WORKFORCE ANALYTICS - INDIVIDUAL ASSIGNMENT : PART 1
MARIA ATHENA B. ENGESAETH (CID 01159179)

The dataset used to answer the two below questions was the list of US patents awarded to firms in the semiconductor industry between 2000-2005. The semiconductor industry is an industry that relies heavily on high performing employees that are experts and can secure patents for their employers. Although some of the results can be extrapolated to other industries, caution should be made in doing so with the below models due to limitations pertained to severe endogeneity issues.

Q 1A: Relationship between the number of times inventors collaborate and project performance

For the analysis I used the average cosine similarity of each team by comparing the inventor ID numbers (inv_num) in teams, as a metric to measure past collaborations. I assumed that collaboration between inventors on patents only happened within the same firm. The performance of the collaborations was measured by using the number of citations as a proxy.

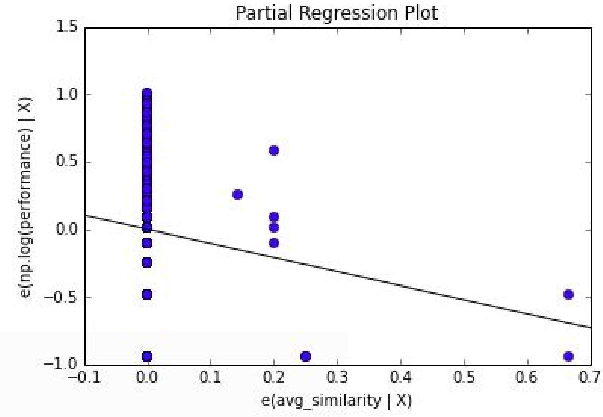
For the analysis of the relationship between the two variable I used OLS regression with performance in the natural logarithmic form to void negative values and get a look at the incremental effect of past collaboration.

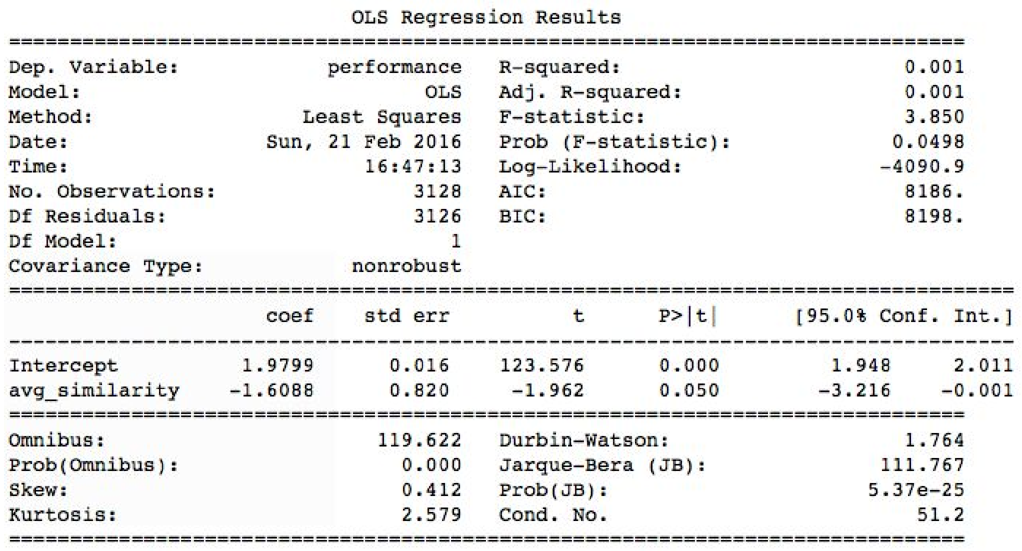
$$\log(\text{performance}) = 1.979 - \underset{[0.820]}{1.608}\ \text{avg\_similarity}$$

R^2 : 0.001
Sample size: 3,128

The results do not seem intuitive as past collaboration appear to have a negative effect on the performance of the team. A narrative around this could be how a high level of specialisation can be detrimental to performance as new ideas are not being explored. Including a squared term of avg_similarity might offset this, however upon inclusion

The p-value associated with the avg_similarity of 0.050 indicates the avg_similarity is of statistical significance. However, given the high standard error and low R^2 indicate there is more to be desired in terms of appropriately explaining the variability of team performance. To improve on the model it would be interesting to add more variables to the regression model to avoid endogeneity issues.





```
                            OLS Regression Results
==============================================================================
Dep. Variable:            performance   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     3.850
Date:                Sun, 21 Feb 2016   Prob (F-statistic):             0.0498
Time:                        16:47:13   Log-Likelihood:                -4090.9
No. Observations:                3128   AIC:                             8186.
Df Residuals:                    3126   BIC:                             8198.
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept          1.9799      0.016    123.576      0.000         1.948     2.011
avg_similarity    -1.6088      0.820     -1.962      0.050        -3.216    -0.001
==============================================================================
Omnibus:                      119.622   Durbin-Watson:                   1.764
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              111.767
Skew:                           0.412   Prob(JB):                     5.37e-25
Kurtosis:                       2.579   Cond. No.                         51.2
==============================================================================
```

Endogeneity issues that are likely to exist in this model would be related to having excluded variables such as the teams' (R&D) budget (or firm size as a proxy for this), salary, team size, experience (number of patents an inventor has appeared on) and education of inventors. Some of this data can be found in the data set, although not all of it.

Although cosine similarity is efficient at measuring the composition of teams in this data set, it is not able to capture team members that were not inventors, yet were involved in the work of the team and by consequence would have a great impact on the team performance. Furthermore performance in the form of citations may also be a metric that increases over time as a publication has more time to circulate and this temporal aspect has not been controlled for here.

Similar analysis could be done on the output in industries that rely on patents or where publication and the centrality of these, such as for example journalism.

Q 1B: Relationship between the ethnic composition of a firm and the likelihood that an inventor will move to that firm

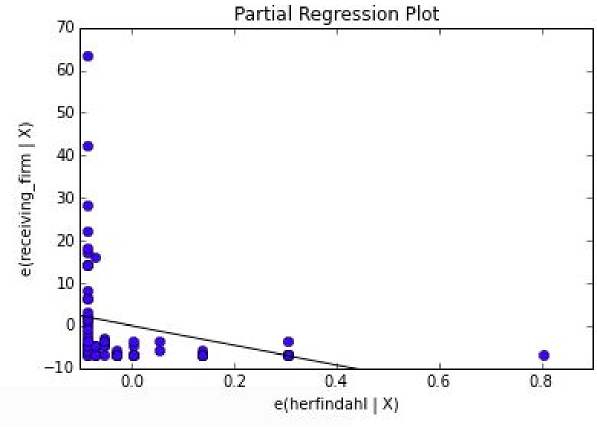
For the analysis I used the herfindahl index on employees as a metric to measure the ethnic composition of the firm. In order to track inventor mobility I looked at an inventor appearing on patents for several different firms.

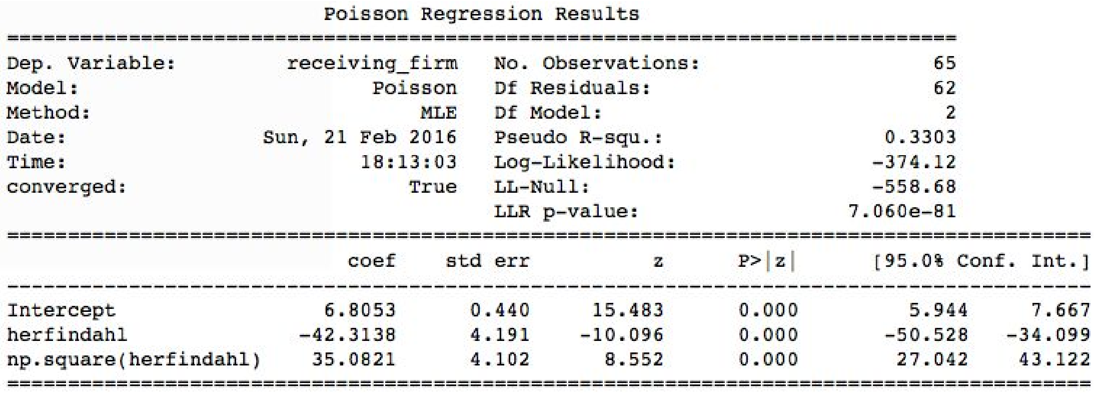
For the regression I used a Poisson model to fit the data as the dependent variable is a tally of the total number of inventors that move to the firm in question. I also included the herfindahl index squared as after a certain level the high diversity of a company could have an adverse effect.

$$\text{receiving\_firm} = \exp\left(5.483 - \underset{[4.191]}{42.314}\ \text{herfindahl} + \underset{[4.102]}{35.082}\ \text{herfindahl}^2\right)$$

Pseudo-R^2 : 0.330
Sample size: 65

The results imply that the ethnic composition of a firm will negatively impact an inventor's decision to move to that firm up until a certain level. The value of the pseudo implies that the model is of very good fit and the p-value of herfindahl and herfindahl ^{2} imply the variables are statistically significant. To improve on the model it would be interesting to add a variable for the increased salary the inventor might be receiving after changing jobs.





```
                          Poisson Regression Results
==============================================================================
Dep. Variable:         receiving_firm   No. Observations:                   65
Model:                        Poisson   Df Residuals:                       62
Method:                           MLE   Df Model:                            2
Date:                Sun, 21 Feb 2016   Pseudo R-squ.:                  0.3303
Time:                        18:13:03   Log-Likelihood:                -374.12
converged:                       True   LL-Null:                       -558.68
                                        LLR p-value:                 7.060e-81
=========================================================================================
                            coef    std err          z      P>|z|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------------
Intercept                 6.8053      0.440     15.483      0.000         5.944     7.667
herfindahl              -42.3138      4.191    -10.096      0.000       -50.528   -34.099
np.square(herfindahl)    35.0821      4.102      8.552      0.000        27.042    43.122
=========================================================================================
```

Endogeneity issues that are likely to exist in this model would be related to the measures firm size, firms' HR and recruitment budget, the firm's reputation, team members or manager turnover to mention a few. Furthermore controlling for the temporal factor that would show how the variables changes over time including the age of the moving inventors would improve the quality of the analysis.

The Herfindahl index may give a good indicator of ethnic composition in the firm, however it is limited by not considering inventors carrying last names that do not accurately describe what they identify as such as 2nd, 3rd or even later generation immigrants. By consequence, basing the measurement of inventors' last name may therefore yield inaccurate measurements.

The true value of this data set is undeniably the ability to track inventor mobility within the industry. However caution should be made in relying too heavily on these results as inventor's may appear less frequently as authors on patents as they go up the ranks into managerial and positions with administrative responsibilities.

APPENDIX 1A

i) Python script for cleaning and reshaping data
ii) Jupyter notebook containing statistical analysis of the data
iii) Parsed data

APPENDIX 1B

i) Python script for cleaning and reshaping data
ii) Jupyter notebook containing statistical analysis of the data
iii) Parsed data