Gene Expression - IMRaD

Douglas Dally - a1767927

5/1/23

Introduction

The dataset is comprised of gene expression values for 8 different cell types. For each cell type, a treatment type was chosen, and a varying concentration of the treatment was applied to 11 samples. The data was provided by the Institute of -Omics in Adelaide, and consisted of 88 observations with variables:

- cell_line: A factor with 2 levels, wild-type or type-101.
- cell_type: Factor with 8 levels denoting the type of cell (GL-CsE, GL-bNo, GL-JZC, GL-fUg, GL-jEK, GL-Hoe, GL-Rza, GL-xpo).
- gene_exp: Gene expression, a continuous numeric value measured for each cell.
- treatment: Treatment type, a factor with 2 levels; placebo (saline) or Activating Factor 42.
- concentration: An integer value from 0 to 10 representing the Concentration of treatment applied to a cell in μg/L.

The key research question was to use this data to produce a predictive model for gene expression.

Methods

This analysis was performed using the packages `tidyverse` [@tidyverse], `lmerTest` [@lmerTest-2017], `gglm` [@gglm-2022] and `MumIn` [@MuMIn-2023] in R [@R] and RStudio. The data was loaded in as .csv file and cleaned.

This involved removing any missing values, which were represented by a -99 entry for gene_exp. This observation was removed with approval from a representative of the institute of omics. This resulted in 87 total observations.

A linear mixed-effects model was chosen to be used to predict gene expression, with cell type as a random effect (intercept), and concentration, treatment and cell line as predictors, along with all corresponding interaction terms between these three predictors.

Starting with a full model for gene expression containing all predictors, a backwards selection algorithm was used to reduce the model size by removing insignificant predictors (p-value < 0.05) and refitting the model. The best model was then chosen using both AIC and R-squared as measures.

Results

The code output below gives the fixed effect coefficients of a linear fixed effects model. The last column of this table are the p-values of each predicitor in the model.

```
                                            Estimate Std. Error        df
(Intercept)                                9.91750000  2.4142839  4.438860
concentration                              3.05140909  0.1088726 75.001751
treatmentplacebo                          -4.92159091  3.4143130  4.438860
cell_linewild                             -0.36156344  3.4151798  4.443351
concentration:treatmentplacebo            -1.40550000  0.1539690 75.001751
concentration:cell_linewild               -0.12145455  0.1539690 75.001751
treatmentplacebo:cell_linewild             0.08179071  4.8291807  4.441105
concentration:treatmentplacebo:cell_linewild -0.96740909  0.2177451 75.001751
                                              t value     Pr(>|t|)
(Intercept)                                4.10784340 1.190974e-02
concentration                             28.02734787 1.748521e-41
treatmentplacebo                          -1.44145863 2.160929e-01
cell_linewild                             -0.10586952 9.202962e-01
concentration:treatmentplacebo            -9.12845797 8.484452e-14
concentration:cell_linewild               -0.78882441 4.327011e-01
treatmentplacebo:cell_linewild             0.01693677 9.872211e-01
concentration:treatmentplacebo:cell_linewild -4.44285088 3.014502e-05
```

The 3-way interaction term is significant with a p-value of 3.01×10^{-5}, however the next two predictors are 2-way interaction terms involving cell line, which have insignificant p-values (both above 0.05). These predictors need to be removed from the model, however, by the principle of marginality the 3-way interaction term must first be removed. The coefficients of the model are reassessed and shown below.

```
                           Estimate Std. Error        df     t value
(Intercept)               8.7082386  2.4103349  4.409502   3.6128749
```

```
                                  Estimate Std. Error       df   t value
concentration                    3.2932614  0.1052702 76.002182  31.2838952
treatmentplacebo                -2.5030682  3.3815281  4.270748  -0.7402181
cell_linewild                    2.0572234  3.3826188  4.276243   0.6081748
concentration:treatmentplacebo -1.8892045  0.1215555 76.002182 -15.5419045
concentration:cell_linewild    -0.6051591  0.1215555 76.002182  -4.9784576
treatmentplacebo:cell_linewild -4.7555189  4.7051095  4.002177  -1.0107138
                                   Pr(>|t|)
(Intercept)                    1.904158e-02
concentration                  3.693155e-45
treatmentplacebo               4.978038e-01
cell_linewild                  5.739021e-01
concentration:treatmentplacebo 2.627029e-25
concentration:cell_linewild    3.898251e-06
treatmentplacebo:cell_linewild 3.692970e-01
```

We see that the interaction term corresponding to a wild type cell line under a placebo treatment is insignificant (p-value = 0.369), and so we need to remove this predictor. Refitting the model:

```
                                  Estimate Std. Error        df     t value
(Intercept)                     9.8967224  2.1084007  5.686288   4.6939476
concentration                   3.2932614  0.1052698 76.003298  31.2840101
treatmentplacebo               -4.8800357  2.4348239  5.688632  -2.0042664
cell_linewild                  -0.3213279  2.4348239  5.688632  -0.1319717
concentration:treatmentplacebo -1.8892045  0.1215551 76.003298 -15.5419616
concentration:cell_linewild    -0.6051591  0.1215551 76.003298  -4.9784759
                                   Pr(>|t|)
(Intercept)                    3.851931e-03
concentration                  3.688667e-45
treatmentplacebo               9.448535e-02
cell_linewild                  8.995498e-01
concentration:treatmentplacebo 2.625474e-25
concentration:cell_linewild    3.897888e-06
```

The output now shows a model with two 2-way interaction terms which are significant, but now the predictor for a wild-type cell line is insignificant. To remove this single predictor, the interaction terms are first removed by the principle of marginality.

```
                 Estimate Std. Error        df   t value     Pr(>|t|)
(Intercept)      16.132424  2.1400329  6.019797  7.538400 2.781392e-04
concentration     2.046080  0.1273222 78.014331 16.070087 1.819603e-26
```

```
treatmentplacebo -14.325644  2.3603645  5.014244 -6.069251 1.736206e-03
cell_linewild     -3.347538  2.3603645  5.014244 -1.418229 2.151699e-01
```

Following this step, the predictor for cell line is still insignificant with a p-value of 0.215, and is removed.

```
                  Estimate Std. Error        df   t value     Pr(>|t|)
(Intercept)       14.46288  1.9140993  7.608812  7.555969 8.546501e-05
concentration      2.04608  0.1273234 78.011473 16.069938 1.822445e-26
treatmentplacebo -14.32986  2.5512632  6.011364 -5.616772 1.351109e-03
```

Finally, the model has been reduced to a point where the p-values for the remaining predictors are significant, as seen above. However, this model now only has 2 predictors.

These 5 models were then compared by computing the AIC and R-squared, summarised in Table 1.

Table 1: Table 1 - AIC and R-squared value (both margionalised and conditional) for each model generated from the backwards selection process.

model	n_pred	AIC	R2m	R2c
step5	2	503.1784	0.7848944	0.8818413
step4	3	502.4759	0.8024080	0.8830156
step3	5	387.5630	0.8846660	0.9733528
step2	6	387.7436	0.8862156	0.9736190
full	7	371.2856	0.8909742	0.9788391

Figure 1 shows the diagnostic plots for the full model, so we can check the assumptions of a linear model.

Table 2 shows the results of the predicted and actual value for gene expression given a wild-type cell of type GL-bNo when given a placebo treatment at a concentration of 8 μg/L.

Table 2: Table 2 - Prediction and Actual gene expression value for a wild-type GL-bNo cell treated with placebo at a concentration of 8 micrograms per litre.

cell_line	cell_type	concentration	treatment	prediction	actual
wild	GL-bNo	8	placebo	8.627612	8.23

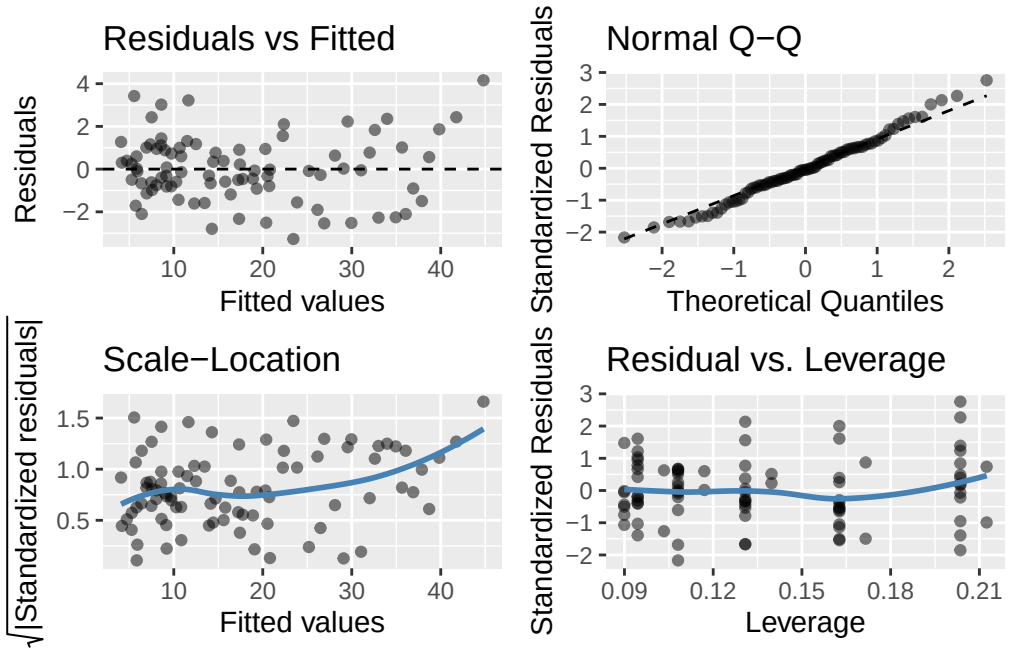

Figure 1: Figure 1 - diagnostics plots for full model.

Discussion

From Table 1, we see that the best model is the full linear mixed effects model, as it has the lowest AIC of 371 and the greatest R-squared values with 0.89 marginalised and 0.97 conditional. This means, going by the marginal R-squared, that the model captures 89% of the variance in the data.

The assumptions of linear modelling can be checked using Figure 1. From the residuals vs. fitted plot, we see there may be some evidence of non-constant variance as seen by a slight curve, but overall is not too bad. The normal QQ plot shows a roughly linear relationship, while the scale-location seems evenly spread. There are no major outliers in the residual vs. leverage plot. Overall, the assumptions for a linear model seem correct.

From Table 2, we see that the model produces a prediction that is close to the actual recorded gene expression for a cell with those conditions.

In future, splitting the data into testing and training sets before fitting the model, and then using the model to predict the testing set would have been a better approach to testing the models accuracy. This would have also helped prevent over fitting.

Appendix

Code for Analysis

```
# Load Libs
pacman::p_load(tidyverse, lmerTest, MuMIn, gt, gglm)

# load data
data <- read.csv(here::here("data", "2023-05-29_cleaned-data-final.csv"))

# clean data
data <- data |>
  mutate(cell_type = as.factor(exp_name),
         cell_line = as.factor(cell_line),
         treatment = as.factor(group))

# obtain necesarry columns
data <- data |>
  select(cell_line, gene_exp, concentration, treatment, cell_type)
data

# Full fixed effect model
M_full <- lmer(gene_exp ~ concentration*treatment*cell_line + (1|cell_type), data = data)
summary(M_full)$coefficients
R2_full <- r.squaredGLMM(M_full)

# remove 3 way interaction term
M2 <- update(M_full, .~. - concentration:treatment:cell_line)
summary(M2)$coefficients
R2_2 <- r.squaredGLMM(M2)

# remove 2 way interaction term with insignificant P-value
M3 <- update(M2, .~. - treatment:cell_line)
summary(M3)$coefficients
R2_3 <- r.squaredGLMM(M3)

# still have insignificant terms as single predictors. hence to remove them we must remove
M4 <- update(M3, .~. - concentration:treatment - concentration:cell_line)
summary(M4)$coefficients
R2_4 <- r.squaredGLMM(M4)
```

```r
# cell_line wild is insignificant
M5 <- update(M4, .~. - cell_line)
summary(M5)$coefficients
R2_5 <- r.squaredGLMM(M5)

# get summary table for 5 models
AIC <- anova(M5, M4, M3, M2, M_full)[2]
R_squared <- rbind(R2_5, R2_4, R2_3, R2_2, R2_full)
n_pred <- c(2,3,5,6,7)

sum_tab <- cbind(n_pred, AIC, R_squared)
sum_tab

# Predicting data
new_data <- data.frame(
  cell_line = "wild",
  cell_type = "GL-bNo",
  concentration = 8,
  treatment = "placebo")
new_data

prediction <- predict(M_full, new_data)
actual <- 8.23
tab2 <- cbind(new_data, prediction, actual)
tab2 |> gt()
```