At the very beginning of the experiment...

characteristics of proteins and transcription factors at time t=0.5h

Group 2; David, Eamon, Joshua & You Yu

Dec 8, 2022

Introduction

The vast data set consists of more than 500 thousand observations from 34 variables. As a whole, it shows the levels of 22 transcription factors and four relevant proteins and how they might change concerning time and different amounts of doses composed by different types of drugs for three repetitions. We are specifically interested in the very start of the experiments. Thus, we picked an exact time spot t=0.5h and tried to draw some interesting conclusions about the relationship of the variables at this time.

Hypothesis Testing: Will 1uM of “Vem” result in a difference in the mean protein level?

Hypothesis Testing - Data Summary

```
finalproject <- read_csv("STA130_Course_Project.csv")
dataset <- finalproject %>% filter(Timepoint == "0.5 h", dose_id == "1" |
                                   dose_id == "2", Drugs == "Vem")
dataset <- dataset %>% select(NGFR, AXL, Sox10, MiTFg, Doses)
dataset <- dataset %>% group_by(Doses)
```

Firstly, we filter out the relevant data at t=0.5h and dose_id = "1"(0uM) or "2"(1uM) and drug type = "Vem." Then we select the proteins and doses we are interested in and delete the rest variables as they are irrelevant. Lastly, we group the data set by Doses(1uM or 0uM) for hypothesis testing.

Hypothesis Testing - Statistical Methods

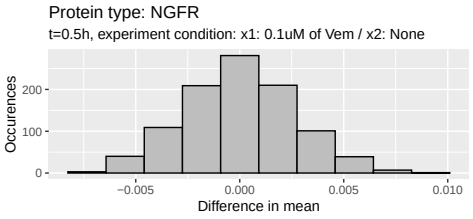
$H_0 : Mean_{0uM} = Mean_{1uM}$; $H_1 : Mean_{0uM} \neq Mean_{1uM}$

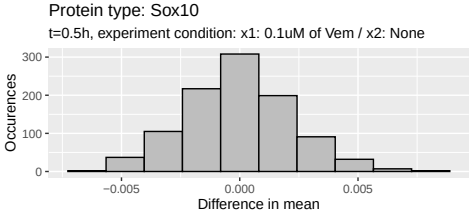
Null Hypothesis: There IS NO difference in the mean protein level, regardless of 1uM of the "Vem" drug.

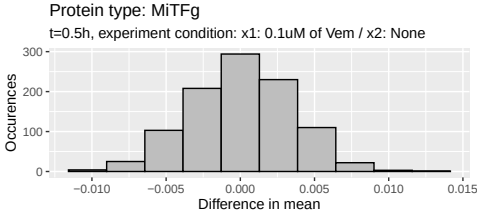
Alternative Hypothesis: There IS a difference in the mean protein level with 1uM of the "Vem" drug.

Then we set the number of repetitions to 1000. In each repetition, we randomly assign the doses into two groups using the sample() function. After that, we will calculate the difference in the groups' mean. By comparing them to the original difference in mean, we could obtain the number of occurrences of more extreme cases. Hence, we could also get the p-value by calculating the probability of more extreme cases.

Hypothesis Testing - Results







Hypothesis Testing - Results

Protein Name	Test Value	p-value
NGFR	-0.003123238	0.242
AXL	0.02071139	0
Sox10	0.02787675	0
MiTFg	0.0210963	0

Hypothesis Testing - Conclusion

From the data above, the p-values are 0.242, 0, 0, and 0, respectively. Therefore, there is extremely strong evidence to reject the NULL hypothesis for the latter three proteins AXL, Sox10, and MiTFg. Furthermore, for the first protein NGFR, the p-value is 0.242, which has no evidence against the NULL hypothesis, so we cannot reject it. Therefore, at t=0.5h, the Vem drug's dose of 0.1uM would result in a difference in the mean protein level for the latter three proteins AXL, Sox10, and MiTFg, but not for the first protein NGFR.

Limitations:

1. We only evaluate the difference in mean values to address whether there is a change in the protein level, which could be biased. For example, the distribution could become more skew and have more extreme values with the same mean.
2. The difference in the type of doses may not be the causation of the mean difference.

Correlation Estimation: What is the relationship between ATF3 and MiTFg at time t=0.5h under the experimental condition 0.1uM of “Vem”?

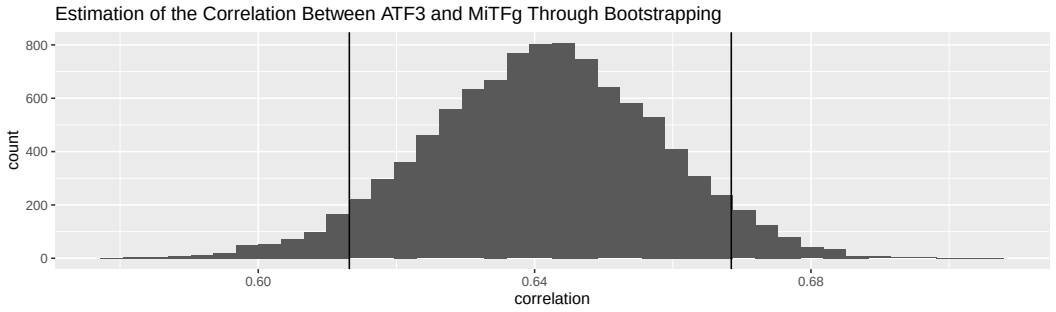
Correlation Estimation - Data Summary

1. CSV file is made into a data frame using read_csv
2. Filtering the observations that do not satisfy time t=0.5 h and experimental conditions Drugs == Vem and Doses == 0.1 uM
3. ATF3 and MiTFg variables selected and stored under 'main_file'

Correlation Estimation - Statistical Methods

1. Correlation between ATF3 and MiTFg estimated through bootstrapping
2. Pseudo-random number generator set up to make the process repeatable
3. Observations of 'main_file' is sampled with replacement
4. correlation between AT3 and MiTFg is calculated and stored in the vector 'bootstrap_cor'; the process is repeated 10 000 times
5. 90% confidence interval derived from 'bootstrap_cor'

Correlation Estimation - Results




```
##         5%        95%
## 0.6131836 0.6684658
```

Correlation Estimation - Conclusion

With 90% confidence, the correlation situates between 0.6131836 and 0.6684658.

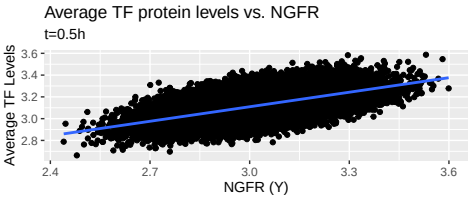
Simple Linear Regression: What is the relationship relationship between the cellular phenotypical outcomes and the transcription factors at time t=0.5h?

Simple Linear Regression - Data Summary & Statistical Methods

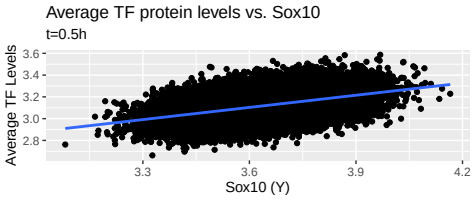
We mutate a new variable called avg.TF, which is the mean of the 22 TF factors. Then we apply simple linear regression to avg.TF and the proteins using the "lm" in ggplot to find a relationship between different cellular phenotypical outcomes.

We can also use the summarise() function to describe the y-intercept and slope of the regression line to understand their relation better.

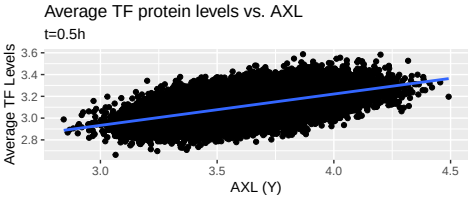
Simple Linear Regression - Results



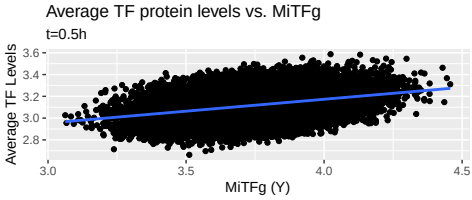
y−intercept: 0.667; gradient: 0.755



y−intercept: 2.209; gradient: 0.459



y−intercept: 0.512; gradient: 1.003



y−intercept: 2.109; gradient: 0.555

Simple Linear Regression - Conclusion

We could discover the protein levels of the cellular phenotypical outcomes using regression equations. According to regression equations, there is a positive relationship between the protein types and the transcription factors for all of the proteins at time t=0.5h, as they all have a positive gradient for the regression line. Sox10 has the slightest gradient of 0.459, while AXL has the largest, 1.003.

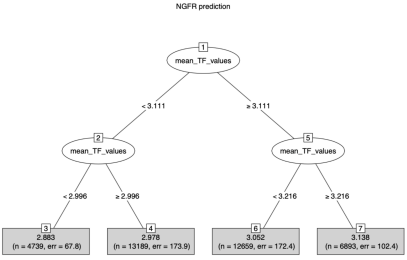
Limitations: 1. Using the mean values for all of the transcription factors may over generalize the results. 2. There may be other contributing factors that may affect the causation of the positive relationship.

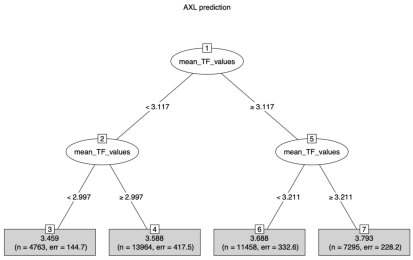
Second Approach, Classification: What is the relationship relationship between the cellular phenotypical outcomes and the transcription factors at time t=0.5h?

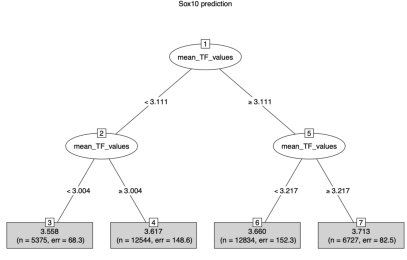
Classification - Statistical Methods

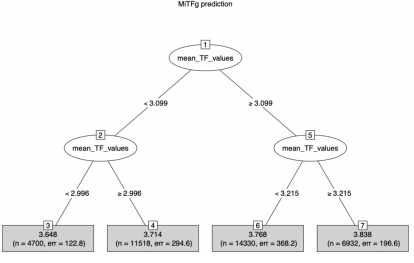
To avoid complexity and over-fitting, we will use the same avg.TF value to classify the proteins' level. We'll divide our data set to two separate data sets "train" and "test" using the 80-20 approach, and we will use the "train" data set to fit the model.

Classification - Results









Classification - Conclusion

From the classification trees, we could understand what approximate values the proteins will have concerning the TF factors. To analyze the Gene types, we could further identify a cutoff and classify the proteins into “High” and “Low” by this cutoff.

The limitations are the same as the simple linear regression part; multivariate classifications should be used if we want to study one specific protein’s relation with other proteins.

Final Conlusion: What do the results mean?

1. ATF3 has a positive correlation to MiTFg
2. The higher average transcription factors cause a rise in AXL, MiTFg, NGFR, and Sox10 levels
3. Since ATF3 having a positive correlation to MiTFg lines up with the regression testing results, we can verify the connection between the results.
4. Therefore, we can establish that an increase in protein levels implies an increase in average transcription factors

Final Conclusion: What do the results mean?

5. 'Vem' causes an increase in the mean difference of the protein levels excluding NGFR
6. This allows us to point out the contradiction in the 'Vem' tests results, in which all the mean protein levels increased, excluding NGFR
7. Either NGFR's correlation to average transcription factors changed, or a new factor in determining the mean protein levels of NGFR is at play
8. This newly apparent relationship can be used as a guideline for turning deleterious cells into non-deleterious cells.