Moe Kyaw Thu  
Final Report for ISyE 7401 Advanced Statistical Modelling  
March 14th, 2024

# Analysis of Citation Impact using Scisci Net Database

## Introduction

In this age of exponential growth in research publications, quantifying their academic productivity has become crucial to understanding their impacts. Among different methods, bibliometric analysis provides a way to measure the impact and productivity of researchers, institutions, and journals with metrics such as citation counts, h-index, and journal impact scores. For example, Web of Science (WoS) and Scopus are widely used databases that consist of scholarly literature across various disciplines, offering researchers comprehensive platforms for conducting bibliometric analysis. Undeniably, while these are great sources for bibliometric analysis, their extensive features often accompany significant expenses, limiting access for those unable to afford it. In light of this, this study introduces Scisci Net database – a largescale open data base for the bibliometric research (Yin et al. 2023).[[1]](#footnote-1)Against this backdrop, this study aims to analyze the citation impact using Scisci Net database.

## Data and Method

Since its launch in 2023, the Scisci Net repository has become a major venue for research dissemination of bibliometrics for various disciplines. As an open-source platform, it provides an equal opportunity for researchers to understand the impact of academic productivity. In total, it provides a total of 20 datasets, ranging from 311 to over one billion observations. Among the 20 datasets, this study chooses ‘SciSciNet\_Authors\_Gender’ database. In essence, this database is chosen because it focuses on an author-level metric that measures both the productivity and citation impact of the publications of an individual researcher. Based on this database, this study asks the following question: How does an author’s citation impact, as defined by the variable H.index,[[2]](#footnote-2) vary based on predictors associated with productivity, average count of 10-year citations, inference source and counts, and probability of an individual being female?

Before answering the question, this study first cleans and subsets the data. Initially, this dataset has a total of over 134 million observations since it is linked to a variety of sources, including Web of Science and Scopus. However, due to computational constraints, this study opts to analyze only 5000 observations, representing a relatively small sample size. From these 5000 observations, this study further eliminates the NA values, which leave only around 4000 observations, a sizeable number. Additionally, for the sake of regression analysis, this study also subsets by leaving only variables H.index, Productivity, Average\_C10, Average\_LogC10, Inference\_Sources, Inference\_Counts, and P.gf, while leaving out variables Author\_ID and Author\_Name. Finally, as the variables Average\_C10, Average\_LogC10, and P.gf. are expressed as decimals, this study rounds them up to two decimal places for consistency.

Additionally, the following table describes the full description of the dataset (the subset variables are bolded):

|  |  |
| --- | --- |
| Variable(s) | Description(s) |
| Author\_ID | Microsoft Academic Graph’s Author ID of the author. |
| Author\_Name | Original name of the author. |
| **Productivity** | Total number of publications of the author |
| **H.index** | H-index of the author |
| **Average\_C10** | Average 10-year citation of an individual author |
| **Average\_LogC10** | Average log of 10-year citation of an individual author |
| **Inference\_Sources** | The number of name-gender inference source datasets |
| **Inference\_Counts** | The number of empirical count of humans with the first name and  gendered label in the source datasets |
| **P.gf.** | The probability that indicates to what extent a name belongs to an  individual gendered female |

**Table.1. Codebook for SciSciNet\_Authors\_Gender Dataset**

To assess the impact of predictors on the response variable H.index, the following regression model is presented:

**Fig.1. Full Regression Model**

In addition to the linear regression analysis for the full model, this study further conducts procedures. Firstly, based on the outcomes, a correlation test is conducted to evaluate the relationships among predictor variables. Subsequently, additional regression procedures are conducted to assess the individual impact of each predictor on the response variable. Diagnostics procedures were conducted to assess the regression model; they include Variance Inflation Factor (VIF) test for potential multicollinearity, Shapiro-Wilk test for normality, Breusch-Pagan test for non-constant variance assumption, and identification for leverage point, outliers, and influential points. In doing so, these tests collectively inform about the model’s performance and potential issues.

The third stage encompasses several procedures. Various model selection techniques, namely regression sub-setting with adjusted R-squared and Cp criteria, and stepwise regression are used to identify the most suitable regression model for predicting the response variable. Outside of the model selection, regularization techniques, namely ridge regression, non-negative Garrote regression, and Lasso regression are applied to address multicollinearity and overfitting while enhancing the accuracy of the model. Overall, these procedures aimed to find a parsimonious model with strong predictive power for the response variable H.index.

## Results

**Linear Regression Modeling**

In the full model, the results indicate that variables Productivity, Average\_C10, Average\_LogC10, and Inference\_Sources significantly predict H.index, whereas Inference\_Counts and P.gf. do not exhibit statistical significance (See appendix figure 3 for full model output). However, it should be noted that the intercept has a negative coefficient of -1.156, meaning when all predictors are zero, the value of H.index is negative. Further analysis of the model reveals that the overall model is statistically significant and explains approximately 83.3% of the variance in H.index, as indicated by the adjusted R-squared value. Overall, the model is able to highlight the impact of academic productivity predictors on citation impact (See table 2 for full model regression output).

A screenshot of a computer

Description automatically generated

**Table.2. Full Regression Model Output**

While the model’s outcomes are statistically significant with the exception of two variables, several regression modelling approaches show different results. For instance, by conducting correlation plot (see appendix pg. 4), Productivity and P.gf. are negatively correlated to each other with -11 percent. In contrast, inference counts and inference sources as well as Average C10 and Average\_LogC10 are moderate to strongly correlated with each other. The correlation results highlight the importance of considering the dynamics between predictor variables when interpreting the results of regression modeling approaches.

Against the backdrop of the correlation plots, tested to evaluate the individual relationships between H.index and P.gf. as well as Productivity (See appendix pg. 10-11). In the full model between P.gf., Productivity, and H.index, P.gf. does not appear to have any significant effect due to the negative correlation between Productivity and P.gf. However, when tested individually with H.index in separate models, both variables are found to be statistically significant. It is important to note that, among the three models, the last model with only H.index and P.gf. explains a very small amount of variance, as indicated by its low R-squared value.

Based on the earlier correlation plot, this study also tests additional regression analysis due to the high correlation between inference counts and sources, and average citation count over 10 years and log of average citation count over 10 years. While these predictors collectively have a significant effect on the response variable, the model explains only a modest amount of the variability for H.index with R-squared value only at 18 percent, as compared to the earlier model with Productivity and P.gf., whose variability explains up to 73 percent (See appendix pg. 10-13).

Additionally, H.index is also tested with predictor variables by using interaction and squared terms. In interaction models, while a lot of the response variables are statistically significant and the model also explains a great variability up to 93 percent, there are a few things to be noted. For instance, while P.gf. alone has a borderline statistical significance, in interaction terms with Productivity, it is highly significant, and this comes as surprising as in the model "lm.prod.h.index", due to the negative correlation between the two, P.gf. does not have any significance at all. Due to this, further testing is performed by regressing the interaction terms of Productivity and P.gf. This study finds that by including squared and interaction term, all of the outcomes are statistically significant including the squared terms for P.gf. compared to the model "lm.prod.h.index". Additionally, the model explains almost 80 percent of the variability with H.index.

Lastly, this study also analyzes the quadratic terms between H.index and the predictor variables. The outcomes show that the quadratic term only explains 53 percent of the variability. In addition, when compared with the interaction model, Average\_C10 is highly significant and in contrast, Inference\_Counts in both of the models are not significant at all (See appendix pg. 13-16 for all model outputs for interaction and quadratic terms).

**Diagnostics**

After the linear regression analysis, this study conducts diagnostic procedures to assess various measures. In terms of multicollinearity, the VIF testing procedures show that there is generally little to moderate multicollinearity among the predictor variables for the original model. In contrast, when tested for normality for its residuals, the Breusch-Pagan test shows that there is heteroskedasticity in the model output. Upon testing for large leverage points, it was surprising to find that only observation 3845 stood out, suggesting that it is the sole standout in terms of its impact on the model. This study also examines Cook’s distance to identify influential observations, assessing model fit, and evaluating the assumptions underlying regression analysis. The results show that similar to leverage procedure, observation 3845 is the most influential point. In contrast and lastly, when detecting outliers, observation 3441 is the only outlier (See appendix pg. 16-18).

**Model Selection**

When conducting subset regression for exhaustive search, forward or backward stepwise, or sequential replacement, this study analyzes adjusted R-squared and Mallow’s Cp criteria. By analyzing the adjusted R-squared, the results show that 5 variables, namely Productivity, Average\_C10, Average\_LogC10, Inference\_Sources, and Inference\_Counts, explain up to 83.25929 percent whereas adding the sixth one only explains up to 83.25746. In contrast, upon evaluating the Cp criteria, this study observes that the subset model with 4 parameters, excluding Productivity, effectively strikes a balance between model complexity and explanatory power (See appendix pg. 19-21).

The stepwise regression process shows a result similar to Cp criteria model selection. Out of the outcomes assessed, the optimal model comprises the variables Productivity, Average\_C10, Average\_LogC10, and Inference\_Sources, with an AIC value of 11927.72. However, it should be noted that the full original model’s AIC value stands at 11927.75, with only 0.03 difference. This demonstrates that the reduced model, despite having fewer variables, offers a comparable level of fit to the data compared to the original model (See appendix pg. 21-23).

**Regularization**

Within regularization, ridge regression is performed to analyze the relationship between the dependent variable and predictors while mitigating multicollinearity issues and reducing overfitting. While scaling the original dataset, all predictor variables are included in the model with a sequence of lambda values ranging from 0 to 4000 specified for regularization. Based on the simulation, the optimal lambda of 1.8 is selected based on Generalized Cross-Validation (GCV) requirement. Based on this lambda value of 1.8, additional ridge regression is performed. The outcomes are as follows: Modified HKB estimator is 0.9182103 and modified L-W estimator is 0.8043325. These results show that the H.index’s potential values are more likely to be centered. The coefficients, after rounding them up to 4 decimals further suggest that the five variables Productivity, Average\_C10, Average\_LogC10, Inference\_Sources, and Inference\_Counts are positively associated with H.index, while P.gf. is not (See appendix pg. 23-25)

As a next step, this study also performs non-negative garrote to estimate the coefficients for predicting the response variable. Based on this, lambda was first chosen using GCV technique. Based on the resulting lambda of 0.00008791419, non-negative garrote is also performed. The resulting coefficients suggest that except for variables Interence\_Counts and P.gf., the rest of the variables have a non-zero impact on the predicted value of the response variable, as indicated by their non-zero coefficients in the fitted model. However, it should be noted that the magnitude of these negative coefficients is relatively small compared to the positive coefficients, suggesting a weak negative influence on H. index. In contrast, variable Productivity has the largest coefficient, suggesting its strong impact on the H. index compared to others (See appendix pg. 24-25).

As the last segment of the regularization, Lasso regression using Cp with ‘lars’ package and leave-one-out cross validation is performed. The Cp statistics are used here as means to measure model fit, where the lower the values with more variables added, the better. In this case, it can be seen that parameters 3 to 6 are the best fit. This means that inclusion of variables other than Productivity demonstrates a good fit. With leave-one-out cross validation, the optimal lambda here is 0.003309913, indicating that while there is less regularization, the model can fit into the data closely. The coefficient outcomes also suggest that apart from P.gf., the rest of the variables are positively associated with H.index; however, similar to non-negative garrote, the magnitude of negative association between H.index and P.gf. are minute (See appendix pg. 25-26).

## Discussion and Conclusion

Based on these testing procedures, this study shows that the author's citation impact, as measured by the variable H.index, varies based on several predictors. Between linear, ridge, and lasso regressions, there are several patterns to be noted. First and foremost, there is a negative association between H.index and P.gf. This means that if a researcher’s gender is likely to be a female, it is likely that they will have less H.index, compared to being a male gender. Secondly, Productivity is always positively associated with H.index; this demonstrates that an author’s academic output in terms of the number of papers can lead to having a higher H.index. Thirdly, the results suggest that variables namely Average\_C10, Average\_LogC10, and Inference\_Sources also play a significant role in predicting H.index. In the case of the former two, this suggests that authors with a higher average citation over a ten-year period tend to have a higher H.index. This likely stems from the fact that increased publication output enhances the visibility of an author’s work, thereby increasing the likelihood of citations. As for Inference\_Sources, it could possibly suggest that authors who draw from a greater number of sources for gender inference tend to have a higher H.index. Lastly, in terms of model performance, regularization techniques show that the use of all predictor variables on H.index yields varying degrees of success.

In terms of model criteria, several things should be noted as well. One challenge is that by using all variables as predictors in the dataset for regression analysis, it poses the challenge of overfitting. In fact, as a point in case, in ridge regression, the small lambda value of 1.8 may show signs of overfitting. However, as the dataset is compiled from across all available scholarly search engines, it offers a comprehensive representation of various academic contexts and disciplines, which helps to mitigate concerns related to overfitting. With regards to the variable P.gf., as it is based on the probability that an author’s gender could be female, one suggestion is to explore alternative methods for gender inference to enhance the accuracy and reliability of this predictor. To achieve this, one way is to incorporate national language processing techniques to refine gender classification.

However, limitations accompanied by future research directions also exist for this study. Primarily, this study is meant to serve as an introduction to the utilization of the Scisci Net database, limiting the analysis to 5000 observations and only around 4000 observations left after eliminating the NAs. However, for future investigations, expanding the analysis to include all 134 million observations could be advantageous, particularly with improved computing capabilities. By doing so, this would enable researchers to conduct more robust analysis between predictors and H.index, and more importantly, due to the large number of observations, it could generalize findings to the entire research community across various disciplines. Moreover, considering the absence of disciplines in the Scisci Net dataset, a potential approach is to associate each author with their respective discipline and investigate how disciplinary differences might impact the relationship between all predictor variables and H.index.

## References

Yin, Yian (2023). SciSciNet: A large-scale open data lake for the science of science research. figshare. Collection. <https://doi.org/10.6084/m9.figshare.c.6076908.v1>

## Appendix

#Reading and cleaning data set  
#Description:   
df <- read.delim("scisci\_gender.tsv", nrow = 5000) #Sample size of 5000 out of 130 million rows for illustrative purposes

#Cleaning data set  
#Omitting NAs  
df.clean <- na.omit(df)  
#Total number of rows without NAs: 3991 rows  
  
#Sub setting to clean out unnecessary columns for better analysis  
df.clean.sub <- subset(select=c("H.index", "Productivity", "Average\_C10", "Average\_LogC10", "Inference\_Sources", "Inference\_Counts", "P.gf."), df.clean)  
#Rounding up numbers  
df.clean.sub$Average\_C10 <- round(df.clean.sub$Average\_C10, 2)  
df.clean.sub$Average\_LogC10 <- round(df.clean.sub$Average\_LogC10, 2)  
df.clean.sub$P.gf. <- round(df.clean.sub$P.gf., 2)

#Data Exploration and Visualization  
visnorm <- function(dataframe) {  
 par(mfrow=c(2, ncol(dataframe) %/% 2 + 1))  
 for (col in names(dataframe)) {  
 hist(dataframe[[col]], main=paste("Histogram of", col), xlab=col, col="blue", border="red")  
 qqnorm(dataframe[[col]], main=paste("QQ Plot of", col))  
 qqline(dataframe[[col]])  
 }  
 par(mfrow=c(4, 4)) # Reset the plotting layout  
}  
#Visualizing Normality  
visnorm(df.clean.sub)

A group of graphs and diagrams

Description automatically generatedA screenshot of a graph

Description automatically generated

#Interpretation: Based on the visualization, the data, after omitting NAs, are not distributed normally. In fact, almost all of the columns are skewed in either of the directions.

#Basic Linear Regression Analysis  
lm.h.index <- lm(H.index ~., data = df.clean.sub)  
summary(lm.h.index)

##   
## Call:  
## lm(formula = H.index ~ ., data = df.clean.sub)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -42.847 -1.494 0.100 0.811 51.399   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.156e+00 2.371e-01 -4.878 1.11e-06 \*\*\*  
## Productivity 1.710e-01 1.385e-03 123.427 < 2e-16 \*\*\*  
## Average\_C10 1.136e-02 1.441e-03 7.887 3.96e-15 \*\*\*  
## Average\_LogC10 1.693e+00 6.341e-02 26.697 < 2e-16 \*\*\*  
## Inference\_Sources 2.229e-02 9.165e-03 2.432 0.0151 \*   
## Inference\_Counts 5.609e-08 7.466e-08 0.751 0.4525   
## P.gf. -1.697e-01 1.626e-01 -1.044 0.2966   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.454 on 3984 degrees of freedom  
## Multiple R-squared: 0.8328, Adjusted R-squared: 0.8326   
## F-statistic: 3308 on 6 and 3984 DF, p-value: < 2.2e-16

#Interpretation: The results show that variables Productivity, Average\_C10, Average\_LogC10, and Inference\_Sources are statistically significant predictors of H.index, while Inference\_Counts and P.gf. are not statistically significant.

#Additional regression procedures  
#Correlation between predictors  
corr <- round(cor(df.clean.sub[,-1]),2)  
corrplot(corr, method = "circle", type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)

A diagram of a number of numbers

Description automatically generated with medium confidence

#Interpretation: Of all the correlations, Productivity and P.gf. are weakly correlated to each other with -11 percent. In constrast inference counts and inference sources as well as Average C10 and Average\_LogC10 are moderate to strongly correlated with each other.  
  
#Relationship between productivity and female gender probability  
lm.prod.h.index <- lm(H.index ~ P.gf. + Productivity, data = df.clean.sub)  
lm.prod.h.index.2 <- lm(H.index ~ Productivity, data = df.clean.sub)  
lm.prod.h.index.3 <- lm(H.index ~ P.gf., data = df.clean.sub)  
summary(lm.prod.h.index)

##   
## Call:  
## lm(formula = H.index ~ P.gf. + Productivity, data = df.clean.sub)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -46.874 -2.411 -1.136 1.230 51.579   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.228522 0.118923 18.739 <2e-16 \*\*\*  
## P.gf. 0.180216 0.186136 0.968 0.333   
## Productivity 0.182253 0.001572 115.948 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.185 on 3988 degrees of freedom  
## Multiple R-squared: 0.7731, Adjusted R-squared: 0.773   
## F-statistic: 6795 on 2 and 3988 DF, p-value: < 2.2e-16

summary(lm.prod.h.index.2)

##   
## Call:  
## lm(formula = H.index ~ Productivity, data = df.clean.sub)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -46.836 -2.486 -1.041 1.240 51.759   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.303594 0.090169 25.55 <2e-16 \*\*\*  
## Productivity 0.182082 0.001562 116.58 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.185 on 3989 degrees of freedom  
## Multiple R-squared: 0.7731, Adjusted R-squared: 0.773   
## F-statistic: 1.359e+04 on 1 and 3989 DF, p-value: < 2.2e-16

summary(lm.prod.h.index.3)

##   
## Call:  
## lm(formula = H.index ~ P.gf., data = df.clean.sub)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.538 -5.939 -3.910 1.462 130.682   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.5381 0.2294 32.85 < 2e-16 \*\*\*  
## P.gf. -2.2426 0.3866 -5.80 7.14e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.84 on 3989 degrees of freedom  
## Multiple R-squared: 0.008363, Adjusted R-squared: 0.008114   
## F-statistic: 33.64 on 1 and 3989 DF, p-value: 7.145e-09

#Interpretation: These three models are tested to evaluate the individual relationships between H.index and that of P.gf. and Productivity. It turns out that due to the negative correlation between Productivity and P.gf., in the model "lm.prod.h.index", P.gf. does not have any significant. However, tested individually with H.index, both variables are statistically significant. However, it should be noted that of all the three models, the last model with H.index and P.gf. only explain very small amount due to low R-squared.  
  
#Relationship between H index, average C\_10 and infer  
lm.h.index.c10.infer <- lm(H.index ~ Average\_C10 + Average\_LogC10 + Inference\_Sources + Inference\_Counts, data = df.clean.sub)  
summary(lm.h.index.c10.infer)

##   
## Call:  
## lm(formula = H.index ~ Average\_C10 + Average\_LogC10 + Inference\_Sources +   
## Inference\_Counts, data = df.clean.sub)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -46.248 -4.281 -1.466 0.609 120.180   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.099e-01 5.103e-01 -0.803 0.421901   
## Average\_C10 1.955e-02 3.185e-03 6.139 9.13e-10 \*\*\*  
## Average\_LogC10 2.862e+00 1.384e-01 20.683 < 2e-16 \*\*\*  
## Inference\_Sources 7.221e-02 2.023e-02 3.569 0.000362 \*\*\*  
## Inference\_Counts -2.781e-07 1.628e-07 -1.708 0.087625 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.855 on 3986 degrees of freedom  
## Multiple R-squared: 0.1811, Adjusted R-squared: 0.1802   
## F-statistic: 220.3 on 4 and 3986 DF, p-value: < 2.2e-16

#Interpretation: While the predictors collectively have a significant effect on the response variable, the model explains only a modest amount of the variability for H.index.  
  
#Full model with consideration for interaction terms  
lm.prod.full <- lm(H.index ~.^2-1, data = df.clean.sub)  
summary(lm.prod.full)

##   
## Call:  
## lm(formula = H.index ~ .^2 - 1, data = df.clean.sub)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -49.389 -1.243 -0.490 0.853 21.698   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## Productivity 5.239e-02 5.548e-03 9.443 < 2e-16 \*\*\*  
## Average\_C10 7.411e-02 7.239e-03 10.237 < 2e-16 \*\*\*  
## Average\_LogC10 1.035e+00 1.431e-01 7.229 5.81e-13 \*\*\*  
## Inference\_Sources 1.768e-02 6.329e-03 2.794 0.005231 \*\*   
## Inference\_Counts 1.044e-06 1.756e-06 0.595 0.552027   
## P.gf. 6.939e-01 3.935e-01 1.763 0.077893 .   
## Productivity:Average\_C10 1.494e-04 3.326e-05 4.493 7.24e-06 \*\*\*  
## Productivity:Average\_LogC10 4.790e-02 1.847e-03 25.937 < 2e-16 \*\*\*  
## Productivity:Inference\_Sources -1.242e-03 1.678e-04 -7.399 1.66e-13 \*\*\*  
## Productivity:Inference\_Counts 1.140e-08 1.392e-09 8.195 3.36e-16 \*\*\*  
## Productivity:P.gf. 3.063e-02 2.962e-03 10.341 < 2e-16 \*\*\*  
## Average\_C10:Average\_LogC10 -1.526e-02 1.044e-03 -14.619 < 2e-16 \*\*\*  
## Average\_C10:Inference\_Sources -1.862e-04 1.952e-04 -0.954 0.340195   
## Average\_C10:Inference\_Counts 4.653e-09 1.379e-09 3.373 0.000750 \*\*\*  
## Average\_C10:P.gf. 2.021e-02 2.880e-03 7.015 2.69e-12 \*\*\*  
## Average\_LogC10:Inference\_Sources 1.883e-02 6.002e-03 3.137 0.001719 \*\*   
## Average\_LogC10:Inference\_Counts -2.341e-07 6.125e-08 -3.822 0.000134 \*\*\*  
## Average\_LogC10:P.gf. -8.394e-01 1.180e-01 -7.114 1.33e-12 \*\*\*  
## Inference\_Sources:Inference\_Counts -2.950e-08 5.289e-08 -0.558 0.577055   
## Inference\_Sources:P.gf. -2.358e-02 1.596e-02 -1.477 0.139710   
## Inference\_Counts:P.gf. 1.017e-07 2.277e-07 0.447 0.655199   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.555 on 3970 degrees of freedom  
## Multiple R-squared: 0.9227, Adjusted R-squared: 0.9223   
## F-statistic: 2258 on 21 and 3970 DF, p-value: < 2.2e-16

#Interpretation: While a lot of the response variables are statistically significant and the model also explains a great variability, there are a few things to be noted. For instance, while P.gf. alone has a borderline statistical significance, in interaction terms with Productivity, it is highly significant and this comes as surprising as in the model "lm.prod.h.index", due to the negative correlation between the two, P.gf. does not have any significance at all. Due to this, we do further testing:  
lm.prod.full.sub <- lm(H.index ~ Productivity + P.gf. + Productivity \* P.gf. + Productivity ^ 2 + P.gf. ^ 2, data = df.clean.sub)  
summary(lm.prod.full.sub)

##   
## Call:  
## lm(formula = H.index ~ Productivity + P.gf. + Productivity \*   
## P.gf. + Productivity^2 + P.gf.^2, data = df.clean.sub)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -57.001 -2.196 -1.144 1.220 42.807   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.443772 0.120936 20.207 < 2e-16 \*\*\*  
## Productivity 0.175025 0.001797 97.393 < 2e-16 \*\*\*  
## P.gf. -0.500524 0.202915 -2.467 0.0137 \*   
## Productivity:P.gf. 0.033206 0.004104 8.091 7.79e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.144 on 3987 degrees of freedom  
## Multiple R-squared: 0.7768, Adjusted R-squared: 0.7766   
## F-statistic: 4625 on 3 and 3987 DF, p-value: < 2.2e-16

#When tested with these three variables by including squared and interaction term, all of the outcomes are statistically significant including the squared terms for P.gf. compared to model "lm.prod.h.index". Additionally, the model explains almost as 80 percent of the variability with H.index.  
  
#Full model with quadratic terms  
lm.prod.full.quad <- lm(H.index ~I(Productivity^2) + I(Average\_C10^2)+I(Average\_LogC10^2)+I(Inference\_Sources^2)+I(Inference\_Counts^2)+I(P.gf.^2), data = df.clean.sub)  
summary(lm.prod.full.quad)

##   
## Call:  
## lm(formula = H.index ~ I(Productivity^2) + I(Average\_C10^2) +   
## I(Average\_LogC10^2) + I(Inference\_Sources^2) + I(Inference\_Counts^2) +   
## I(P.gf.^2), data = df.clean.sub)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -83.110 -3.147 -1.883 1.512 58.867   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.416e+00 2.736e-01 8.832 < 2e-16 \*\*\*  
## I(Productivity^2) 3.572e-04 6.042e-06 59.113 < 2e-16 \*\*\*  
## I(Average\_C10^2) -2.332e-06 1.916e-06 -1.217 0.22361   
## I(Average\_LogC10^2) 5.697e-01 2.231e-02 25.530 < 2e-16 \*\*\*  
## I(Inference\_Sources^2) 1.148e-03 3.521e-04 3.261 0.00112 \*\*   
## I(Inference\_Counts^2) -8.558e-15 2.417e-14 -0.354 0.72329   
## I(P.gf.^2) -1.285e+00 2.713e-01 -4.736 2.25e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.417 on 3984 degrees of freedom  
## Multiple R-squared: 0.5364, Adjusted R-squared: 0.5357   
## F-statistic: 768.2 on 6 and 3984 DF, p-value: < 2.2e-16

#Interpretation: Compared to the interaction terms, R-squared for quadratic terms only explain for over 53 percent of the variation, whereas the interaction models explain up to 93 percent. Additionally, there are a few more things to notice. the quadratic term for Average\_C10 is not significant (p = 0.22361), whereas in the interaction model, Average\_C10 is highly significant (p < 0.001), as well as some of its interaction terms. When logged with Average\_LogC10, the terms become significant. For Inference\_Counts, none of them are significant in both of the models.

#Diagnostics  
#VIF testing  
print(vif(lm.h.index))

## Productivity Average\_C10 Average\_LogC10 Inference\_Sources   
## 1.066610 1.404166 1.459449 1.259957   
## Inference\_Counts P.gf.   
## 1.286424 1.047192

#Interpretation: The VIF values indicate that there is generally little to moderate multicollinearity among the predictor variables for the original model.  
  
#Normality test  
print(shapiro.test(lm.h.index$residuals))

##   
## Shapiro-Wilk normality test  
##   
## data: lm.h.index$residuals  
## W = 0.78048, p-value < 2.2e-16

#Interpretation: We reject the null hypothesis that the residuals are normally distributed.  
  
#Constant Variance assumption  
plot(lm.h.index$fitted.values, lm.h.index$residuals)

A graph of a number of dots

Description automatically generated with medium confidence

print(bptest(lm.h.index))

##   
## studentized Breusch-Pagan test  
##   
## data: lm.h.index  
## BP = 1412, df = 6, p-value < 2.2e-16

#We can reject the null hypothesis constant variance and assume that there is evidence of heteroscedasticity in the model output. The plotting of residuals and fitted values further confirm that there is hetereoskedascity.   
  
#Large leverage points.  
X <- model.matrix(lm.h.index)  
H <- X%\*%solve(t(X)%\*%X)%\*%t(X) #Hat Values  
plot(diag(H))  
abline(h=2\*7/5000,col=1)  
text(x = 1:length(diag(H)), y = diag(H), labels = 1:length(diag(H)), pos = 2) #Identification

A graph with numbers and lines

Description automatically generated

#Interpretation: The largest leverage point is the observation 3845.  
  
#Influential Points  
cook.dist.prod <- cooks.distance(lm.h.index)  
halfnorm(cook.dist.prod,2,ylab="Cook’s distances")

A graph with numbers and a line

Description automatically generated

#Interpretation: Same goes for Cook's distance as observation 3845 is the most influential point.   
  
#Searching for outliers  
estar <- rstudent(lm.h.index)  
plot(lm.h.index$fitted.values, estar, xlab = "Fitted values", ylab = "Studentized residuals")  
abline(h=2\*7/5000,col=1)  
text(x = lm.h.index$fitted.values, y = estar, labels = 1:length(lm.h.index$fitted.values), pos = 2)

A graph with numbers and a black dot

Description automatically generated

#Interpretation: It seems the outliers is the observation 3441.

#Model Sub setting  
require(leaps)

## Loading required package: leaps

subset <- regsubsets(H.index ~ ., data = df.clean.sub)  
rs <- summary(subset)  
print(rs$which)

## (Intercept) Productivity Average\_C10 Average\_LogC10 Inference\_Sources  
## 1 TRUE TRUE FALSE FALSE FALSE  
## 2 TRUE TRUE FALSE TRUE FALSE  
## 3 TRUE TRUE TRUE TRUE FALSE  
## 4 TRUE TRUE TRUE TRUE TRUE  
## 5 TRUE TRUE TRUE TRUE TRUE  
## 6 TRUE TRUE TRUE TRUE TRUE  
## Inference\_Counts P.gf.  
## 1 FALSE FALSE  
## 2 FALSE FALSE  
## 3 FALSE FALSE  
## 4 FALSE FALSE  
## 5 FALSE TRUE  
## 6 TRUE TRUE

#Choosing Model based on adjusted-r squared  
print(which.max(rs$adjr2))

## [1] 5

plot(2:7,rs$adjr2,xlab="No. of Parameters",ylab="Adjusted R-square")

A graph with numbers and lines

Description automatically generated

#Interpretation: 5 variables explain up to 83.25929 percent whereas adding the sixth one only explains up to 83.25746.  
  
#Choosing Model based on Cp criteria  
names <- colnames(rs$which)  
plot(2:7, rs$cp, xlab="No. of Parameters",ylab="Cp Statistic")  
abline(a = 0, b = 1, col = "red")  
text(2:7, rs$cp, labels = names, pos = 4, col = "blue", cex = 0.5)

A graph with numbers and lines

Description automatically generated

#Interpretation: 5 parameters explain with the Cp criteria. Thus, the selected subset model with 5 parameters does a good job of maintaining a balance between model complexity and explanatory powers.

#Stepwise Regression using AIC as an identification  
stepwise.lm.h.index <- stepAIC(lm.h.index)

## Start: AIC=11929.75  
## H.index ~ Productivity + Average\_C10 + Average\_LogC10 + Inference\_Sources +   
## Inference\_Counts + P.gf.  
##   
## Df Sum of Sq RSS AIC  
## - Inference\_Counts 1 11 79031 11928  
## - P.gf. 1 22 79041 11929  
## <none> 79020 11930  
## - Inference\_Sources 1 117 79137 11934  
## - Average\_C10 1 1234 80254 11990  
## - Average\_LogC10 1 14137 93157 12585  
## - Productivity 1 302162 381182 18208  
##   
## Step: AIC=11928.31  
## H.index ~ Productivity + Average\_C10 + Average\_LogC10 + Inference\_Sources +   
## P.gf.  
##   
## Df Sum of Sq RSS AIC  
## - P.gf. 1 28 79059 11928  
## <none> 79031 11928  
## - Inference\_Sources 1 187 79218 11936  
## - Average\_C10 1 1238 80269 11988  
## - Average\_LogC10 1 14235 93266 12587  
## - Productivity 1 302993 382023 18215  
##   
## Step: AIC=11927.72  
## H.index ~ Productivity + Average\_C10 + Average\_LogC10 + Inference\_Sources  
##   
## Df Sum of Sq RSS AIC  
## <none> 79059 11928  
## - Inference\_Sources 1 190 79249 11935  
## - Average\_C10 1 1242 80301 11988  
## - Average\_LogC10 1 14213 93272 12586  
## - Productivity 1 308323 387382 18268

summary(stepwise.lm.h.index)

##   
## Call:  
## lm(formula = H.index ~ Productivity + Average\_C10 + Average\_LogC10 +   
## Inference\_Sources, data = df.clean.sub)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -42.903 -1.491 0.112 0.805 51.198   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.273318 0.221667 -5.744 9.92e-09 \*\*\*  
## Productivity 0.171138 0.001373 124.680 < 2e-16 \*\*\*  
## Average\_C10 0.011400 0.001440 7.914 3.20e-15 \*\*\*  
## Average\_LogC10 1.690772 0.063161 26.769 < 2e-16 \*\*\*  
## Inference\_Sources 0.025504 0.008236 3.097 0.00197 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.454 on 3986 degrees of freedom  
## Multiple R-squared: 0.8327, Adjusted R-squared: 0.8326   
## F-statistic: 4961 on 4 and 3986 DF, p-value: < 2.2e-16

#Interpretation: The lowest AIC value is 11927.72. The variables are Productivity, Average\_C10, Average\_LogC10, and Inference\_Sources. However, it should be noted that the original model with all variables have an AIC of 11927.75; this indicates that the reduced model, despite having fewer variables, provides a similar level of fit to the data compared to the original model.

#Ridge regression  
df.clean.sub.scaled <- data.frame(scale(df.clean.sub))  
lm.h.index.ridge <- lm.ridge(H.index ~., data = df.clean.sub.scaled, lambda=seq(0,4000,.01))  
#Choosing GCV based on maximum and minimum values  
GCV.lambda <- which.min(lm.h.index.ridge$GCV)  
print(lm.h.index.ridge$lambda[GCV.lambda])

## [1] 1.8

#Plotting  
matplot(lm.h.index.ridge$lambda, coef(lm.h.index.ridge), type = "l", xlab = expression(lambda), ylab = expression(hat(beta)), col = 1)  
abline(v = lm.h.index.ridge$lambda[GCV.lambda], col = "red")

A graph of a number of numbers

Description automatically generated with medium confidence

#Interpretation: The optimal lambada here is 1.8. This small lambada value means that the model fit the data more closely.  
  
#Running ridge regression based on lambda value of 1.8  
select(lm.h.index.ridge)

## modified HKB estimator is 0.9182103   
## modified L-W estimator is 0.8043325   
## smallest value of GCV at 1.8

lm.h.index.ridge <- lm.ridge(H.index~.,lambda=1.8,df.clean.sub.scaled)  
coef(lm.h.index.ridge)

## Productivity Average\_C10 Average\_LogC10   
## 5.523934e-17 8.253552e-01 6.059364e-02 2.088798e-01   
## Inference\_Sources Inference\_Counts P.gf.   
## 1.770234e-02 5.504902e-03 -6.958837e-03

round(coef(lm.h.index.ridge),4)

## Productivity Average\_C10 Average\_LogC10   
## 0.0000 0.8254 0.0606 0.2089   
## Inference\_Sources Inference\_Counts P.gf.   
## 0.0177 0.0055 -0.0070

#Interpretation: Modified HKB estimator is 0.9182103 and modified L-W estimator is 0.8043325.

#Fitting Non negative Garrote using GCV  
scaled.df.clean.sub <- scale(as.matrix(df.clean.sub))  
X <- scaled.df.clean.sub[,-1]  
y <- scaled.df.clean.sub[,1]  
a.cv <- cv.nnGarrote(x=X,y=y)

## Performing cross-validation for 101 values of the shrinkage parameter "lambda.nng":  
## 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 | 80 | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 97 | 98 | 99 | 100 | 101 |

outcome <- nnGarrote(x=X,y=y,lambda.nng = a.cv$optimal.lambda.nng)  
names.garrote <- c("(Intercept)", colnames(X))  
name.df <- data.frame(Parameter = names.garrote, Coefficient = c(outcome$betas[1, 1], outcome$betas[-1, 1]))  
print(name.df)

## Parameter Coefficient  
## 1 (Intercept) 5.511187e-17  
## 2 Productivity 8.256694e-01  
## 3 Average\_C10 6.049699e-02  
## 4 Average\_LogC10 2.088836e-01  
## 5 Inference\_Sources 1.761974e-02  
## 6 Inference\_Counts 5.479097e-03  
## 7 P.gf. -6.843552e-03

#Interpretation: Except for variables Interence\_Counts and P.gf., the rest of the variables have an non-zero impact on the predicted value of the response variable, as indicated by their non-zero coefficients in the fitted model.

#Lasso Regression using Cp criteria  
y.2 <- scale(df.clean.sub$H.index)  
x.2 <- as.matrix(scale(df.clean.sub[,2:7]))  
a.lasso <- lars(x.2,y.2)  
summary(a.lasso)

## LARS/LASSO  
## Call: lars(x = x.2, y = y.2)  
## Df Rss Cp  
## 0 1 3990.0 19842.513  
## 1 2 1247.9 3466.276  
## 2 3 745.4 466.984  
## 3 4 673.2 37.886  
## 4 5 668.5 11.682  
## 5 6 668.0 10.725  
## 6 7 667.0 7.000

names.lasso <- colnames(x.2)  
plot(1:7,a.lasso$Cp)  
abline(0,1)  
text(2:7, a.lasso$cp, labels = names.lasso, pos = 3, col = "blue", cex = 0.5)

A graph of numbers and a rectangle

Description automatically generated

#Interpretation: Overall, the model demonstrates good performance in terms of model fit, predictor selection, and optimal model complexity. It also explains the observed data well.  
  
#Lasso with glmnet  
a.lasso <- glmnet(x.2,y.2,family="gaussian")  
a.cv <- cv.glmnet(x.2,y.2,family="gaussian",nfolds=5000)

## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per  
## fold

a.cv$lambda.min

## [1] 0.003309913

round(coef(a.lasso,s=a.cv$lambda.min),5)

## 7 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) 0.00000  
## Productivity 0.82345  
## Average\_C10 0.05859  
## Average\_LogC10 0.20745  
## Inference\_Sources 0.01561  
## Inference\_Counts 0.00386  
## P.gf. -0.00407

#Interpretation: The optimal lambda here is 0.003309913. The coefficients show that apart from P.gf., the rest are positively associated with the H.index.

1. For full database access, see: <https://springernature.figshare.com/collections/SciSciNet_A_large-scale_open_data_lake_for_the_science_of_science_research/6076908/1> [↑](#footnote-ref-1)
2. In this study, H-index is referred to an author-level metric that measures both the productivity and citation impact of the publications. [↑](#footnote-ref-2)