PREDICTING THE SCIENTIFIC SUCCESS OF NOBEL LAUREATES

A BRIEF STUDY OF THE TIME EVOLUTION OF THE $h-$INDICES OF NOBEL HONOURED SCIENTISTS

BY

Xabier Oyanguren Asua

Universitat Autònoma de Barcelona

COMPLEX DATA ANALYSIS

Bachelor's degree in
Computational Mathematics and Data Analytics

2019-2020

Contents

Predicting the Scientific Success of Nobel Laureates

A brief study of the time evolution of the *h*-index of Nobel Laureates

Abstract

In what follows, first of all, we assert that the quality of a researcher can quite reasonably be captured by the h-index. Then, a time analysis is performed to study the h-index of different Physics and Chemistry Nobel Laureates in search of a pattern in three different temporal directions: the h-index succession of different laureates in time and the h-index of the same laureates in different times, among other additional analyses. In the end, attending the h-index they had when they published the awarded paper, the h-index at prize time and ten years after, we try to build a predictive model.

1 Motivation

In 2005 Hirsch JE [1] proposed the following bibliometric index to assert the quality of scientific production of a researcher in time: the maximum number x such that this particular researcher has x papers with x or more citations. This was called the **h-index** of that researcher. Ever since it has become the nightmare of many, as it has become one of the standard criteria to assert the impact of the career of a scientist. As a joke, it is said that Newton for instance, used to take several years of recheck, before publishing his discoveries: nowadays he would be fired for lack of publishing. As it has been hardly criticized in the literature [2], [3], an unfortunate aphorism has become true in today's scientific community: *"Publish or Perish"*.

First things first, it is obvious that h-indices of academics should not be compared across disciplines [1], not even inside a same science. For instance, it is clear that the research output in experimental sciences is easily more boiling hot than the long term theoretical discoveries. Also, many times theoretical discoveries become part of the proper knowledge and are no longer necessarily cited when used (see the Schrödinger Equation for instance). Additionally, the h-index does not contemplate very largely cited papers. Once a publication counts for the x we were mentioning, its magnitude is not important. Therefore, this punishes punctual (but maybe profoundly transcendental) discoveries. It doesn't either check if the researcher's name is first in the paper (which in many fields indicates the highest contribution), or the number of authors of the paper in general.

"Potayto potahto", if a researcher has a high h-index, there is no doubt about the quality of his/her research output. Actually it means that for a long time, the paper production has been very successful in citations, which means was key in the development of many other studies. Therefore, although it has its drop-backs, the magnitude of the h-index is clearly correlated with a life-long success.

It is worth noting that the h-index is a time dependent scalar. Elder academics have had more years to generate knowledge than younger ones. Thus, in general there will be a correlation between the age and the h-index of a researcher. It should also be noted though, that the h-index and in general, the number of citations, are a reliable number only since IT times began. As such, if we want to study these bibliometric indices as a measure of success, it seems convenient to restrict ourselves to the last 30 years or so.

With all this in mind, we could study the time variation of this index in some distinguished scientific population: the Nobel Laureates for instance. The awards scientist receive are also indicative of their career's impact. Thus, being the quintessential award, the Nobel Prize, they seem to be an in teresting group to study. So our main question will be: *Can we predict the future scientific career of a Nobel Laureate?*

2 The Data

The most optimal thing would have been to find a comprehensive database of worldwide scientists, with their h-indices, number of citations and whether they were awarded a Nobel Prize or not. Of course, this does not exist. But the following two databases were found instead:

(α) On the one hand, there is a database by Ioannidis et al.[4], where they offer several bibliometric indicators (among them the total number of citations and h-index, excluding self-citations) of the 105.000 best scientists in the world across disciplines, according to a composite index they generate using the different indicators. The numeric values of the indicators were collected in *Scopus*, restricting the citation count from 1996 to 2018. Apparently, Scopus has only recently began to collect reference lists for papers published from 1960 to 1996. Therefore, for papers prior to 1996, only citations from 1996 to 2018 are included. Anyhow, for active scientists who began their carriers in the IT age, these are possibly the years of their academic activity.

(β) A database by Li, J., Yin, Y., Fortunato, S. et al.[5] of the Nobel Prizes in Physics, Chemistry and Medicine, that contains their names, the year they won the prize and the year they published the awarded discovery paper (and the paper's information).

Clearly, these two sets are not enough for the analysis we want to perform. Sure, we could try to find manually the hundreds of laureates of the last 30 years in the first database, but this would be absurd, as the indicators given in that dataset are the ones recorded at 2018. But our aim is to compare the laureates in equal footing, taking their indices in the year they were awarded. Therefore, motivated by these two databases, it was decided to build up some additional ones. A whole day of field work, using the recorded names and dates in database (β)[1] and searching for the h-indices till certain dates given by the *Web of Science* (WOS) database, correctly filtering the names (taking into account homonymous authors) and so on, the following four databases were generated:

(γ) Two Tables: one for Physics and one for Chemistry Laureates from 1990 to 2018, collecting the surname and names, the year they won the prize and the h-index they had that year according to WOS and the year in which they published the awarded paper. There are 72 different laureates in the Physics Table and 65 in the Chemistry Table.

(δ) Two Tables, one for Physics and one for Chemistry Laureates from 2000 to 2010, collecting the surname and names, the year they won the prize, the year they published their awarded discovery paper, the h-index they had in each of those two years and the h-index they had ten years after they were awarded the Nobel. There are 30 different laureates in Physics and 28 in Chemistry.

These data-sets and the R script used can be found in the following GitHub repository:

https://github.com/Oiangu9/Final_Practice_CDA_MatCAD

Also, some additional data was collected for the test in Section 3.6, which will be tabled in the same section.

[1] The databases will be referenced by the Greek letters we set next to their descriptions

3 Results and Discussion

3.1 Do we need to care about both, the h-index and the number of citations?

Before jumping into using the h-index as the only indicator of research quality, it seems interesting to legitimate such a decision, by noting that for most of the top researchers, the other big bibliometric index: the total number of citations, is positively correlated with it. As we mentioned previously, it is true that the h-index fails at recognizing authors with few but very cited papers. But let us hypothesize that at least among the recent top researchers, this is not generally the case. This will legitimize the use of the h-index for the rest of the present work.

In particular we will have enough by proving that the correlation is monotonically positive. That is, a big h-index implies a big number of citations and vice versa. We don't really care if the correlation is linear or has a particular non-linear shape. Thus, the contrast hypothesis will be the following: Denoting by (h_i, c_i) the h-index and total number of citations of the i-th researcher, we will assume that $\forall i \; h_i$ and c_i are *iid* (independent and identically distributed) according to the respective random variables $\mathcal{H}$ and C. Then, we seek to reject the following null hypothesis:

$$H_0 : \mathcal{H},\ C \ \textit{are not correlated}$$
$$H_1 : \mathcal{H},\ C \ \textit{are Monotonically Positively Correlated}$$

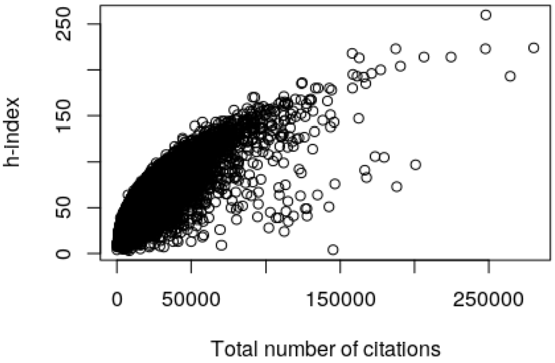

Figure 1: h-index vs total number of citations of the top 105.000 scientists according to [4]. Data of dataset (α).

In order to check this, dataset (α) was used, the one containing the h-index and the total citations (excluding self-citations) of the 105.000 top scientists of the world. You can see a plot showing the analyzed data in Figure 1. A correlation permutation test was considered using two different statistics:

- On the one hand, the not so restrictive Spearman Correlation Coefficient [6] was checked. This statistic takes a value in $(-1, 1)$ and only contemplates if data are monotonically correlated or not (the magnitude of the statistic) and the relation direction (its sign).
- We would have enough with that, but we still checked the Pearson correlation coefficient to see if there was actually a linear correlation (which would give more weight to the result).

For both cases an approximate permutation test was performed using 10.000 rearrangements of the h-indices (there are 105.000 samples, we could not take into account every possible permutation).

We obtained a Spearman correlation coefficient of 0.9076 and a Pearson correlation of 0.8442 ($R^2 = 0.7127$). Both with a p-value obtained form the permutation test of exactly 0. That is, non of the 10.000 permutations gave a Spearman nor Pearson correlation coefficient more positive than the one of the observed set. You can see the histograms of the statistics obtained in the rearrangements in Figure 2.

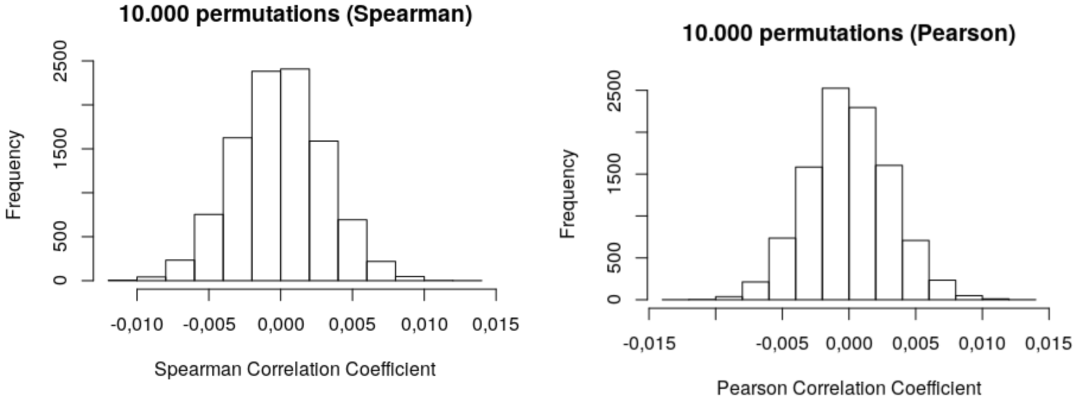

Figure 2: Histograms of the correlation coefficients obtained in the Permutation Test with 10.000 rearrangements of Section 3.1.

Therefore, we could actually reject the null hypothesis with practically any significance level. And thus, we conclude that we can safely use the h-index alone, as it is significantly, even linearly, positively correlated with the total number of citations. One of them being big implies the other is proportionally big as well. Thus, tracking both indices would not give any additional value to the following tests. This result is what legitimated to build the datasets (γ) and (β) only using the h-index.

We won't care about it, but we could build a linear least square model relating h-index and the total number of citations, and we would get that:

$$h_i = \beta_0 + \beta_1 c_i = 28.63 + 1.466e - 3c_i$$

with a classical regression model p-value $< 2e - 16$ for β_1, indicating its high significance (non-nullity). We see that effectively, the coefficient is positive.

3.2 Any pattern in time for h-indices of Nobel Laureates?

In a first approach to the h-indices of Nobel Laureates, they should not be mixed, as it could happen that they are time dependent. That is, there might be a pattern in the h-index of Nobel laureates of different years. For instance, one could easily suggest a hypothesis that as time goes on, the scientific community tends to give more value to a life-long work (higher h-index) than to punctual discoveries young scientist could find (rather a lower h-index). Alternatively, one could think that the laureate selection mechanism has some kind of bias in time. If any of these was true, then we could abstract the pattern and maybe be able to predict the h-index of the scientists that will win the award in the coming years!

Therefore, a first test we shall make is to check if the time series presents any kind of autocorrelation. That is, to test if we can reject the following null hypothesis:

H_0 : *Observations follow a white noise* $\rightarrow h_{Nobel} = \mu + Z_i$ *with* Z_i *iid random variables and* $\mu \in \mathbb{R}$
H_1 : *There is a temporal correlation of any of the orders in* $(1, 2, 3, 4, 5)$

Before anything, we acknowledge that each year more than one person can share the award. In order to have a proper time series, we will consider that the laureates of each year belong to a same random variable and thus, we'll use their average h-index as the observed value for that year. You can see the time series resulting from averaging each year in the datasets of (γ) in Figure 3, for the Physics and Chemistry laureates. The autocorrelation coefficients for both groups were also calculated and can be seen graphically represented in Figure 4. None of the coefficients seem to be significantly big at first glance, except the third order one in the chemistry case.

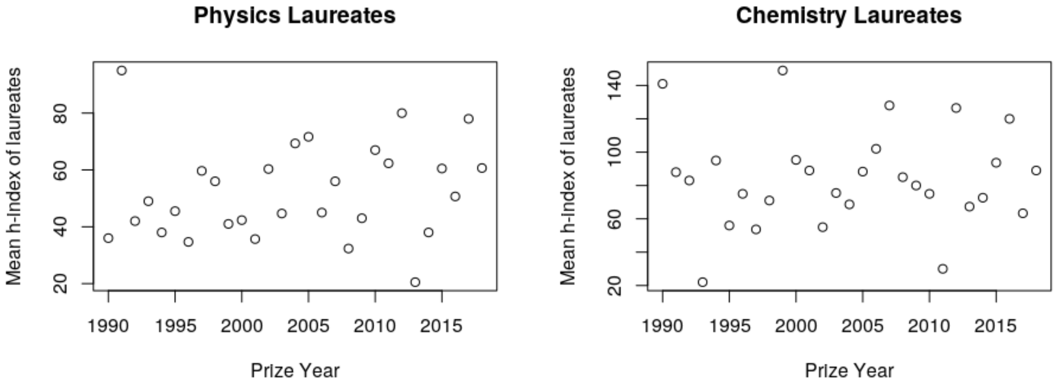

Figure 3: The mean h-index of the laureates per year.

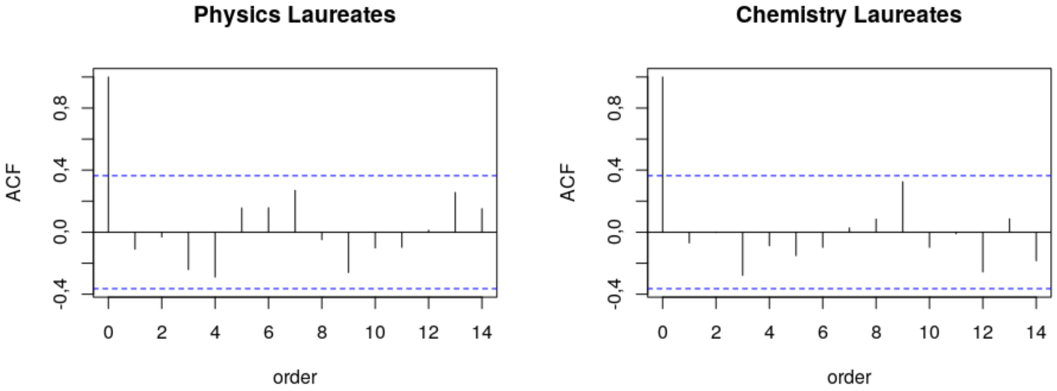

Figure 4: The autocorrelation coefficients (ACF) of different orders for the mean h-indices of different year laureates.

Performing a permutation test with 100.000 rearrangements of the time series for Physics and for Chemistry, we obtain the p-values for each alternative hypothesis considered, listed in Table 1. Clearly, non of them is smaller than a significance 0.05[2], which means that we have no sufficient evidence as to assume that the h-indices follow any clear temporal trend.

Then, we first conclude that we cannot predict next year's laureates' expected h-index. But most importantly, we conclude that h-indices for Nobel Prizes are not a year dependent phenomenon, nor for the Chemistry group, nor the Physics group. Thus, we can now safely mix the data of different year laureates!

[2] We'll consider a confidence of 95% throughout all the work.

Table 1: The Permutation Test p-values obtained with 100.000 rearrangements for different autocorrelation order alternative hypothesis. The mean h-index of each year's laureates were studied (Section 3.2)

Autocorrelation Order	**Permutation Test p-value**	
	Physics Laureates	Chemistry Laureates
1	0.33924	0.42825
2	0.50888	0.57253
3	0.10974	0.07156
4	0.06006	0.36967
5	0.13116	0.22954

3.3 Are h-indices across disciplines comparable?

We now know that we don't need to consider the temporality of laureates, but we should check the statement we assured in the introduction that h-indices across sciences should not be directly compared. If they could be compared, apart form time, we could mix the laureate h-index observations across disciplines for the remaining analysis (which would be helpful in having more observations for the tests).

We first note that the mean h-index among the 65 chemistry laureates of γ is 82.12 (st.dev. 30.46), while for the 72 physicists is 51.25 (st. dev. 21.51). This suggests that if they are really two different populations, then the expectancy for the chemists' random variable would be bigger. As such, denoting by h_{phys} and h_{chem} the random variables generating the h-indices of the respective fields, we generate the following pair of contrast hypothesis:

$$H_0 : h_{phys},\ h_{cehm}\ \textit{represent the same random variable}$$
$$H_1 : E(h_{chem}) > E(h_{phys})$$

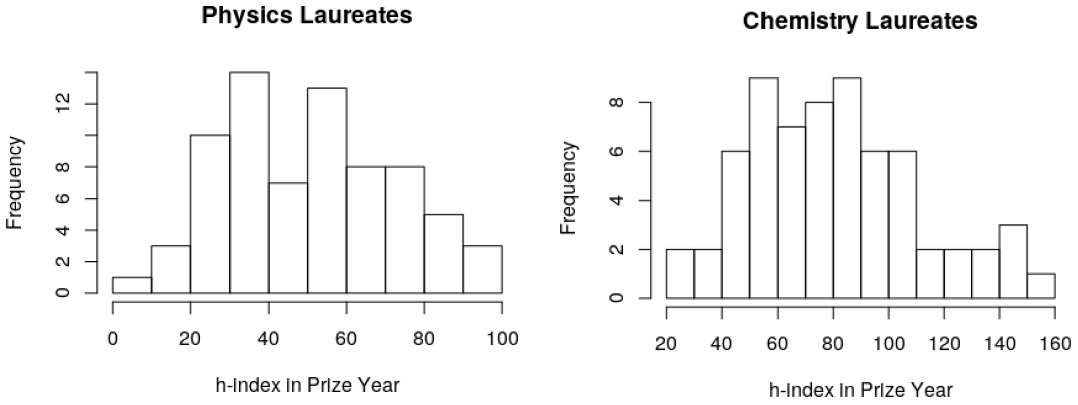

Figure 5: h-index observation histograms per group. See how visually a normal pattern can be glimpsed.

Looking at the distribution of the observed h-index values in each discipline (Figure 5), a normal distribution appears to be shaped. To assert the normality we performed a Shapiro-Wilk normality test (which was found by Monte Carlo search to be more significative than a typical Kolmogorov Smirnov Test [7]). We indeed find a p-value of 0.4 for the physics laureate data and 0.1452 for chemistry. This test is done on a null hypothesis that the data follow a normal distribution, as such, we find that we cannot reject it, and thus we can safely assume that the data proceed from normal distributions. In order to compare the means of the two apparently different populations, this allows us to perform a Parametric Bootstrap Test if we arrive to estimate the parameters of the normal distributions.

As we need to have explicit expressions for the probability distributions, we'll hypothesize that the mean and the variance of the normal distributions are the sample mean and variance:

$$H_0' : \begin{cases} h_{phys} \sim N(mean(h_{observed\ phys}), var(h_{obs_phys})) \\ h_{chem} \sim N(mean(h_{obs_chem}), var(h_{obs_chem})) \end{cases}$$

A Kolmogorov-Smirnov test will now help us assert this assumption, as it assumes for the null hypothesis that we know the mean and the variance of the normal distribution (which was not necessary for the Shapiro-Wilk test). Performing the test we get a p-value of 0.6274 for physics and 0.899 for chemistry. Then we have no reason against using H_0' as true, which will be crucial for a parametric bootstrap to test H_0 vs H_1.

Performing 100.000 simulations, we get with a 95% confidence that $E(h_{phys}) \in (46.30, 56.20)$ and $E(h_{chem}) \in (74.69, 89.53)$. There is no overlap in the confidence intervals with a 0.05 significance. Therefore, we conclude with this significance that the expected h-index for chemistry laureates is indeed higher than for physics ones. This would allow us reject H_0, and thus, we have asserted that we cannot mix the data. They are clearly two distinct groups, and thus we proved that cross field h-index comparison is meaningless.

3.4 Is there a pattern in the time laureates wait until their discovery is recognized?

In general, researchers need to wait years till their discoveries are recognized. In particular we have the numerical data of the time laureates had to wait from the publication of their discovery to the honor of the Nobel Prize. In this section, we'll be interested on looking for a pattern in this magnitude as a function of time, just as we did in section 3.2 with the h-index. That is, our starting point will be the pair of hypothesis:

H_0 : *Observations follow a white noise:* $\Delta t_{tillRecog} = \mu + Z_i$ *with* Z_i *iid random variables and* $\mu \in \mathbb{R}$

H_1 : *There is a temporal correlation of any of the orders in* $(1, 2, 3, 4, 5)$

Once again, we group the laureates of each year in a same discipline and obtain a mean value for their waited times. Thus, we are left after the averaging of data in dataset (γ) with 28 observations. These can be seen in Figure 6. The autocorrelation of different orders can be seen in figure 7. We see that only the magnitude of the autocorrelation of second order in the physicist group seems to be significant visually.

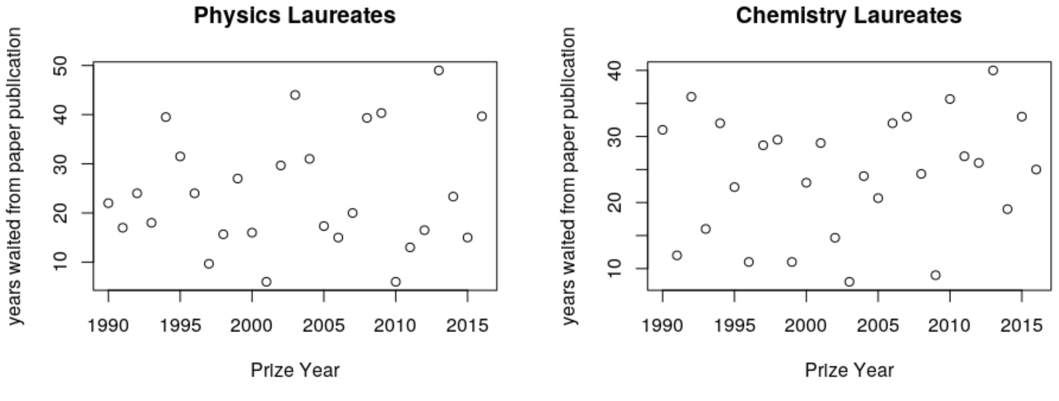

Figure 6: The mean number of years waited from the paper publication until they receive the Nobel Prize, as a function of the award year.

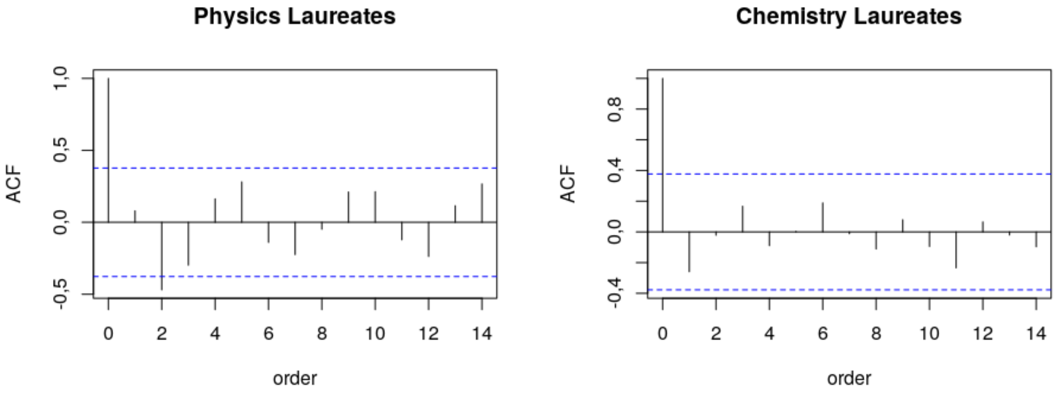

Figure 7: The autocorrelation coefficients (ACF) of different orders for the mean number of years waited until the award.

We perform a permutation test with 100.000 rearrangements of the time series for each of the groups, and obtain the p-values listed in Table 2 per autocorrelation order.

Table 2: The Permutation Test p-values obtained with 100.000 rearrangements for different autocorrelation order alternative hypothesis. We are testing the year difference from the publication of the awarded paper and the proper award as a function of the award year (Section 3.4).

Autocorrelation Order	**Permutation Test p-value**	
	Physics Laureates	Chemistry Laureates
1	0.26508	0.1152
2	0.00515	0.53439
3	0.0672	0.13277
4	0.13335	0.37741
5	0.03444	0.41646

With the values in Table 2, we see that only the p-value for the autocorrelation of 5-th order and the second order for physicists (0.03444 and 0.00515) might allow us reject H_0. However, the permutation test was approximate, so we still need to calculate the confidence interval for the p-value with the classical expression:

$$p - value = \hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p}$ is the estimated p-value, n the number of simulations (100.000) and $z_{1-\alpha/2} = 1.96$ if $\alpha = 0.05$. We get that the p-values are now 0.0344 ± 0.0009 and 0.0052 ± 0.0005. They both lie entirely in the rejection zone (with significance 0.05), and even if we could discuss the validity of the p-value for the fifth order correlation, it appears clear that there is an autocorrelation of second order.

The Nobel Laureate selection follows a similar process in both sciences. The Nobel Comitte for Physics and for Chemistry respectively, sends confidential correspondence to academics who are competent to nominate laureates. There is a list of conditions an academic needs to fulfill in order to be a valid nominator (which might seem a bit restrictive at first, but in the end apparently worldwide researchers tend to be chosen). See the page https://www.nobelprize.org/nomination/ for the details.[3]

[3] Apparently nominator and nominee information is confidential until 50 years after the selection. In this page you can find the archive of revealed nominator-nominees, the fascinating intricacies of which would deserve a whole analysis work.

Attending that some nominators are always chosen to be earlier Nobel laureates, it could be argued that laureates tend to nominate scientists of their own generation, and thus they tend to give merit to discoveries made when themselves where publishing their own discoveries. However, the autocorrelation has negative sign! Which means that if the year difference from the publication of the discovery until the Nobel prize is bigger than the average in one year, two years after the difference will be smaller than the average and so on. Thus, for this hypothesis to be trtue we should check if there is a pattern in the selection of nominators themselves. That is, an earlier nobel is not selected to be a nominator until a period, 2-3 years, pass. We cannot assure this, as nominators are confidential until 50 years later. However, this suggests that the time trend follows a rather sinusoidal trend. Additionally, this would actually be interpretable as follows: in an attempt to award both younger and older scientists, that is, scientists who published their main discovery earlier on time or further in time, the post-nomination selection committee, unintentionally ends up sequentially giving merit first to former discoveries and then each time to earlier discoveries and so on, with a period of 2 or 3 years.

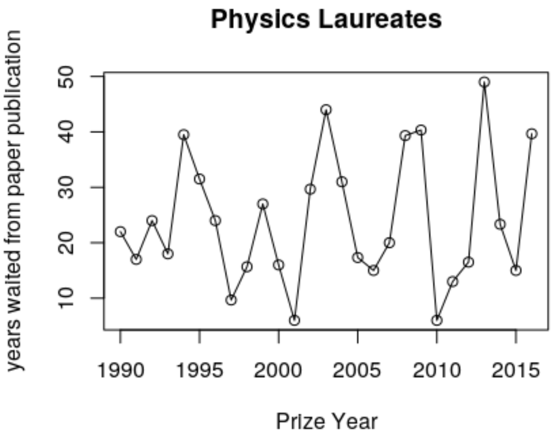

Figure 8: How many years prior to the award the discovery of the laureates was discovered as a function of time. We plotted the polygonal line joining the observations in order to evoke the sinusoidal non linear model we suggest.

In addition, there is a positive 5-th order autocorrelation, which means that there is a positive incremental pattern each 5 year step. This would coincide with one peak of the sinusoid we hypothesized (of period 2-3) with the peak two periods further. Looking this way into Figure 8, where we joined by lines the observed points, everything seems to make sense. It seems like we could model the evolution using a sinusoid with fixed period of around 2-3 years, the amplitude of which is modulated by another sinusoid the offset of which seems to increase in time. Anyhow, this non-linear analysis stays outside the scope of this work.

In short, in the case of chemistry, there is no apparent pattern in the time evolution of the waited times till recognition, while there seems to be a non-linear pattern in the case of physics. However, as the autocorrelation is negative, the values in physics do not follow a white noise, but they do seem to be around a central value. Therefore, justified for chemists, and maybe more empirically guided for physicists, in order to make a symmetric analysis for both groups in what follows, we will forget about the temporal dependence of the laureates.

3.5 Predicting the boost given by the award

It appears clear that receiving a Nobel Prize gives visibility to a scientist and in general to the particular studied field. In order to assert this, a more exhaustive study should be done by comparing the evolution of the h-indices of laureates with those of standard scientists. Anyhow, with the available data in dataset (δ), restricted to the Nobel laureate population, we could try to build a model to predict the boost in h-index the award supposes (if such a thing is indeed true). We will build the model as a function of the h-index they had when they published the awarded discovery paper $h_{whenPaper}$, the h-index the year they received the award $h_{whenPrize}$ and the years they had to wait till this recognition $\Delta t_{tillRecognition}$. In particular we'll check the impact the three of those parameters have on the h-index of the scientist 10 years after they receive the award $h_{10years\ later}$.

That is, we want to adjust a least square multidimensional linear fit as:

$$h_{10years\ later} = \beta_0 + \beta_1 h_{whenPaper} + \beta_2 h_{whenPrize} + \beta_3 \Delta t_{tillRecognition}$$

We first perform a classical regression model on the Physics and Chemistry datasets and we obtain the parameters on Table 3. Certainly, all of them except $h_{whenPaper}$, would be considered significant using a classical test. The $h_{whenPaper}$ in the physics group seems to be significative as well with 0.05 significance, but in the chemistry group it fails to be so.

Table 3: Coefficients of the linear regression model (section 3.5) fitted by least squares. The p-values of the contrast hypothesis $H_0 : \beta_i = 0$ using the classical linear model assumptions are also given.

Variable	β_i **regression coefficients**		**Classical test p-values**	
	Physics	Chemistry	Physics	Chemistry
Intercept	23,4679	12,90468	-	-
h-index paper year	-0,3757	-0,09346	0,027765	0,2530
h-index prize year	1,2139	1,09139	6,88e-11	4,33e-16
time waited	-0,6203	-0,38710	0,000761	0,0295

We will assert the significance of the parameters using a Non Parametric Bootstrap to obtain the confidence intervals for the coefficients. In particular, as we have a standard error estimator for them (the classical regression model slope standard error), we will be able to use the most convenient of the three methods studied in the lectures: the Bootstrap-t method.

Using 100.000 simulations, we find the results in Table 4. We see that zero is in the confidence interval of the coefficient for $h_{whenPaper}$ in both groups. Therefore, we clearly conclude that it is not an influential variable in the determination of the h-index 10 years after the award.

Table 4: Coefficients β_i of the linear regression model and their confidence intervals obtained using Non-Parametric Bootstrap with 100.000 simulations. Significance 0.05. (Section 3.5)

Variable	**Confidence Intervals for** β_i		**Bootstrap Estimates of** β_i	
	Physics	Chemistry	Physics	Chemistry
Intercept	(3.4567, 38.6956)	(-3.3740, 27.9258)	21.08 ± 17.62	12.28 ± 15.65
h-index paper year	(-0.7360, 0.10545)	(-0,2347, 0.0548)	-0.31 ± 0.42	-0.09 ± 0.15
h-index prize year	(0.9623, 1.4631)	(0.9510, 1.2290)	1.21 ± 0.25	1.09 ± 0.14
time waited	(-0.9790, -0.1633)	(-0.6869, -0.0218)	-0.57± 0.40	0.36± 0.33

For the other two variables we note the following: We get that the influence of the index the year they receive the award is positively influencing the one they'll have 10 years later. Its interpretation is straight-forward. What is more interesting is to note that the sign of β_3 is negative. This suggests that the more recently the paper was released when they received the award, the more attention of the scientific community working on that field they receive, and thus their papers get more citations. This could also be explained apart form the visibility, as follows: the award supposes a motivation for the researchers in order to work deeper into their fields, which in the end allows them to produce more scientific content. An alternative explanation could also be that the more recent the award winning paper was published, if it was published when the researcher was young, then they still have a longer scientific life in front. This or that, we conclude the construction of the predictive model.

3.6 Testing the Predictive Model

In order to assert the robustness of the designed model, we could obtain a non-parametric Boostrap estimate of the h-index for 2021 and 2022 of some Nobel Laureates awarded in 2011 and 2012 respectively. The point is that we can only compare these values with the ones recorded in 2020. Alternatively, we could extrapolate in the other direction of time, and predict the h-index of the laureates of 1999 and 1998. In this case, we would be able to contrast the prediction with a true value (recorded in 2009 and 2008 respectively).

We randomly choose some laureates in both directions of time for the physicst group and the chemist group independently. Their names are listed in Table 5.

We'll make the predictions using a Non Parametric Bootstrap-t method, using the fact that we know an equation for the standard error of a multidimensional linear regression prediction (see Section 5). Using 100.000 simulations for each honoured, we obtain the confidence intervals for the predictions listed in the same table. The closest possible true h-indices of those researchers to the predicted time can be found as well in that table.

Table 5: Observed parameters of the test laureates and the prediction of their h-index 10 years after receiving the award, contrasted with a true observation in the year indicated in parenthesis. The predictions and the interval confidences were obtained by Non-Parametric Bootstrap with 100.000 simulations.

Field	**Laureate**	**Award**	**Paper**	**h-index in**	**h-index predict.**	**10 year later**	**Observed h-index**	**Guess?**
	Name	**year**	**Publs. in**	**Prize year**	**CI**	**Range**	**(at year)**	
Physics	Riess, Adam G.	2011	1998	61	(72.91, 86.35)	79.63 ±6.73	74 (2020)	Y
	Schmidt, BP	2011	1998	59	(70.97, 84.32)	77.64 ± 6.68	64 (2020)	N
	Perlmutter, S	2011	1998	67	(78.73, 92.68)	85.71 ± 6.97	71 (2020)	N
	Veltman, M	1999	1972	29	(34.50, 45.80)	40.15 ± 5.65	35 (2009)	Y
	Thooft, G	1999	1972	53	(60.45, 68.88)	64.67 ± 4.21	57 (2009)	N
	Stormer, HL	1998	1982	57	(68.11, 80.18)	74.14 ± 6.03	78 (2008)	Y
Chemistry	Shechtman, D	2011	1984	30	(29.25, 39.95)	34.62 ± 5.33	30 (2020)	Y
	Lefkowit, R	2012	1986	151	(152.91, 172.24)	162.69 ± 9.07	153 (2020)	Y
	Llevitt, M	2013	1975	52	(51.14, 57.78)	54.47 ± 3.32	53 (2020)	Y
	Warshel, A	2013	1972	64	(63.05, 69.95)	66.50 ± 3.45	66 (2020)	Y
	Zewail, A	1999	1988	92	(99.43, 109.34)	104.39 ± 4.95	108 (2009)	Y
	Kohn, W	1998	1964	66	(68.26, 72.77)	70.52 ± 2.26	70 (2008)	Y

We find that the model performs truly well in the case of chemists. It guesses the value for 6 out of 6 scientists! This proves the goodness of fit.

When it comes to the physicists, only half the indices are correctly guessed and the failed ones are not that close to the prediction as to say that in the one or two remaining years they will arrive to those indices. The model doesn't fit that well. This suggests that more data should be considered in the model construction for physicists, or rather the assumption we made in approximating the waited time as time independent was wrong.

4 Conclusions

In a nutshell, we first concluded that the h-index was enough in order to assert the quality of a scientific career (it somewhat included the main information on the total number of citations). This legitimated the use of the h-index alone for the study of the Nobel laureates.

When it comes to the time series of the laureate h-indices, we found that there was no apparent temporal trend among chemistry awards nor in physics awards. This allowed us to mix the laureate observations and forget about the year they received the price.

We then confirmed the statement that h-index of two scientists of different fields is not comparable. In particular, we tested that chemistry and physics laureates belong to two different groups when it comes to the index.

Afterwards, we checked whether the time each laureate had to wait till they were awarded followed any time pattern. We found that the collected data was compatible with a white noise in time for Chemistry Laureates. Therefore, this variable as well was found to be time independent for them. In the case of physicists, we found there was a second and fifth order autocorrelation and concluded that a sinusoidal non-linear model would be best suited for a predictive fit. However, the autocorrelation was negative, which allowed us to approximate the data as a white noise around a mean value. This was a fundamental assumption for an equal footing treatment with respect to the chemistry data in the final analysis.

Having asserted this final ingredient, we built a linear model to predict the h-index of a laureate 10 years after the reception of the award. We did it separately for chemists and physicists following the previous conclusions. Also, the previous conclusions let us drop time dependences both in the h-indices of the laureates and the times they waited till recognition, which made the model more general in a sense.

Finally in order to check the prediction capacity of the model, we evaluated the values for some laureates from 2011, 2012, 1999 and 1998, as the model was built using data from 2000 to 2010. We predicted their h-index 10 years later, and checked their true h-index in a close date after a decade of their award. The closeness of the predictions for the chemistry group suggested the model was very fine. Meanwhile, the discrepancy in the physicist group made us see that more data points are required for a model within this group or rather time should also be considered in the analysis (the assumption of time independence for the waited times was wrong). This concluded the analysis.

5 Employed Statistical Data Analysis Methods

5.1 Classical Multidimensional Linear Regression Prediction Standard Error

For the computation of the Standard Error for the predicitons done using the linear mulitvariate model $\hat{Y} = \hat{\beta}X = \hat{\beta}_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_k x_k$ in Section 3.6, the following formula was used, following the proves in Ref.[9].:

$$SE^2_{Prediciton} = Var(E(Y_0) - \hat{Y_0}) = \sigma^2 \left[X_0^t (X^t X)^{-1} X_0 \right]$$

where $X_0 = (1, x_{02}, ..., x_{0k})$ is the support point we want to predict, $\hat{Y_0}$ is the prediction using: $\hat{Y_0} = \hat{\beta}X_0$ and:

$$\sigma = \left[\frac{\sum_{i=1}^{n} (Y_i - \sum_{j=1}^{k} \hat{\beta}_j X_{ij})^2}{n-k} \right]^{1/2}$$

with Y_i the i-th observation used to build the model (which was observed when the support point was $(X_{i2}, .., Xik)$ and $X_{i1} = 0 \; \forall i \in \{1..k\}$.

5.2 Normality Tests

5.2.1 Shapiro-Wilk Normality Test

As it is explained in [8], the Shapiro-Wilk test, tests the null hypothesis that a given set of observations proceed form a normal distribution. The test statistic usually called W, can be found in the cited paper. It depends fundamentally on the sample values. The interesting part is that there is no name for the particular distribution of W, and the thus the cutoff values for the test statistics are calculated using Monte-Carlo simulations (see the Wikipedia article on the test).

The point on using this test instead of a classical Kolmogorov-Smirnov (KS) goodness of fit test is that it was proven to be the most powerful method among the Anderson-Darling, KS and Lilliefors tests for normality assertion [7]. Also, it doesn't assume that the mean and variance of the normal distribution are known to build the null hypothesis, which was our case here.

5.2.2 Kolmogorov-Smirnov Test

Once the normality was asserted, we used a KS test in order to check whether there was any evidence in the sample against assuming that the mean and variance of the originating normal distribution could be acceptably estimated by the sample mean and variance.

The test basically works as follows: the empirical distribution function F_n is calculated using the n iid observations $\{x_i\}$ as:

$$F_n(x) = \frac{1}{n} \sum U(x - x_i)$$

where $U(x - x_i)$ is the Heaviside step function centered at x_i. Then, considering the cumulative distribution function $F(x)$ of the distribution in the null hypothesis (the particular normal distribution in a normality test), the test statistic is:

$$TS = supremum_x |F_n(x) - F(x)|$$

which is then contrasted with the Kolmogorov distribution critical values.

5.3 Permutation Test

This is the method used in Sections 3.1, 3.2 and 3.4 to obtain the p-values for the considered hypothesis contrasts. The method simply follows the next steps:

1. Consider a statistic that takes big absolute values if the alternative hypothesis is true (which must be logically incompatible with the null). In Sections 3.2 and 3.4 we used the autocorrelation coefficients, as they should be big if the observations in time do not follow a white noise. In Section 3.1 we chose the Spearman and Pearson correlation coefficients as they should be as furthest as possible in (-1,1) form 0 if the data are monotonically (and linearly in the case of Pearson coefficient) correlated. We assert the positive correlation checking additionally the sign of the coefficients being positive.

2. Rearrange the observations as to build new sets of observations in a way that are not rearrangeable if the null hypothesis is false. In Sections 3.2 and 3.4, we shaffled the observations in time, as this would only be possible if the observations were time independent (which satisfied H_0 but was incompatible with H_1). In Section 3.1 we shafled the h-indices of different researchers, as this would only be possible if the h-index was not bounded to the total number of citations (which is what H_0 stated in that case).

3. Then, under the null hypothesis the observations will be rearrangable in this way. Which means that under the null hypothesis, we could as well have obtained any of the possible permutations of the observed set. If this was true, we could simulate as many rearrangements as possible in order to mimic the whole support of the observations. Now, taking the statistic that measures the closeness to the null hypothesis (or furtherness), we will compute it for every possible rearrangement, aka for every possible point in the support. Now, plotting with them a histogram, we would obtain an approximation of the probability distribution of the resultant statistics we could have obtained (if the null hyp. was true).

4. Now, we look at the original sample's statistic. If it lies in the outermost zone of this probability distribution, then this would mean that if the null hypothesis was true, then we have obtained a very very improbable observation. Then, if it lies within the most unlikely α percent of possible observations, we will assume that it was too unlikely to happen as for the H_0 to be true. That is, by Ockham's razor the null hypothesis must be false, we consider that our observation was not a statistical fluctuation. In particular, the outermost area the true observation accumulates in this distribution will be the p-value we'll be referring to.

5. If you don't only want to reject the null hyp. you can consider not a two tailed outermost zone, but only check one of the sides. This is what we did in Sections 3.1, as we were interested in also testing that the correlation was positive. Also in Sections 3.2 and 3.4, we were only interested in asserting the sign of the autocorrelation. Thus, for each order we made a one tail rejection zone as a function of the sign of the autocorrelation coefficient.

5.4 Parametric Bootstrap

This is the method we used in Section 3.3 to compare the h-indices across groups. The basic idea follows as:

1. Assuming there is enough evidence as to assume that the data comes from a particular probability distribution and if we can safely estimate the parameters of the distribution: estimate them.

2. Simulate randomly as many new samples as possible following this estimated probability distribution, maintaining the sample set size as the original one.

3. Calculate the statistic of interest for each simulation; you will have built the approximate probability distribution for the statistic if the sample really came form the hypothesized distribution. Using the quantiles you can get the confidence interval.

5.5 Non-Parametric Bootstrap-t

In this case, we won't assume we know the probability distribution of the observations. Instead, we will estimate this distribution by assuming that the dataset itself is a representative discretization of the probability distribution.

1. Simulate as many re-samplings form the dataset as possible with reposition (that is, assume it is the discretization of the random variable). For each of them calculate the statistic of interest (we'll call it the bootstrap estimate $\hat{\theta}_{boots}$).

2. Assuming you know an expression for the standard error of the statistic, calculate it for each bootstrap sample $\hat{se}_{boots}$. Then compute the following parameter in each of the simulations:
$$t_b = \frac{\hat{\theta}_{boots} - \hat{\theta}}{\hat{se}_{boots}}$$
where $\hat{\theta}$ is the estimation obtained using the true observed set ($\hat{se}$ will be its standard error).

3. Defining $t_{\alpha/2}$ and $t_{1-\alpha/2}$ as the α outermost quantiles of the accumulated bootstrap t_b parameters, the confidence interval for the desired statistic will be:
$$(\hat{\theta} + t_{\alpha/2}\hat{se},\ \hat{\theta} + t_{1-\alpha/2}\hat{se})$$
We used this method in Section 3.5 and 3.6 to estimate confidence intervals for the coefficients of the linear model and the predictions of ordinate for new support points, respectively.

References

[1] Hirsch JE (2005) *An index to quantify an individual's scientific research output.* Proceedings of the National Academy of Sciences 102:16569-16572, doi: 10.1073/pnas.0507655102

[2] *"Publish or perish".* Nature. 467 (7313): 252. 2010. Bibcode:2010Natur.467..252.. doi:10.1038/467252a

[3] Fanelli, D. (2010). Scalas, Enrico (ed.). *"Do Pressures to Publish Increase Scientists' Bias? An Empirical Support from US States Data".* PLOS ONE. 5 (4): e10271. doi:10.1371/journal.pone.0010271.

[4] Ioannidis, J. P. A., Baas, J., Klavans, R., Boyack, K. W. (2019). *A standardized citation metrics author database annotated for scientific field.* PLoS Biology, 17, e3000384.

[5] Li, J., Yin, Y., Fortunato, S. et al. *A dataset of publication records for Nobel laureates.* Sci Data 6, 33 (2019). https://doi.org/10.1038/s41597-019-0033-6

[6] (2008) Spearman Rank Correlation Coefficient. In: *The Concise Encyclopedia of Statistics.* Springer, New York, NY by Dodge, Yadolah.

[7] Razali, Nornadiah; Wah, Yap Bee (2011). *"Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors and Anderson–Darling tests".* Journal of Statistical Modeling and Analytics. 2 (1): 21–33. Retrieved 30 March 2017.

[8] Shapiro, S. S.; Wilk, M. B. (1965). *"An analysis of variance test for normality (complete samples)".* Biometrika. 52 (3–4): 591–611. doi:10.1093/biomet/52.3-4.591. JSTOR 2333709. MR 0205384. p. 593

[9] Joseph S. Desalvo (1971). *Standard Error of Forecast in Multiple Regression: Proof of a Useful Result.* The American Statistician Vol. 25, No. 4 (Oct., 1971), pp. 32-34 (3 pages)

Appendix: Used R-Scripts

For Importing the datasets

```
#IMPORT DATA using readxl------------------------------

library(readxl)
# The worlds best 100.000 scientist dataset conatining h-index, citation number etc at
#        the end of 2018
best_scientists<-read_excel("Best_100000_scientists.xlsx", sheet="Career")

# For the Physics Laureates, their h-index the year the prize was given to them (from
#        1990 to 2018)
hindexPhysTime<-read_excel("Nobel_hIndex_Physics - Years.xlsx")
# same for Chemistry Laureates
hindexChemTime<-read_excel("Nobel_hIndex_Chemistry - Years.xlsx")

# For Physics Laureates awarded from 2000 to 2010, the h-index they had when they
#       published the award winning paper,
#  their h-inbex the year they received the Nobel, and their h-index 10 years after
hindex3Phys<-read_excel("Nobel_hIndex_Physics.xlsx")
# same for Chemistry Laureates
hindex3Chem<-read_excel("Nobel_hIndex_Chemistry.xlsx")

keys<-read_excel("Best_100000_scientists.xlsx", sheet="Key")

# the extra data to check the model
extra<-read_excel("ExtraPrediction.xlsx")
```

For Section 3.1

```
# PREPARE DATA ARRAYS AND PLOT THEM------------------

h<-best_scientists$‘h18 (ns)‘
c<-best_scientists$‘nc9618 (ns)‘

plot(c,h, xlab=’Total number of citations’, ylab=’h-index’) #PLOTIEU

#EXECUTE PERMUTATION TEST FOR PEARSON AND SPEARMAN CORRELATION COEFFICIENTS--------------

n<-length(h)

rearrangement_number<-10000

piersonRear<-numeric(rearrangement_number)
spearmanRear<-numeric(rearrangement_number)

originalPierson<-cor(c, h, method="pearson") # 0.844244
originalSpearman<-cor(c, h, method="spearman") # 0.9076

piersonCount<-0
spearmanCount<-0

for(i in 1:rearrangement_number){
  rear<-sample(h, n)
  piersonRear[i]<-cor(c, rear, method="pearson")
  spearmanRear[i]<-cor(c, rear, method="spearman")
  if(piersonRear[i]>=originalPierson){ piersonCount=piersonCount+1}
  if(spearmanRear[i]>=originalSpearman){ spearmanCount=spearmanCount+1}
}

hist(spearmanRear, xlab="Spearman Correlation Coefficient")
spearmanCount/rearrangement_number
#Non of the rearrangements gives a superior R -> p-value=0

hist(piersonRear,  xlab="Pearson Correlation Coefficient")
piersonCount/rearrangement_number # p-value=0

# a linear model
h_c_model<-lm(h~c)
summary(h_c_model)
```

For Section 3.2

Example code given for the Physics group. The code for Chemistry is the same but substituting every “Phys” string by “Chem”.

```
#PREPARE DATA ARRAYS, PLOT THEM AND CALCULATE THE AUTOCORRELATION----------------------

hYearPhysTable<-data.frame(hindexPhysTime)
hYearPhys<-aggregate(hYearPhysTable[,3], list(hindexPhysTime$‘Prize Year‘), mean)
plot(hYearPhys, xlab="Prize Year", ylab="Mean h-index of laureates", main=
       \ "Physics Laureates") #plotieeu -> OUTLIER???
acf(hYearPhys[2], main="Physics Laureates", xlab="order")

# PERMUTATION TEST FOR AUTOCORRELATION----------------------

n<-length(hYearPhys$Group.1)
rearrangement_number<-100000
acTrue<-acf(hYearPhys[2])
ac1<-numeric(rearrangement_number)
ac2<-numeric(rearrangement_number)
ac3<-numeric(rearrangement_number)
ac4<-numeric(rearrangement_number)
ac5<-numeric(rearrangement_number)
ac6<-numeric(rearrangement_number)
for(i in 1:rearrangement_number){
  rear<-sample(hYearPhys$x, n)
  acfRear<-acf(rear, plot=FALSE)
  ac1[i]<-acfRear$acf[2]
  ac2[i]<-acfRear$acf[3]
  ac3[i]<-acfRear$acf[4]
  ac4[i]<-acfRear$acf[5]
  ac5[i]<-acfRear$acf[6]
  ac6[i]<-acfRear$acf[7]
}
sum(ac1<=acTrue$acf[2])/rearrangement_number # 0,33924
sum(ac2<=acTrue$acf[3])/rearrangement_number # 0,50888
sum(ac3<=acTrue$acf[4])/rearrangement_number # 0,10974
sum(ac4<=acTrue$acf[5])/rearrangement_number # 0,0631
sum(ac5>=acTrue$acf[6])/rearrangement_number # 0,13116
sum(ac6>=acTrue$acf[7])/rearrangement_number # 0.12633

\subsection*{For Section 3.3}
```

```
\addcontentsline{toc}{subsection}{For Section 3.3}
\vspace{-0.3cm}
Example code given for the Physics group. The code for Chemistry is the same but substituting ever
\begin{Verbatim}[fontsize=\small, xleftmargin=0.5cm]
#VISUALIZE THE h-INDEX HISTOGRAMS--------------------

hist(hYearPhysTable$h.index.Prize.Year, xlab="h-index in Prize Year", main="Physics Laureates")

#SHAPIRO-WILK TEST-------------------------------
shapiro.test(hYearPhysTable$h.index.Prize.Year)
#pvalue 0.4

#KOLMOGOROV-SMIRNOV TEST---------------------------

n_phys<-length(hYearPhysTable$h.index.Prize.Year)
mu_phys<-mean(hYearPhysTable$h.index.Prize.Year)
sd_phys<-sqrt(var(hYearPhysTable$h.index.Prize.Year))
ks.test(hYearPhysTable$h.index.Prize.Year, "pnorm", mean=mu_phys, sd=sd_phys)
# p-value 0.6274 -> cant reject the hypothesis!

#MEAN ESTIMATION BY PARAMETRIC BOOTSTRAP---------------

simulation_number<-100000

bootstrapMeansPhys<-numeric(simulation_number)

for(i in 1:simulation_number){
  physSimul<-rnorm(n_phys, mu_phys, sd_phys)
  bootstrapMeansPhys[i]<-mean(physSimul)
}
# we get 95% bootstrap confidence intervals for both and see if they overlap
quantile(bootstrapMeansPhys, probs=c(0.025, 0.975))
#    2,5%    97,5%
# 46,29834 56,19707
```

For Section 3.4

Example code given for the Physics group. The code for Chemistry is the same but substituting every “Phys” string by “Chem”.

```
#PREPARE DATA ARRAYS, PLOT THEM AND CALCULATE THE AUTOCORRELATION------------------

timeTillRecognitionPhys<-aggregate(hYearPhysTable[,4][hindexPhysTime$‘Paper Year‘!=0],
        \ list(hindexPhysTime$‘Prize Year‘[hindexPhysTime$‘Paper Year‘!=0]), mean)
timeTillRecognitionPhys[2]<-timeTillRecognitionPhys[1]-timeTillRecognitionPhys[2]
plot(timeTillRecognitionPhys, xlab="Prize Year", ylab="years waited from paper
        \ publication", main="Physics Laureates")
acf(timeTillRecognitionPhys[2], main="Physics Laureates", xlab="order")

# PERMUTATION TEST FOR AUTOCORRELATION---------------------

n<-length(timeTillRecognitionPhys$Group.1)
rearrangement_number<-100000
acTrue<-acf(timeTillRecognitionPhys[2])
ac1<-numeric(rearrangement_number)
ac2<-numeric(rearrangement_number)
ac3<-numeric(rearrangement_number)
```

```
ac4<-numeric(rearrangement_number)
ac5<-numeric(rearrangement_number)
for(i in 1:rearrangement_number){
  rear<-sample(timeTillRecognitionPhys$x, n)
  acfRear<-acf(rear, plot=FALSE)
  ac1[i]<-acfRear$acf[2]
  ac2[i]<-acfRear$acf[3]
  ac3[i]<-acfRear$acf[4]
  ac4[i]<-acfRear$acf[5]
  ac5[i]<-acfRear$acf[6]
}
sum(ac1>=acTrue$acf[2])/rearrangement_number # 0,26508
sum(ac2<=acTrue$acf[3])/rearrangement_number # 0,00515
sum(ac3<=acTrue$acf[4])/rearrangement_number # 0,0672
sum(ac4>=acTrue$acf[5])/rearrangement_number # 0,13335
sum(ac5>=acTrue$acf[6])/rearrangement_number # 0,03444
```

For Section 3.5

Example code given for the Physics group. The code for Chemistry is the same but substituting every “Phys” string by “Chem”.

```
# ORIGINAL MODEL EVALUATION--------------------------------

timeTillRecognitionPhys<-hindex3Phys$‘Prize Year‘-hindex3Phys$‘Paper Year‘
originalModelPhys<-summary(lm(hindex3Phys$‘h-index 10 years later‘ ~ hindex3Phys$‘h
        \ -index Paper Year‘+hindex3Phys$‘h-index Prize Year‘+timeTillRecognitionPhys))
originalModelPhys

# NON PARAMETRIC BOOTSTRAP FOR REGRESSION COEFFICIENTS--------------

b_inter<-originalModelPhys$coefficients[1]
b_paper<-originalModelPhys$coefficients[2]
b_award<-originalModelPhys$coefficients[3]
b_recogt<-originalModelPhys$coefficients[4]

sd_inter<-originalModelPhys$coefficients[5]
sd_paper<-originalModelPhys$coefficients[6]
sd_award<-originalModelPhys$coefficients[7]
sd_recogt<-originalModelPhys$coefficients[8]
n<-length(hindex3Phys$‘Laurete Name‘)

simulation_number<-100000
indxs<-seq(1, n)
t_inter<-numeric(simulation_number)
t_paper<-numeric(simulation_number)
t_award<-numeric(simulation_number)
t_recogt<-numeric(simulation_number)

for(i in 1:simulation_number){
  rearIndex<-sample(indxs, n, replace=TRUE)
  bootsModel<-summary(lm(hindex3Phys$‘h-index 10 years later‘[rearIndex]
          \ ~ hindex3Phys$‘h-index Paper Year‘[rearIndex]
           \ +hindex3Phys$‘h-index Prize Year‘[rearIndex]
            \ +timeTillRecognitionPhys[rearIndex]))

  t_inter[i]<-(bootsModel$coefficients[1]-b_inter)/bootsModel$coefficients[5]
  t_paper[i]<-(bootsModel$coefficients[2]-b_paper)/bootsModel$coefficients[6]
  t_award[i]<-(bootsModel$coefficients[3]-b_award)/bootsModel$coefficients[7]
  t_recogt[i]<-(bootsModel$coefficients[4]-b_recogt)/bootsModel$coefficients[8]
}
#CI for the beta coefficients
b_inter+sd_inter*quantile(t_inter, probs = c(0.025, 0.975))
# 2,5%     97,5%
# 3,456727 38,695626

b_paper+sd_paper*quantile(t_paper, probs = c(0.025, 0.975))
#      2,5%      97,5%
# -0,7359812  0,1054549

b_award+sd_award*quantile(t_award, probs = c(0.025, 0.975))
#     2,5%     97,5%
# 0,9622919 1,4631034

b_recogt+sd_recogt*quantile(t_recogt, probs = c(0.025, 0.975))
#      2,5%      97,5%
#-0,9789771 -0,1633193
```

For Section 3.6

Example code given for the Physics group. The code for Chemistry is the same but substituting every “Phys” string by “Chem”.

```
# NON PARAMETRIC BOOTSTRAP ESTIMATION OF THE PREDICTION BY THE MODEL
FOR EACH OF THE PHYS LAUREATES IN THE DATASET CALLED EXTRA--------------------

timeWaited<-(extra$‘Prize Year‘-extra$‘Paper Year‘)[extra$Field=="Phys"]
hindexPrize<-(extra$‘h-index Prize year‘)[extra$Field=="Phys"]

num_boots<-100000
lowerBounds<-numeric(length(timeWaited))
upperBounds<-numeric(length(timeWaited))

# we build the new model without the h index at papers year
originalModelPhys<-summary(lm(hindex3Phys$‘h-index 10 years later‘
        \ ~ hindex3Phys$‘h-index Prize Year‘+timeTillRecognitionPhys))
X<-model.matrix(originalModelPhys)
for(i in 1:length(timeWaited)){
  originalPrediction<-originalModelPhys$coefficients[1]
          \ +originalModelPhys$coefficients[2]*hindexPrize[i]
           \ +originalModelPhys$coefficients[3]*timeWaited[i]
  sigma<-sqrt(sum((hindex3Phys$‘h-index 10 years later‘-originalModelPhys$coefficients[1]
          \ +originalModelPhys$coefficients[2]*X[,2]
                  \ +originalModelPhys$coefficients[3]*X[,3])^2)/(n-3))
  new_x<-c(1, hindexPrize[i], timeWaited[i])
  se_origPred<-sigma*sqrt(t(new_x)%*%solve(t(X)%*%X, tol=1e-12)%*%new_x )
  t_boots<-numeric(num_boots)
  for(k in 1:num_boots){
    rearIndex<-sample(indxs, n, replace=TRUE)
    bootsModel<-summary(lm(hindex3Phys$‘h-index 10 years later‘[rearIndex]
                    \ ~ hindex3Phys$‘h-index Prize Year‘[rearIndex]
                            \ +timeTillRecognitionPhys[rearIndex]))
    bX<-model.matrix(bootsModel)
    bootPrediction<-bootsModel$coefficients[1]+bootsModel$coefficients[2]*hindexPrize[i]
```

```
                  \ +bootsModel$coefficients[3]*timeWaited[i]
    bsigma<-sqrt(sum((hindex3Phys$‘h-index 10 years later‘-bootsModel$coefficients[1]
            \ +bootsModel$coefficients[2]*bX[,2]
                    \ +bootsModel$coefficients[3]*bX[,3])^2)/(n-3))
    se_bootPred<-bsigma*sqrt(t(new_x)%*%solve(t(bX)%*%bX, tol=1e-12)%*%new_x )
    t_boots[k]<-(bootPrediction-originalPrediction)/se_bootPred
  }
  lowerBounds[i]<-originalPrediction+quantile(t_boots, 0.025)*se_origPred
  upperBounds[i]<-originalPrediction+quantile(t_boots, 0.975)*se_origPred
}

#WE OBTAIN THE LOWER AND UPPER BOUNDS OF THE CONFIDENCE INTERVALS IN THE ARRAYS
#                 lowerBounds and upperBounds
```