10	-174.5644	369.1289			
(1 | GL)	9	-214.1471	446.2942	79.16527	1	5.712558e-19

The statistical significance of the random effect terms in the mixed effects models can also be tested, as shown in Table 5 and Table 6. From these tables, it can be seen that the random effect term in each model is statistically significant, and should therefore be retained.

Comparing models

Table 7: The AIC, R^2 values and root mean squared errors for each of the three fitted models.

Name	AIC	R2	conditional R2	RMSE
lm_step	443.9494	0.9327916		2.798396
m2	408.8027		0.9647610	1.942813
m4	372.5428		0.9788391	1.500209

The fixed effects model and mixed effects models can be compared to each other using AIC values, R^2 values and RMSE values (shown in Table 7). From these values, it can be seen that the model with interaction terms between concentration, treatment and cell line has the best AIC. The other two mixed effects models, where there is either no interaction with cell line or cell line is not a predictor, have very similar AIC values, whilst the model without random effects has the worst AIC. The conditional R^2 values, which take into account both the fixed effects and the random effects, are very similar for all models, but the model with all interaction terms is still slightly better. The R^2 value for the fixed effects model is worse than the conditional R^2 values for the mixed effects models. The root mean square errors of the first two mixed effects models in Table 7 are very similar, whilst the root mean squared error for the fixed effects model is much greater than the other values. The lowest root mean squared error occurs for the mixed effects model with interaction terms between all three predictors, suggesting that this model is the best. Thus, the model with interaction terms between all three predictors, and with gene line as a random effect, appears to be the best model.

Discussion

The chosen model is the mixed effects model with concentration, treatment and cell line as predictors, along with all interaction terms between the three predictors, and gene line as a random effect. This model is shown as the lines in Figure 8, where each line is the fitted model for a different gene line. This figure shows how the gene lines with the placebo treatment (in darker brown and lighter blue) have a flatter slope than the gene lines with the activating factor 42 treatment (in lighter brown, grey and darker blue). The slope of the fitted model for the wild-type cell lines with the placebo treatments (in darker brown) is also flatter than the slope of the fitted model for the cell-type 101 cell lines with the placebo treatments (in lighter blue).

The coefficients of the fitted model are shown in Table 8, and the random intercepts are shown in Table 9. The intercept for each gene line is found as the overall intercept (in Table 8) plus the gene line specific intercept in Table 9.

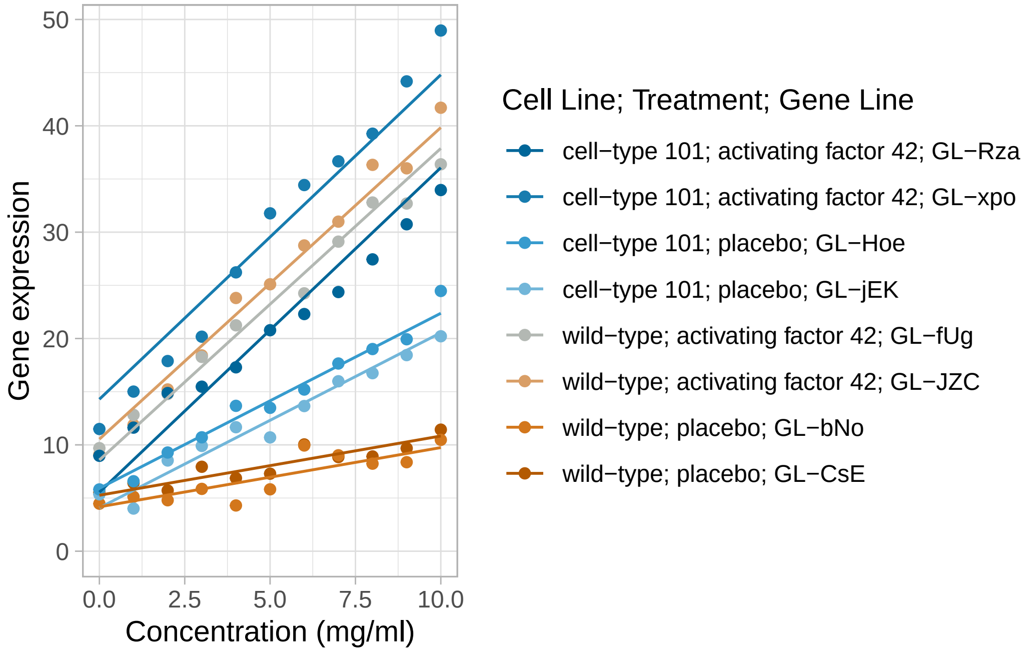

Figure 8: A plot of gene expression as a function of concentration, coloured by gene line (with cell line and treatment also indicated), and with the fitted model indicated by the lines.

Table 8: The coefficients of the chosen model. The value of the intercept is the overall intercept, which is added to the values in Table 9 to find the intercept for each gene line.

	value
(Intercept)	9.9175000
concentration	3.0514091
treatmentplacebo	-4.9215909
CLwild-type	-0.3615634
concentration:treatmentplacebo	-1.4055000
concentration:CLwild-type	-0.1214545
treatmentplacebo:CLwild-type	0.0817907
concentration:treatmentplacebo:CLwild-type	-0.9674091

Table 9: The difference from the overall intercept for each gene line.

	value
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	-4.3688926
GL-xpo	4.3688926

From these tables, it can be seen that as growth factor concentration increases, so does gene expression. It can also be seen that the placebo treatment has a smaller intercept and flatter slope than the activating factor 42 treatment does. Similarly, the wild-type cell line has a lower intercept and flatter slope than the cell-type 101 cell line does. Thus, gene expression is higher for higher concentrations of the growth factor, the activating factor 42 treatment and cell-type 101 cell line. Conversely, lower concentrations of the growth factor, the placebo treatment and wild-type cell line results in lower gene expression.

Appendix: Code

```r
pacman::p_load(tidyverse, readr, lme4, knitr, performance, sjPlot, lmerTest)
options(knitr.kable.NA = "")
theme_set(theme_light())
data <- read_csv("data/2023-03-01_gene-data.csv")
data_long <- data %>%
  mutate(CL = `cell line`,
         treat = treatment) %>%
  unite(`cell line`, `treat`, sep = "; ", col = "grouping") %>%
  pivot_longer(cols = 4:14, names_to = "concentration", values_to = "GE") %>%
  filter(GE >= 0) %>%
  mutate(concentration = as.integer(concentration),
         GL = as.factor(sheet_names),
         CL = as.factor(CL),
         treatment = as.factor(treatment),
         grouping = as.factor(grouping))
data_long %>%
  ggplot(aes(x = CL, y = GE, col = CL)) +
  geom_boxplot() +
  theme(legend.position = 'none') +
  harrypotter::scale_color_hp_d("Ravenclaw") +
  labs(x = "Cell line",
       y = "Gene expression")
data_long %>%
  ggplot(aes(x = treatment, y = GE, col = treatment)) +
  geom_boxplot() +
  theme(legend.position = 'none') +
  harrypotter::scale_color_hp_d("Ravenclaw") +
  labs(x = "Treatment",
       y = "Gene expression")
data_long %>%
  ggplot(aes(x = concentration, y = GE, color = grouping)) +
  geom_point() +
  ylim(0, NA) +
  harrypotter::scale_color_hp_d("Ravenclaw") +
  labs(x = "Concentration (mg/ml)",
       y = "Gene expression",
       color = "Cell Line; Treatment")
lm_null <- lm(GE ~ 1, data = data_long)
scope <- GE ~ concentration*treatment*CL
```

```
lm_step <- step(lm_null, scope = scope, direction = "both", trace = 0)
lm_step <- lm(GE ~ concentration*treatment*CL, data = data_long)
anova(lm_step) %>% kable(digits = c(0, 5, 5, 6, 6))
m1 <- lmer(GE ~ concentration + (1|GL), data = data_long, na.action = na.omit)
plot(m1)
m2 <- lmer(GE ~ concentration*treatment + (1|GL), data = data_long, na.action = na.omit)
plot(m2)
m3 <- lmer(GE ~ concentration*treatment + CL + (1|GL), data = data_long, na.action = na.om
plot(m3)
m4 <- lmer(GE ~ concentration*treatment*CL + (1|GL), data = data_long, na.action = na.omit
plot(m4)
anova(m3) %>% kable(digits = c(5, 5, 0, 6, 6, 6))
anova(m2) %>% kable(digits = c(5, 5, 0, 6, 6, 6))
anova(m4) %>% kable(digits = c(5, 5, 0, 6, 6, 6))
ranova(m2) %>% kable(digits = c(0, 4, 4, 5, 4, 25))
ranova(m4) %>% kable(digits = c(0, 4, 4, 5, 4, 25))
compare_performance(lm_step, m2, m4) %>%
  select(c("Name", "AIC", "R2", "R2_conditional", "RMSE")) %>%
  rename("conditional R2" = "R2_conditional") %>%
  kable(digits = c(0,4,7,7,6))
data_long %>%
  mutate(group = grouping,
         geneline = GL) %>%
  unite(group, geneline, sep = "; ", col = "grouping2") %>%
  ggplot(aes(x = concentration, y = GE, color = grouping2)) +
  geom_point() +
  geom_line(aes(y = predict(m4))) +
  ylim(0, NA) +
  harrypotter::scale_color_hp_d("Ravenclaw") +
  labs(x = "Concentration (mg/ml)",
       y = "Gene expression",
       color = "Cell Line; Treatment; Gene Line")
fixef(m4) %>%
  data.frame() %>%
  rename(value = ".") %>%
  kable()
random_effects <- ranef(m4)$GL
random_effects %>%
  rename(value = `(Intercept)`) %>%
  kable()
```

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.

Lüdecke, Daniel. 2023. *sjPlot: Data Visualization for Statistics in Social Science*. https://CRAN.R-project.org/package=sjPlot.

Lüdecke, Daniel, Mattan S. Ben-Shachar, Indrajeet Patil, Philip Waggoner, and Dominique Makowski. 2021. "performance: An R Package for Assessment, Comparison and Testing of Statistical Models." *Journal of Open Source Software* 6 (60): 3139. https://doi.org/10.21105/joss.03139.

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, 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.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.