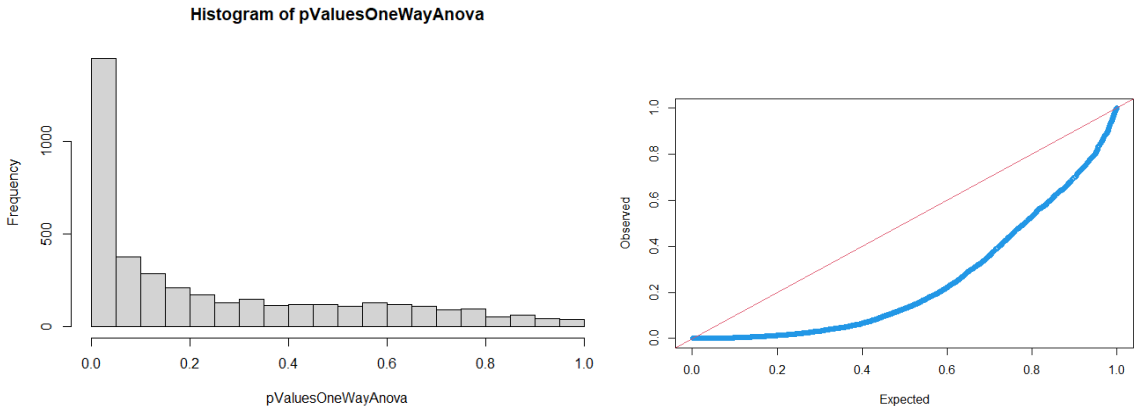
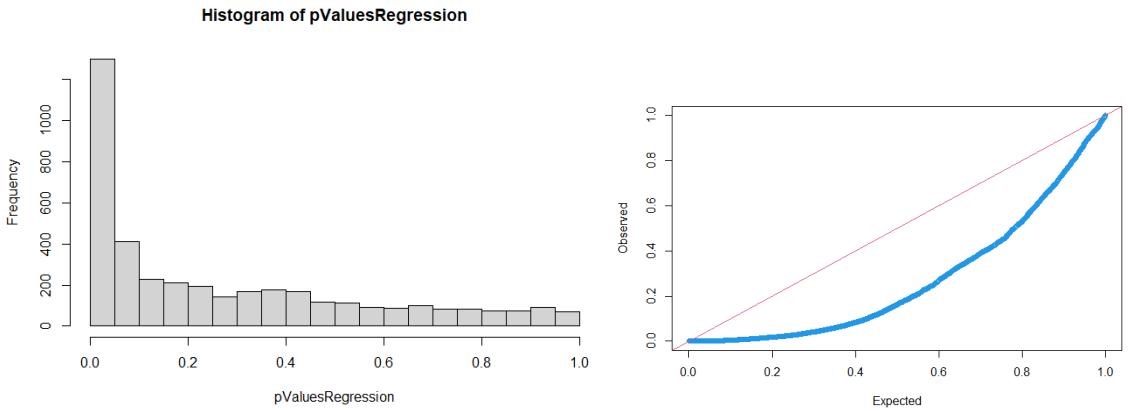
1) **Problem (1A):** For each row in the spreadsheet, perform a one-way ANOVA with categories “day 2”, “week 12” and “week 18”. Plot out the histogram of all p-values. How many genes are significant at a BH FDR-corrected 0.05 threshold.



pValuesOneWayAnova_bh <- p.adjust (pValuesOneWayAnova, method='BH')
print(sum(pValuesOneWayAnova_bh < 0.05)) ## 612

612 genes are significant out of **3983.**

2) **Problem (1B):** Next make an ANOVA as a linear regression as a function of time (so 2 days, 86 days and 128 days). Plot out the histogram of all p-values. How many genes are significant at a BH FDR-corrected 0.05 threshold.

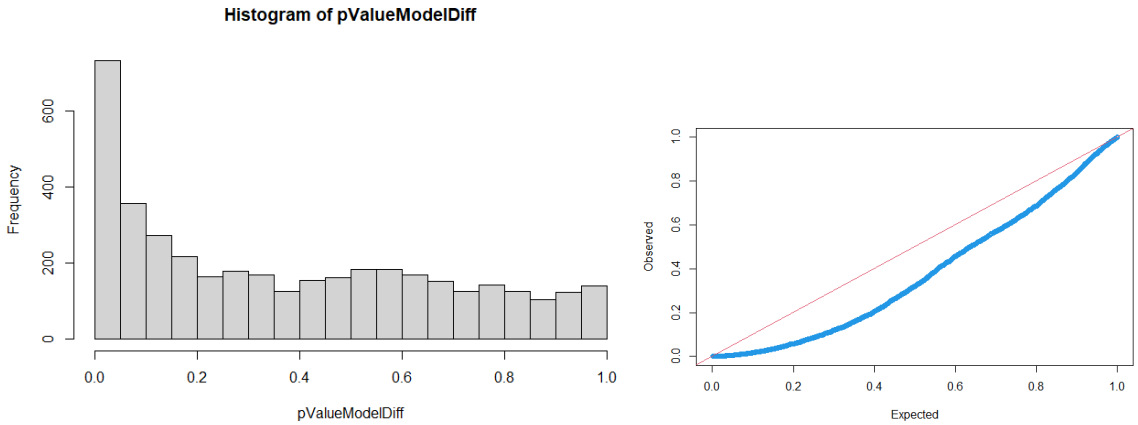


pValuesRegression_bh <- p.adjust (pValuesRegression, method='BH')
print(sum(pValuesRegression_bh < 0.05)) ## 448

448 genes are significant out of **3983.**

3) **Problem (1C):** Finally, for each row in the spreadsheet perform an ANVOA comparing the three-parameter model from (A) and the two-parameter model from (B). Plot out the histogram of all

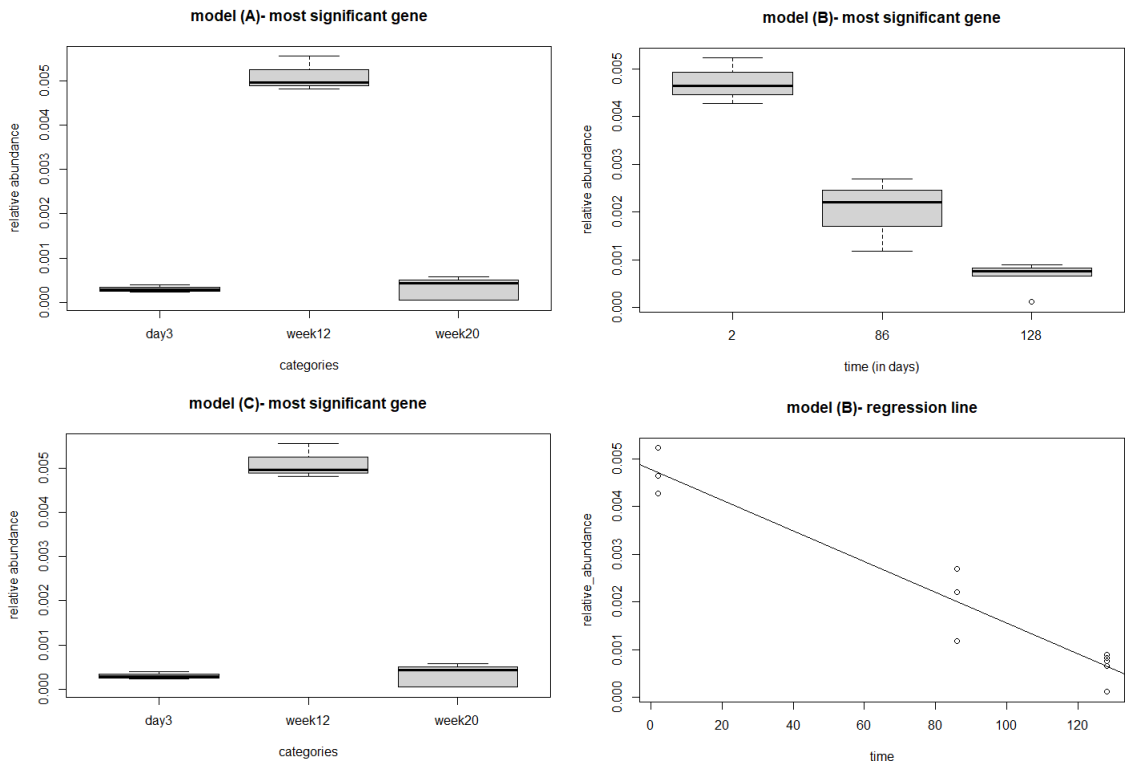
p-values. For how many genes is there a significant difference between these two models at a BH FDR-corrected threshold.



pValueModelDiff_bh <- p.adjust (pValueModelDiff, method='BH')
print(sum(pValueModelDiff_bh < 0.05)) ## 51

51 genes are significant out of **3983.**

4) **Problem (1D):** Make three graphs showing the relative abundance of the most significant gene under each of the three ANOVA models. For (A) and (C), the x-axis will the category (day 3, week 12 and week 18) and the y-axis will be the relative abundance. Be sure to properly label and title all graphs and axes. For (B) the x-axis will be time (in days) and the y-axis will be the relative abundance. For the graph of the top hit from (B), include the regression line for the plot from (B).



For, model A (one way anova) and C (model diff) I have found the same gene to be the most significant one, it's with index = 2896. However, for model B (regression) the index of the most significant gene was 2915.

5) **Problem (1E):** Overall, do you think the three parameter model in (A) or the two-parameter model in (B) is more appropriate for these data? Justify your answer.
Answer: Overall, I think the three param model (A) is more appropriate for these data, cause from the p values by ANOVA we have seen more number of significant values/genes for model (A) than the model (B), which I assume shows that with the three param model, for most of the genes, we are learning additional information.