10.25097/rep.n25.2017.02

Artículo Original

Bases for alternative nonparametric Mincer function

Bases para una alternativa noparametrica de la funcion de Mincer

Alejandro Brondino[1]

Matías Nicolás Sacoto Molina[2]

[1] Independent econometric modeling Consultant. Email address: ale_brondino@hotmail.com. Cordoba, Argentina.
[2] Professor at University of Cuenca. Email address: matias.sacoto@ucuenca.edu.ec. Cuenca, Ecuador.

Resumen

Este trabajo realiza una regresión no paramétrica con el fin de probar la viabilidad de esta técnica para modelar una versión simplificada de la función de ganancias de Mincer aplicada a los salarios de los jugadores de la NBA. Las principales ventajas del uso de esta técnica es que no se basa en supuestos y la inferencia estadística no es sensible a perturbaciones de distribuciones debido a violaciones de estos supuestos. Los resultados de la estimación no paramétrica se comparan con una regresión OLS clásica. Se encontró evidencia de que la regresión OLS no cumplió con los supuestos que este método requiere, por lo tanto, inferencia estadística en base a esta regresión podría llevar a establecer conclusiones incorrectas (debido a la ineficiencia del estimador), a menos que se apliquen las correcciones al modelo que permitan solucionar los problemas con los supuestos. Por otro lado, los intervalos de confianza obtenidos de la regresión no paramétrica son más precisos y menos sensibles a la variabilidad y magnitud de las variables. En consecuencia, la estimación no paramétrica sería una alternativa para modelar el comportamiento de los salarios evitando supuestos muy estrictos que potencialmente conducirán a conclusiones de inferencia estadística erróneas.

Palabras clave: econometría no paramétrica, inferencia estadística, estimación no paramétrica, función Mincer, intervalos de confianza.

Abstract

This work undertakes a nonparametric regression in order to assess the viability of this technique in modeling a simplified Mincer Function of earnings applied to the NBA players'

wages. The main advantages of using this technique is that it does not rely on assumptions and the statistical inference is not sensitive to distributions disturbances due to violations of the assumptions. The results of the nonparametric estimation are compared to a classical OLS regression. We found evidence that the OLS estimator did not fulfilled the assumptions that this method requires, therefore, the statistical inference form this estimation could lead to wrong conclusions (due to lack of efficiency), unless some correction that solves the violation to the assumptions is applied to the model. On the other hand, the confidence intervals obtained from the nonparametric regression are more accurate and less sensitive to variability and magnitude of the variables. Consequently, the nonparametric estimation would be an alternative to model the behaviour of the wages avoiding strong assumptions that could lead to wrong statistical inference conclusions.
Key words: nonparametric econometrics, statistical inference, nonparametric estimation, Mincer function, confidence intervals.

Código JEL: C52, C14.

Recibido: 04/07/2017
Aceptado:17/10/2017

1. Introduction

Several studies have been conducted in order to describe potential factors that might explain the behavior of the wages in an economy. Mainly, the Mincer earnings function[1], Jacob Mincer (1974), has been applied to different samples, even to various countries and industries. Moreover, in the sports industry it would be interesting to develop a model that enable us to describe how experience interacts with wages of sportsmen, specifically, the NBA players' wages.

However, frequently the models used to investigate this equation stand on many assumptions, in some cases really strong (e.g. exogeneity, homoscedasticity, etc.) and violations to these assumptions can affect, to some extent, the conclusions derived from these models. Nonetheless, it is important to mention that there is plenty of bibliography related to techniques

[1] Due to Jacob Mincer (1974): $ln\omega_i = ln\omega_0 + \rho s_i + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, where: ω_i= earnings (wage), s_i=years of schooling, x_i =years of potential labor market experience, ϵ_i = mean zero residual and ρ, β_1, β_2 are coefficients.

that allow researchers to overcome many estimations problems that might arise from the violation of these assumptions[2].

The main reasons why a model estimated using a technique of estimation that relies on assumptions can lead to wrong conclusions (due to not fulfillment of these) can be: a) heteroscedasticity, which underestimates the variance of the coefficients; b) not normal distribution of the error terms, which also affects the variances of the estimates; c) autocorrelation; d) endogeneity; among others. Thus, statistical inference will present skewness. Specifically, the confidence intervals of the fitted values of the dependent variable won't be precise.

Fixing the specification issues that an OLS estimation might present can be burdensome, therefore, as an alternative, a nonparametric estimation, e.g. using a *k* nearest weighted neighbor, is proposed to get more reliable confidence intervals without the need for further corrections to the original estimation method.

In particular, in the present work we are interested in showing how a nonparametric estimation represents appropriately the shape of the relation of the logarithm of the wages of a sample of players of the NBA and their years of experience. Furthermore, a confidence interval is estimated from the nonparametric estimation in order to pursue reliable statistical inference. Additionally, the obtained results are compared to the classical OLS estimation, in which we also included the years of experience squared in order to have no constant relation in the model. As a result, it can be seen that all assumptions of the OLS estimation are violated, so that the confidence interval presents problems.

2. Data and descriptive statistics.

The econometric modelling of wages is usually based on the assumption that a person's pay is correlated to their personal skill. Nevertheless, since direct measures of the level of skill are hard to find, most models tend to approximate it by the level of education, IQ or (like in the present paper) the level of work experience. Sports are a field of study that allow for specific measures of work performance and empirical studies lead to the result that better performing athletes tend to earn more money (Rose, S., Sanderson, A., 2000). Taking into consideration these reasons, we decided to perform the model comparison for a dataset that contains 267 observations[3] of the NBA professional players. Specifically, it has information of annual salary

[2] The reader can refer to Gujarati, D. N. (2009) or Cameron, A., Trivedi, P. (2009) for further information about techniques used to correct problems that the OLS model might present.

[3] The data comes from: http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html.

and years of experience at the time that the information was gathered. It is important to note that in this particular dataset the years of experience are measured as a discrete variable. Table 1 summarizes the principal statistics of the sample.

Table 1. Descriptive statistics of data.

	min. Value	max. Value	mean	median	mode	standard Deviation
Salary	1.500.000	57.400.000	14.189.000	11.860.000	1.500.000	9.879.219
Experience	1	13	5,0262	4	2	3

Source: The Authors.

The range of the data is really large, for salaries is 55900000 and for years of experience is 12. Furthermore, based on the variance coefficient, the annual salary has a variation with respect to the mean of 69,63% and the wages have a dispersion of 64,42%.

Additionally, from table 2 we could affirm that, since there not many observations of the players with 12 and 13 years of experience, the estimates for this part of the dataset might have a large variance. Moreover, the low number of observations for higher level of experience suggests that the extreme fitted values may be underestimated with the nonparametric algorithm.

Readers might feel that the data is not ideal, however, it is important to remark that the purpose of this work is to compare efficiency of estimates rather than finding specific economical results.

Table 2. Frequency distribution of years of experience.

Years of experience	Number of observations
1	36
2	41
3	28
4	32
5	26
6	24
7	15

8	18
9	16
10	10
11	12
12	4
13	5
Total	**267**

Source: The authors.

3. Methodology.

3.1. OLS estimation.

The first model that is studied is a simplification of the Mincer function:

$$ln\omega_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + u_i \qquad (1)$$

Where $ln\omega_i$ represents the logarithm of the wage of player *i* and x_i the years of experience and u_i represents the error term of observation *i*. The model is estimated with classical OLS.

The reason why we reduced the original equation is because we preferred to keep this work more parsimonious to facilitate the analysis and comprehension of the estimation method. Nevertheless, upcoming studies will have deeper analysis in which other variables (like years of schooling) are included to the equation in order to test the robustness of the results.

About the model, it has been broadly discussed whether the Mincer function is too simplistic. Even though the quadratic variable enables the model to have a variation that depends on the magnitude of the independent variable, for example, Lemieux (2003), shows that higher order polynomials enhance the capacity of prediction of the model.

Several tests are also applied in order to prove whether the estimated equation fulfill the assumptions on the classical OLS estimation. Specifically, we test the normality of the residuals with the Jarque-Bera test, the autocorrelation of the residuals using a Durbin-Watson test, and we use the White test to analyze heteroscedasticity in the model.

Finally, an asymptotic confidence interval is computed. The aim of this, is to compare this interval to the resultant from the nonparametric regression. A MonteCarlo simulation was undertaken for this purpose[4].

[4] In specific, we applied bootstrapping to dataset in order to compute the empirical confidence intervals. For details in bootstrapping refer to Cameron and Trivedi (2009).

The simulation consists on repeating the estimation process thousands of replications. For each iteration, a new dataset is generated from the original sample, so that the characteristics of the original data are kept. Afterwards, in each replication the fitted dependent variable is computed. Finally, the 0,025 sorted fitted value is taken as the asymptotic lower limit interval and the 0,975 is considered the upper limit interval. These values are considered due to a significance level of 5%.

3.2. Nonparametric estimation.

In the second part of this work we present an estimation of $m(x_i)^5$ using a nonparametric regression model:

$$y_i = m(x_i) + \varepsilon_i \qquad (2)$$

We apply the k-nearest neighbors regression technique, which takes averages in neighborhoods "$N_k(x)$" of a point *x*. We selected this technique because, even though it is easy to implement, it exhibits remarkable flexibility while modeling low dimensional data (Altman, N. S., 1992). The neighbors are defined in such a way as to contain a fixed number *k* of observations (which means that we are not necessarily using the same bandwidth for each of the bins).

We find the *k* observations with x_j values closest to x_i, and average their outcomes. Basically, the idea is that if $m(x_i)$ is relatively smooth, it does not change too much as *x* varies in a small neighborhood. Afterwards, taking an average over values close to *x*, $\hat{m}(x_i)$ should give an accurate approximation.

$$\hat{m}(x_i) = \sum_{j=1}^{k} \frac{1}{k} y_j \qquad (3)$$

Where the y_j are the realization of the *k* observations in which $|x_i - x_j|$ is smaller. In the case in which there are several y_j with the same $|x_i - x_j|$ and adding those to the previously selected observations would lead to a number higher than k we apply the following algorithm:

- Divide the observations that satisfy the conditions mentioned above in two vectors, based on the sign of $x_i - x_j$.
- Randomize the order of the elements of each of the vectors, to avoid following a pattern that might generate additional BIAS in the estimations.
- Combine those two vectors into a new one, created by intercalating the first element from each vector as long as it is possible, and then the remaining elements of the vector with the higher number of observations (if necessary).

- Select the remaining number of observations for $\hat{m}(x_i)$ from the first set of elements of this vector.

This algorithm guarantees that we are not over-representing players with more (or less) experience than the bin value of x, unless it is strictly necessary due to the data set used. It is necessary to implement it because of the discrete nature of the variable x, that implies that (in case we do not have a clear aleatory criterion for data selecting in certain situations) we might be systematically selecting more players with more (or less) years of experience than the estimation point, leading to BIAS that could have been avoided.

It is important to notice that the selection of *k* can dramatically change the outcome of the model in different ways. For example, if *k = n*, we are using all the observations, and $\hat{m}(x)$ just becomes the sample average of y_i. Graphically, we will have a perfectly flat estimated function. Furthermore, a large *k* leads to a relatively low variance, nonetheless, the estimated $m(x)$ is biased for many values of *x*, thus, the estimation is inconsistent. Whereas when *k = 1* we are using one observation to estimate the value of each bin. This dramatically reduces the bias but, as we are using few observations, the variance is high.

The literature formally does not stablish a way to select an optimal value of *k*, however, one possible appropriate way would be by trying different values of *k* and picking the one that minimizes the Cross Validation estimate of the MSE (Henderson, D., Parmeter, C., 2015).

In our case we choose *k* using as a reference the choice presented in the section 9.4.2 of Cameron, A. C., & Trivedi, P. K. (2005) which is a value such that: $k = \frac{1}{4} N$.

Since the intuition behind the k nearest neighbor methodology is that objects close in distance are potentially similar, we decided to apply the distance weighting refinement. We choose the Euclidian distance metric defined as:

$$D(x_1, x_2) = \sqrt{(x_1 - x_2)^2} = |x_1 - x_2| \qquad (4)$$

In order to use this distance as weight for each of the y_j, we decided to use the exponential weighting defined as:

$$W_j = \frac{e^{-D(x_i, x_j)}}{\sum_{j=1}^{k} e^{-D(x_i, x_j)}} \qquad (5)$$

After applying the distance weighting refinement, the estimate for $m(x_i)$ is replaced for the following equation:

$$\hat{m}(x_i) = \sum_{j=1}^{k} W_j \; y_j \qquad (6)$$

Additionally, it is important to note that, since the *x* (years of experience of the player) are discrete variables, it is essential to take this into consideration while choosing the number and location of each bin in order to avoid creating skewed neighborhoods.

4. Results and Discussion.

First, model (1) was estimated with OLS. The result of this estimation is presented in table 3.

Table 3. Estimation output of model (1).

	Coefficent	**Standard Deviation**	**t-statistic**
Constant	6,02995	0,12897	46,75608
x	0,27805	0,04531	6,13635
x^2	-0,01328	0,00331	-4,01166
R squared	0,21265		
Jarque-Bera	17,20940		
D-W	2,16671		
LM	21,04284		

Source: The Authors.

We can see that all the coefficients are significant. Furthermore, the signs of the estimates for years of experience and for squared years of experience are as expected, given that the curve of this equation is concave.

However, this model presents: not normal residuals, positive residual autocorrelation and heteroscedasticity, as the tests shows. Consequently, based on classic econometric theory, we can affirm that the variance of the estimates won't be efficient. Therefore, statistical inference, and more precisely, the confidence interval is not reliable.

Indeed, from figure 1, it is easy to realize that at the extreme, the intervals start to explode away from the fitted values.

Source: The Authors.

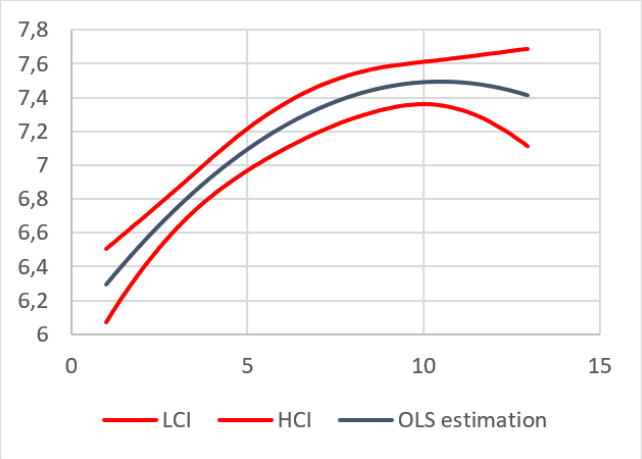

Figure 1. OLS estimation and 95% confidence interval.

Secondly, we estimated the *k* nearest neighbor regression, using distance weighting. As this model is sensitive to the value of *k*, we ran it for different values of this parameter in order to graphically asses the tradeoff between BIAS and variance that was describe in the methodology section. The comparison of these results can be seen in section 7.1. The *k* nearest neighbor regression for different values of *k*. After applying this procedure we selected *k* = 65. It is important to remark that the ratio of the selected $k\left(\frac{k}{N} = 0.24\right)$ is between the ratios of the *k*s recommended by Cameron and Trivedi (0.05 and 0.25).

As it was mentioned in the Methodology section, before estimating the model, it is necessary to take into consideration that the variable "Years of experience" follows a discrete distribution. A discrete distribution of *x* implies that an arbitrary placement of the bins will lead to additional problems in the model outcome. This is due to the fact that (since many of the observations consist in the same value of *x*) a bin placed in a certain position (e.g. $x_i = 4.49$) will have the same *k* nearest neighbors as a bin placed relatively far away (in this example the furthest bin with the same neighborhood will be close to $x_i = 4.01$). This is one of the reasons why we decided to use the distance weighting[5] refinement that partially solves the problems that may arise from this situation. The problems are mitigated because, in spite of the fact that both bins have the exact same *k* nearest neighbors, the weights for each observation will vary based on the value of x_i and therefore $\widehat{m(x_i)}$ will change as well. Nevertheless, we decided to place the bins either for $x_i = a$ or $x_i = a + 0.5$ where *a* is an integer, in order to guarantee that each bin is associated to a different neighborhood.

[5] The distance weighting diminishes boundary issues in line with the conclusions from Hechenbichler, K., Schliep, K. (2004).

Now we present the results from the k-neighbor regression for k = 65 with the corresponding confidence intervals computed through bootstrapping:

Source: The Authors.

Figure 2. K nearest estimation and 95% confidence interval

As it can be seen in figure as the relationship captured by the non-parametric technique applied leads a result that resembles a logarithmic relationship between the variables considered. It resembles the logarithmic function in the sense that it grows at a higher rate in the first years of experience and after some years it starts to grow at a decreasing rate. Nevertheless, there are two main differences when we compare it to a logarithmic function.

The first one is that at the beginning of the function the growth rate is relatively small. This is one of the characteristics of the *k* nearest neighbors algorithm because (since a lot of the nearest neighbors for $x_i = 1$ are related to $x_j > 1$ and non are related to $x_j < 1$) the first bin usually leads to overestimation of $\widehat{m(x_i)}$.

The second difference is found for $x_i = 7$, that yields an estimate $\widehat{m(x_i)}$ that is smaller than it should be for a logarithmic function. After analyzing the data set, we realized that there are not many observations for x = 7, which might lead to a significant difference between the sample distribution and the population distribution of the variable. Additionally, as it can be seen in the confidence intervals for x = 7, the lack of observations in this point also increases the amplitude of the interval.

The first difference is inevitable and inherent to the estimation technique used, but the second one could be solved by utilizing a bigger data-set. Additionally, it is important to note that

(similarly to the case of the first bin) the last bin tends to be underestimated since it is a function of observations associated to lower levels of experience.

4.1. Comparison between the models

Source: The Authors.

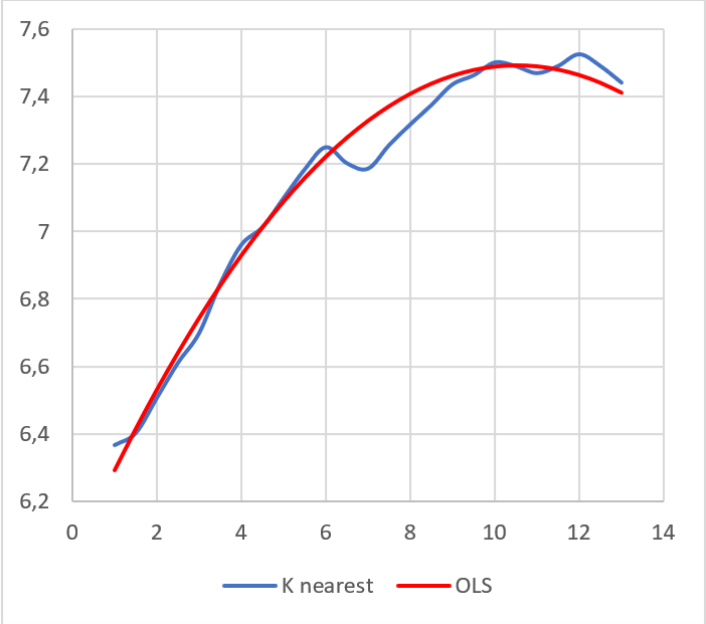

Figure 3. OLS and K-nearest estimations.

Both the estimates that comes from the OLS and the K nearest neighbors are remarkably similar. We detect three main differences:

a) The K nearest estimation is higher in the first bins. This is probably due to the fact that this algorithm tends to overestimate the first bins.

b) Close to $x = 7$ we see that the K nearest estimation suddenly lowers its value but later it returns to values similar to those of the OLS. As it was stated in the previous section, this is probably a feature of the data that could be solved if the data set could be expanded. There is no theoretical reason for this and since both functions behave similarly for the following values of x, this is probably just an issue with the dataset.

c) For the last bins, the OLS estimation is lower than the K nearest one. Since the K nearest neighbors algorithm tends to underestimate monotonically increasing functions in the last bins, could be an indication that the OLS is underestimating the function even more.

5. Conclusions

Both estimation techniques yield similar outcomes for the dataset analysed. Nevertheless, the lack of assumptions behind the K nearest neighbors algorithm makes it easier to implement, especially in a context in which the OLS violates several assumptions. Not addressing the assumption violation in the OLS can lead to underestimating the variance of the coefficients, rendering the model unable to perform trustworthy inference. Addressing these problems might be time intensive and troublesome. This paper exhibits the K nearest algorithm with the distance weighting refinement as an alternative, due to results presented in previous sections, that might provide fewer estimation issues and lead to similar results. Additionally, there is evidence that this model might perform better than the OLS for players with more than 11 years of experience. It would be interesting to conduct a study that could provide further evidence in this matter, especially if the dataset analysed contains players with more than 13 years of experience.

Bibliography.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), 175-185.

Boston College (2014 october 19th). Wooldridge data sets. Retrieved from: http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html.

Cameron, A., Trivedi, P. (2009). *Microeconometrics, Methods and Applications.* New York: Cambridge University Press.

Gujarati, D. N. (2009). Basic econometrics. Tata McGraw-Hill Education.

Hechenbichler, K., & Schliep, K. (2004). Weighted k-nearest-neighbor techniques and ordinal classification.

Henderson, D., Parmeter, C. (2015). *Applied Nonparametric Econometrics.* New York: Columbia University Press.

Lemieux, T. (2003). 11. The "Mincer Equation" Thirty Years after Schooling, Experience, and Earnings. In S, Grossbard. (Ed.), *Jacob Mincer a Pioneer of Modern Labor Economics* (pp 127-147). New York: Springer.

Mincer, Jacob (1974), *Schooling, Experience and Earnings*, New York: Columbia University Press.

Rose, S., Sanderson, A. (2000). Labor Markets in Professional Sports. *NBER* Working Paper No. 7573.

Annex.

1.1. The *k* nearest neighbor regression for different values of *k*

Source: The authors.

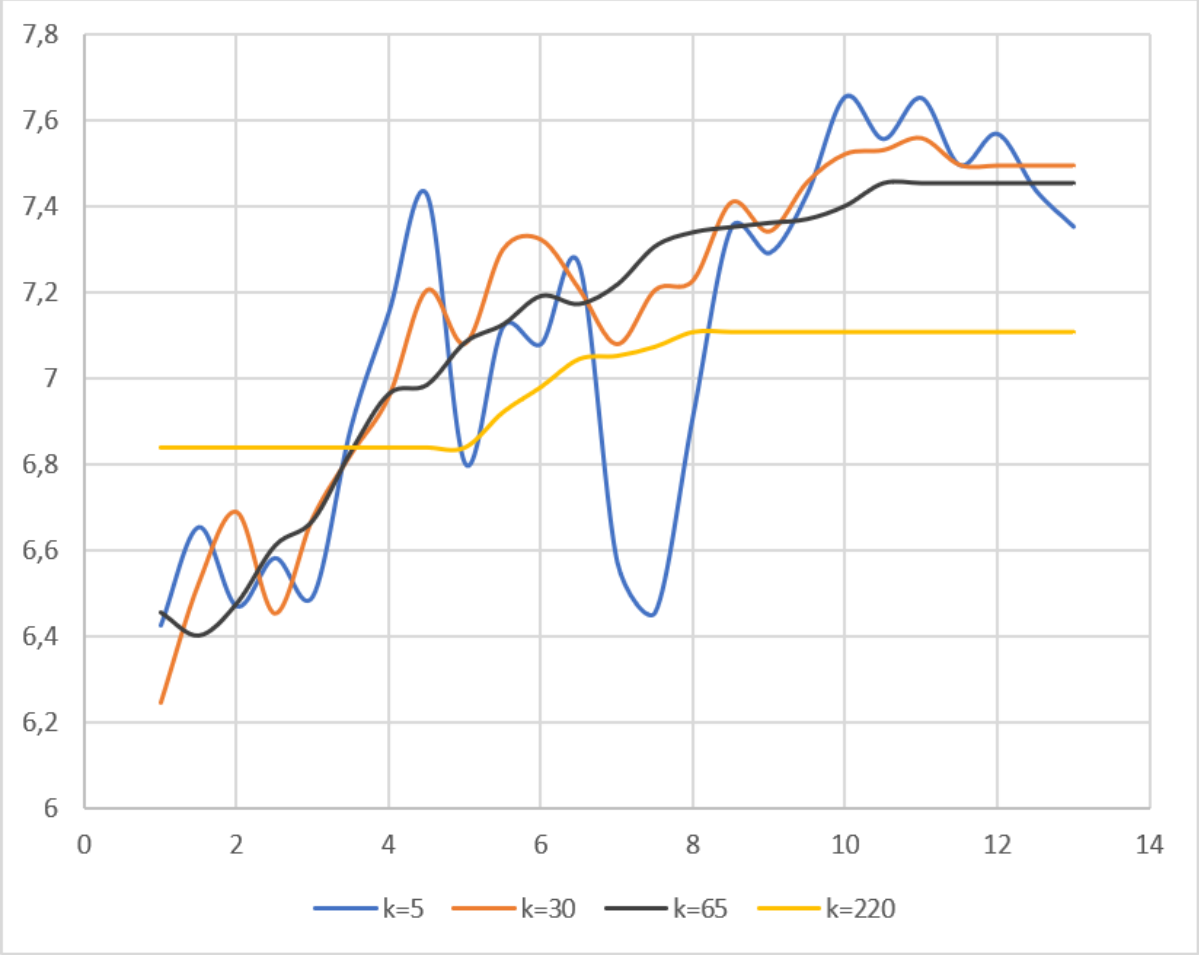

Figure A1. K nearest neighbor for different values of K (No distance weighting)

As it can be seen in figure A1 different values of *k* lead to completely different estimates of the parameter of interest. It is important to note that a higher *k* leads to oversmoothing and larger boundaries issues. This can be seen in the yellow line, corresponding to $k = 220$, in which the first eight bins have the same estimation and the same occurs for the last twelve bins. For a monotonically increasing function this leads to over-estimation in the first group of bins and under-estimation for the last group of. In the case of $k = 5$ we have a clear example of undersmoothing, which leads to a smaller bias but is associated to higher volatility. Both $k = 30$ and $k = 65$ are intermediate cases in which we see a more subtle example of the trade-off between BIAS and variance.

Applying the distance weighting refinement leads to the following results:

Source: The authors.

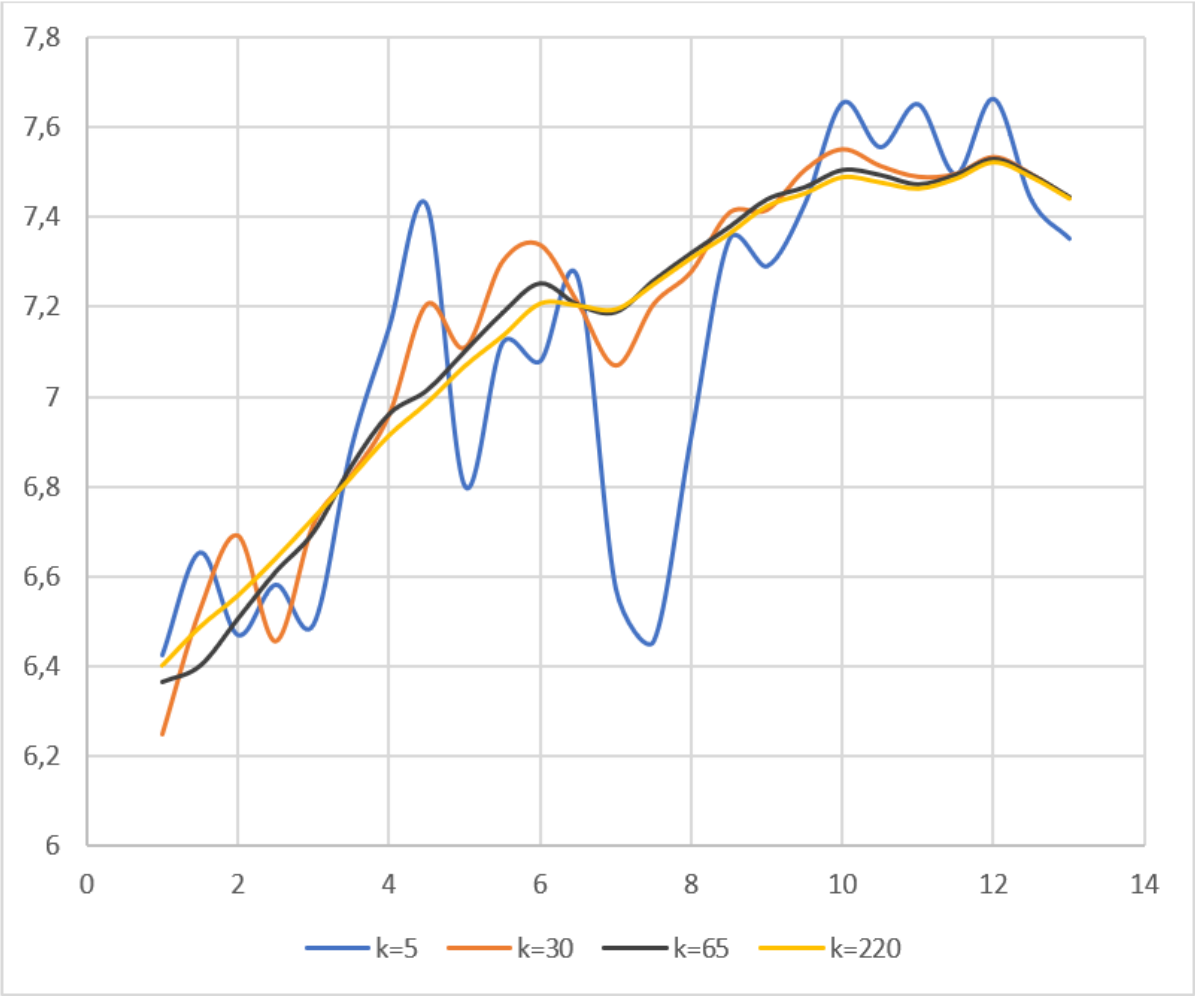

Figure A2. K nearest neighbors for different values of K (distance weighting)

As it can be seen in figure A2 there is no clear case of oversmoothing, even for k = 220. The distance weighting prevents this from happening because even if two bins are estimated using the same neighbors, the weights associated to each observation will differ based on the location of the bin. Therefore, high values of k present less boundaries issues if we apply the distance weighting refinement, in line with the conclusions from Hechenbichler, K., & Schliep, K. (2004).

We decided to choose k using as a reference the choice presented in the section 9.4.2 of Cameron, A. C., & Trivedi, P. K. (2005) which is a value such that: $k = \frac{1}{4} N$. We selected k = *65* after we found no evidence of undersmoothing or oversmoothing based on the graphic analysis.