Regression Results

To view the full ANOVA results for each regression:
https://github.com/Rahul-Khanna/comp_gene/blob/master/Regression.ipynb
^ to quickly skip from one result to the next -> Ctrl + F and search for &&&&

To view condensed ANOVA results for each regression:
https://docs.google.com/spreadsheets/d/1-UjCZGFgkHl2hwE-BJYeOJRLfLrtlBUoqX_zZ5Y3tPE/edit?usp=sharing
^ there are two tabs, one tab is for all regressions run with Type (Case vs Control) as an independent variable, the other is for all regressions run with PD Duration as an independent variable

Note: In each regression the dependent variable is protein levels… if not stated otherwise it is protein levels irrespective of the time of measurement

Note : Every patient in the Control did not have a value for PD Duration. In an attempt to see how these variables affected the protein levels I had to run separate regressions, one where Type was in the independent variables, and one where PD Duration was in the independent variables for each split of the data, hence why some of the regressions will look like duplicates, but aren't as the Independent Variables change.

Significance Numbers

Format of below :
Variable - %of_time_variable_was_significant_using_alpha_of_5%,
%of_time_variable_was_significant_using_alpha_of_10%

Regressions with Type as Independent Variable

- Type - 4/14, 5/14
- Gender - 11/12, 12/12
- Age - 4/9, 7/9
- MonthsFromEval (BL, 4, 6 etc) - 2/8, 5/8

- F- stat - 13/14, 14/14

Regressions with PD Duration as Independent Variable

- PD Duration - 0/14, 0/14
- Gender - 11/12, 12/12
- Age - 3/9, 3/9
- MonthsFromEval (BL, 4, 6 etc) - 1/8, 2/8

- F - stat - 11/14, 12/14

At least from a linear relationship perspective there is definitely a relationship between Gender and Protein levels (which was evident from bar graphs in the first ipython notebook). There also seems to be a relationship between Age and Protein levels, as well as MonthsFromEval (the time the protein level was taken at) and Protein levels. Finally for certain cuts of the original dataset the type of patient (case vs control) has a linear relationship with Protein levels.

Even though every measure of how good of a predictive tool these regressions would be is really low, the f-stat is consistently significant, which indicates using linear combinations of these variables have a relationship with Protein Levels

Performance Numbers

So in general the regression performed pretty poorly, with the max adjusted R^2 square being 0.139, the maximum log-likelihood being -300, and the best AIC and BIC values being in the 500s.

In general the regression looked better when regressing using the Type variable instead of the PD Duration variable.

Surprisingly (given the performance of the classifiers), the regressions performed better when only looking at male data.

The regressions performed better on people between the ages of 60 and 65, and for the regressions using PD Duration, the regression also performed better for people between the age of 65 and 70.

Finally when splitting the dataset by monthsFromEval, i.e. only looking at protein levels taken as baseline, four months after baseline, 6 months after baseline, etc, the regression performed better when only looking at proteins levels after the 8 month mark, and after the 10 month mark.

Conclusion

The goal of this exercise was not to predict protein levels, but understand what variables affect the protein levels of patients.

We know for certain that Gender plays a part in the varying levels of protein.

There is a strong inclination that Age also plays some role in protein levels.

What's interesting is the time since first evaluation doesn't play too big of a role in protein levels, which suggests the protein levels stay relatively stable over time.

Whether a patient is Case or Control doesn't seem to have too big of an impact either, or at least not in a pure linear fashion (piece-wise linear is still possible)

It looks likely that PD Duration has no impact on protein levels, as it very rarely had a significant coefficient

We know that using combinations of these variables are not effective for predicting protein levels, but on the flip side linear combinations of these variables do have a relationship with protein levels as evident by the f-stat.

Moving Forward

1. Run a few more regressions : Single variable regressions as well as using just Gender, Age, MonthsFromScreening and seeing the effect of that on protein levels

2. Look at bar plots over time of protein levels for individuals belonging to different age brackets

3. Add the age of the person to the data points for the classifiers

4. Fix data imbalance issue between Case vs Control