UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Econometrics Master's thesis

UFR02 École d'économie de la Sorbonne

Master 1, Économétrie, Statistiques.

Monte Carlo methods : An application to the White test for Heteroscedasticity

Authors:
Mehdi Ferhat
Jerrold Nemba
Alexis Vignard

Supervisor:
Dr. Philippe De Peretti

Abstract

In this study, we tried to stress the White test using Monte Carlo algorithms to generate data for different combinations of correlation and sample size, expecting it to shatter at some point. Surprisingly, the test does not seem to respond to different degrees of correlation, retaining excellent robustness under either the null hypothesis of homoscedasticity or the alternative hypothesis. More precisely, while it retains an error threshold for the H_0 (Type I), at a certain sample level the test for the heteroscedastic case seems to be flawless, always rejecting H_1. Multicollinearity, on the other hand, is not a problem either, hence even completely ignoring it to satisfy its initial purpose of detecting a non-constant variance of residuals.

Contents

1 Introduction

This paper proposes to use Monte Carlo methods to verify the reliability of the White test for heteroscedasticity. The White Lagrange multiplier test statistic is the product of the R-squared value (Coefficient of determination) and sample size and is supposed to be distributed like a Chi-squared. In this respect, the interest of using a Monte-Carlo based algorithm would be to verify its robutsness by playing with the sample size and the correlation between the explanatory variables. These latter will be generated from various marginal distributions using the Iman-Conover algorithm.

The literature on this study topic is quite sparse. The articles dealing with White's test are very similar to each other and are more or less in-depth explanations of this test. Insofar as it does not question the assumptions stated by Halbert White in his work, the dissertation may be of interest because it will seek through data to question some of his hypotheses. Moreover, some of the concepts used will be considered in an "Econometric analysis". For example, a model with collinear variables will not cause any inconvenience on its forecasting quality but may generate interpretation problems in relation to the estimation of its parameters. Therefore, it will be necessary to define the concepts in order to understand their meaning and purpose. Indeed, explaining the distinction between collinearity and correlation will be emphasized as well as the purpose of Monte Carlo simulations through the use of Iman Conover's algorithm. Using these notions, we therefore propose to generate four explanatory variables under the null hypothesis of homoscedasticity and the alternative, for different sample sizes (N= 50, 100, 150, 250, 500), different levels of correlation (ρ = 0, 0.5, 0.75, 0.99) and finally for different trehsholds (p = 0.85, 0.90, 0.95, 0.99).

2 Concepts and process

2.1 White test for heteroscedasticity

Linear regressions can sometimes be confronted with the problem of heteroscedasticity, i.e. when the variance of their residuals is dependent on explanatory variables, resulting in their variance to not be constant anymore. In this case, the study of the significance of the parameters is no longer possible, the Gauss-Markov theorem is distorted ($\hat{\beta}$ is no longer BLUE) as well as the usual variance formulas. The White heteroscedasticity test, named after Halbert White and proposed in 1980, is a statistical test which aims to determine whether a linear model presents heteroscedastic residuals.

Let an arbitrary linear regression model :

$$y = X\beta + \varepsilon = a_0 + b_1 \times x_{1t} + b_2 \times x_{2t} + \dots b_p \times x_{pt} + \varepsilon_t$$

The first step is about estimating this simple model by the OLS method, and get the estimated residuals $\hat{\varepsilon}_n$. White[1] proposes to regress its square by the explanatory variables, their square and their cross products :

$$\hat{\varepsilon}_t^2 = a_0 + \sum_{p=1}^{P} b_p \times x_{pt} + \sum_{p=1}^{P} b_{P+p}(x_{pt}x_{pt}) + \sum_{p=1}^{P}\sum_{z>p}^{Z} b_{(2\times P+p)}(x_p \times x_z) + u_t \tag{1}$$

Assumptions : Under the null hypothesis H_0, the model is homoscedastic and the variance of the residuals is constant, the $\hat{\beta}_{OLS}$ is still usable. Under the alternative hypothesis H_1, the residuals are heteroscedastic : $\mathbb{V}(\varepsilon) \neq \hat{\sigma}^2$

$$\begin{cases} H_0 : b_p = 0 \ \forall p \in \{1 \dots k\} \\ H_1 : b_p \neq 0 \ \exists p \end{cases} \quad \begin{cases} TR^2 \xrightarrow[H_0]{\mathcal{L}} \chi^2(k) \\ with \ k = p + \frac{2\times(2p-1)}{2} \end{cases}$$

The White test efficiency is measured by using Type I and II errors. These errors are defined as follows :

- **Type I error** : H_0 is wrongly rejected, meaning that when an homoscedastic model is generated, White "gets it wrong", by not confirming the homoscedasticity of the model
- **Type II error** : H_1 is wrongly rejected, meaning that when an heteroscedastic model is generated, White "gets it wrong" by not confirming the heteroscedasticity of the model

[1] Tripathi, G. (2000). ECONOMETRIC METHODS: By Jack Johnston and John DiNardo, McGraw Hill, 1997. Econometric Theory, 16(1), 139-142. doi:10.1017/S0266466600001092

2.2 Correlation and Mutlicollinearity

Correlation and collinearity[2] should not be confused. Indeed, although collinear variables are by definition correlated, it is not always true the other way around. The necessary condition for the existence of collinearity is the fact that the variables measure the same phenomenon.

2.2.1 Correlation

Correlation coefficients are indicators that determine the size and the direction of the variation of two variables taken together. A high degree of variation between variables refers to the strength of the relationship that could exist between two or more variables. In other words, this reflects the fact that the correlation coefficient can be positive or negative. In the positive case, two variables vary in the same direction. In the negative case, two variables vary in the opposite direction. We can also speak of opposing variables.

There are several types of correlation coefficients:

- Pearson correlation : This index illustrates the linear relationship between two continuous variables. This coefficient (noted r) can take any values between 1 and -1 and can possibly take the value of 0.
$$\rho_{XY} = \frac{\mathbb{C}ov(X,Y)}{\sigma_X \sigma_Y}$$
- Spearman correlation : In this case it is a little bit different. When Pearson study the linear relationship, the Spearman index focus more on the existence of a monotonic function between two continuous or ordinal variables. Likewise, this coefficient varies between 1 and -1. Also called, "Spearman's rho", it is used to estimate the correlation coefficient based on the rank. Indeed, the Spearman correlation coefficient is based on the ranked value for each variable, rather than the actual value of the variable (its raw data). It is more commonly known as a non-parametric version of the Pearson coefficient.
$$\rho_s = \rho_{XY} = \frac{\mathbb{C}ov(rk_X, rk_Y)}{\sigma_{rk_X}, \sigma_{rk_Y}} \tag{2}$$

2.2.2 Multicollinearity

Collinearity[3] is basically defined by the existence of a linear relationship between two variables : In other words, two explanatory variables are "collinear" if they measure the same phenomenon in the model. These relations are called "linear combinations". Thus, there will be collinearity between X1 and X2 if there exists a relation such as :

$$X_1 = \lambda_0 + \lambda_2 X_2$$

The multicolinearity phenomon is described with the existence of a linear combination between several other explanatory variables :

$$X_1 = \lambda_0 + \lambda_2 X_2 + \lambda_3 X_3$$

[2]Econométrie : Théorie et applications, Valérie Mignon, 2008
[3]BOURBONNAIS, Régis, et al. Econométrie. Paris, France : Dunod, 2011.

One of the well-known consequences of multicollinearity is the artificial increase in the value of the coefficient of determination R^2. The R^2 is a well-known indicator for judging the rate of variance explained by a model, which is theoretically supposed to show how well the explanatory variables describe the dependent variable. However, if the model suffers from the presence of multicollinearity, this indicator should be taken with great caution.

Moreover, it should be noted that there are different stages of multicollinearity, and that "perfect" multicollinearity will make the estimation of the model impossible: $(X'X)$ matrix would then become singular (i.e if the columns of the design matrix X are linear combinations between-themselves, then X will, by definition, no longer be a "full-rank" matrix). We will therefore deal with the issue of "almost" multicollinearity, i.e. when the relation between our columns is "approximate", but not perfect (or exact).

2.3 White and correlated variables

As seen previously in the section on White's heteroscedasticity test, we can intuitively think that White is closely related to collinearity.

Firstly, considering the form of the model, the fact that it reuses the variables, their cross-products as well as their squares, the presence of collinearity seems intrinsically linked to White.

Secondly, it is possible to predict the possibility that multicollinearity may be a problem with the NR^2 test statistic. Indeed, if with the presence of multicollinearity the R^2 is supposed to grow : while converging to 1, the statistic of White will obvioulsy converge to N, and the decision rule will always be to reject H_0 (thus biasing the test).

Thereby, it is worth trying to explain why the coefficient of determination could increase in the case of a White with correlated variables. White's method consists in adding to the original explanatory variables their squares and their crossed products.The model is more likely to be more complex than necessary, and this may cause an increase of the R^2 value when the number of explanatory variables becomes larger and larger. This is the phenomenon of overtraining[4] (or overfitting), and it has the effect of drastically increasing the R^2. A model that over-learns or overfits is a model that has a greater variability explained by regression (and the squared sum of the errors will be greater): the model looks better because it is made up of more variables but the problem is that an overfitted model could not predict correctly new data.

$$R^2 = \frac{SSE}{SST} \left\{ \; lim_{SSE \to SST} \; \Leftrightarrow \; R^2 \to 1 \right.$$

These intutions may be a bit naive, it will be appropriate to verify them by using Monte Carlo simulations.

]

[4]Neural network studies. 1. Comparison of overfitting and overtraining Igor V. Tetko, David J. Livingstone, and Alexander I. Luik Journal of Chemical Information and Computer Sciences 1995 35 (5), 826-833 DOI: 10.1021/ci00027a006

The Monte Carlo methods have revolutionised the way in which economics and mathematics conceptualise data analysis. There is no single definition of the Monte Carlo principle, but more generally, it represents a way of estimating a numerical quantity while using random numbers or, alternatively, a statistical tool that allows the average of a random variable to be determined. Nicknamed "Monte Carlo" by Stanisław Ulam and John von Neumann to refers to gambling in casinos, Monte Carlo simulations played a major role in the creation of the first atomic bomb during the Second World War (the Manhattan Project). However, if a simulation of a bank loan is carried out, it cannot be considered as a Monte Carlo simulation because it is determined by known variables (the number of months and the interest rate); no random phenomena are involved in these calculations.

In our study, two Monte-Carlo based algorithms will be used, in one hand,The Iman-Conover one which allows us to generate data and control their correlation, and in the other hand an algorithm to generate several White's LM.

2.4 The Iman-Conover method

Spearman's correlation consists in showing the implication of movements of one distribution in relation to an other, not by observing their values but by observing their ranks. It is on this type of correlation that Iman Conover's algorithm[5] is based: this is one of its greatest advantages: it is not necessary to calculate densities, but simply to know the ranks.

$$Target_{Rankcorr} = \begin{pmatrix} 1 & Cov(x_1, x_2) & Cov(x_1, x_3) & Cov(x_1, x_4) \\ Cov(x_2, x_1) & 1 & Cov(x_2, x_3) & Cov(x_2, x_4) \\ Cov(x_3, x_1) & Cov(x_3, x_2) & 1 & Cov(x_3, x_4) \\ Cov(x_4, x_1) & Cov(x_4, x_2) & Cov(x_4, x_3) & 1 \end{pmatrix} \tag{3}$$

Iman Conover will attempt to reproduce a correlation given manually (2) in a target matrix. Let W = $N \times p$, a matrix made up of p univariate laws (these laws are therefore independent and uncorrelated with each other) The correlation of W is arbitrary.

$$X_N = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{41} \\ x_{12} & x_{22} & \cdots & x_{42} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1t} & x_{2t} & \cdots & x_{4t} \end{pmatrix} \begin{cases} x_{1t} \sim \mathcal{N}(\mu, \sigma^2)\,. \\ x_{2t} \sim Lognormal(\mu, \sigma^2)\,. \\ x_{3t} \sim exp(\lambda). \\ x_{4t} \sim U(a, b). \end{cases} \tag{4}$$

Iman Conover's algorithm will transform W (the matrix composed of the p univariate laws) into a new matrix X_N (3), which is multivariate. In order to do it, the algorithm reproduces Spearman by rearranging the columns until the correlation of the Target is approximately obtained (4) (not perfect but really close). It is precisely because the distributions contained in W are univariate that the algorithm works: in fact, since the laws are independent (univariate), the algorithm can move the rows without affecting the marginal probability laws.

[5]Ronald L. Iman W. J. Conover (1982) A distribution-free approach to inducing rank correlation among input variables, Communications in Statistics - Simulation and Computation, 11:3, 311-334, DOI: 10.1080/03610918208812265

$$Rankcorr_{X_N} = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{41} \\ x_{12} & x_{22} & \cdots & x_{42} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1t} & x_{2t} & \cdots & x_{4t} \end{pmatrix} \approx Target_{Rankcorr} \tag{5}$$

The random component of this algorithm relies solely on the choice of the marginal distributions used. This is the great strength of this algorithm, if we choose to set a certain correlation in the Target (1) and run the algorithm in a loop, the generated X_N (3) matrix will always be the same: this simulation method is purely deterministic.

According to Iman and Conover's original paper, 100 repetitions would be enough to get close enough to the desired correlation rank.

2.5 A Monte-Carlo based protocol to test Heteroscedasticity

Using the Iman-Conover method defined above, we generate[6] a matrix X_N of 4 variables for the same marginal laws (3) and the associated Target rank of correlation (2). Necessary parameters and residuals are generated randomly following a normal distribution (5). Let y_t a non-random vector of multivariate measurements, it is simply obtained by a general linear model of the form $X\beta + \varepsilon$ as follows :

$$y_t = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_t \end{pmatrix} = \underbrace{X_N \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} + \begin{pmatrix} e_0 \\ e_1 \\ \vdots \\ e_t \end{pmatrix}}_{X\beta+\varepsilon|\beta,\varepsilon\sim\mathcal{N}(\mu,\sigma^2)\,.} \tag{6}$$

This step finished, last part of the process is about getting the Lagrange multiplier test statistic. Using generated X_N design matrix, we apply the white transformation getting the cross product, squares (1) of original regressors columns and adding the $\bar{1}$ constant vector missing.

$$X_{N_{white}} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{11}^2 & \cdots & x_{31}x_{41} \\ 1 & x_{12} & x_{22} & \cdots & x_{12}^2 & \cdots & x_{32}x_{42} \\ \vdots & \vdots & \vdots & & \vdots & & \vdots \\ 1 & x_{1t} & x_{2t} & \cdots & x_{1t}^2 & \cdots & x_{3t}x_{4t} \end{pmatrix} \tag{7}$$

The output is a matrix of $\left[2 \times 4 + \frac{2\times(2\times4-1)}{2}\right] + 1 = 15$ columns and T rows. The next step of the process now is to regress X_N by the OLS and obtain the squared residuals.

$$\hat{\beta} = (X'_N X_N)^{-1} X'_N y \tag{8}$$

[6] Wicklin, R., SAS Institute. (2013). Simulating data with SAS. Cary, N.C: SAS Institute.

$$\hat{\varepsilon} = y - X\hat{\beta} \tag{9}$$

Subsequently the regression is performed by using $X_{N_{white}}$:

$$\hat{\beta_\varepsilon} = (X'_{N_w} X_{N_w})^{-1} X'_{N_w} \hat{\varepsilon}^2 \tag{10}$$

$$SSE = (X_{N_w}\hat{\beta_\varepsilon})' X_{N_w}\hat{\beta_\varepsilon} - T\bar{Y}^2 \tag{11}$$

$$SSR = \hat{\varepsilon}^{2\prime}\hat{\varepsilon}^2 - (X_{N_w}\hat{\beta_\varepsilon})' X_{N_w}\hat{\beta_\varepsilon} \tag{12}$$

$$SST = SSR + SSE, R_N^2 = \frac{SSE}{SST}. \tag{13}$$

Then, the TR^2 LM is finally stocked in a vector $\bar{S}$. For 1 to N iteration, our algorithm will execute the same process from (3) to (12) so that we obtain the $\bar{S}$ matrix of each score test statistic.

$$\bar{S}_{1toN} = \begin{pmatrix} TR_1^2 \\ TR_2^2 \\ TR_3^2 \\ \vdots \\ TR_N^2 \end{pmatrix} \tag{14}$$

3 The resilience of White's test to multicollinearity when the generated data is homoscedastic

3.1 Verification of the distribution of the data

In order to verify the efficency of the Iman Conover algorithm applied to White's model, we run a "test series" applied to the Homoscedastic case with uncorellated data. So, by putting a correlation of 0 in our Target matrix, we check if Iman Conover succeeds in generating data without any correlation between the explanatory variables. To begin with, we display the graphical representation of the NR^2 distribution to check if the generated data follows a Chi-squared distribution. We represent it with the Gamma law, which is a distribution family including the χ^2:

$$\chi \sim \Gamma(k = \frac{14}{2}, \sigma = 2)$$

It gives us a $\chi^2(k = 14)$ distribution, represented by the blue curve overlaying the histogram below:

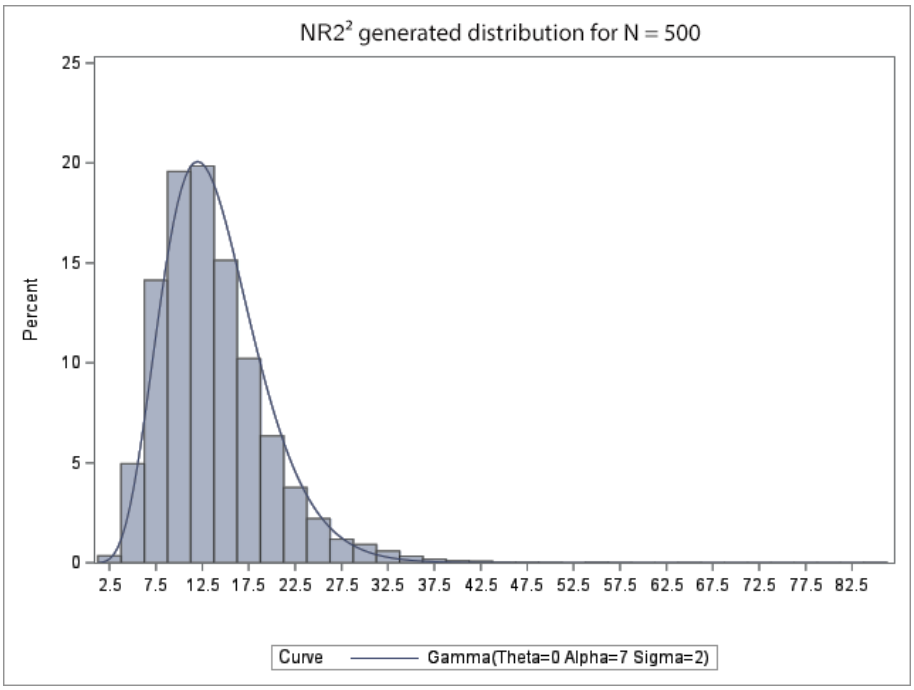

Figure 1: Comparison between the generated data NR^2 and the Chi-squared distribution law

A first intuition is that graphically, our data seems to follow a "Chi-squared" distribution. We verify this assumption by using the Kolmogorov-Smirnov[7] test. This test checks whether or not a sample follows a known law (normal, exponential, chi-square, etc.). This is a "Goodness Of Fit" test. Under the null hypothesis, the empirical cumulative distribution function (ECDF) is supposed to converge to the theoretical function. Because the test is based on the distribution functions, it allows us to test the similarity at each point between two different data (and it takes at least a million point for that to be conclusive).

The theorical distribution function is given by:

$$\begin{aligned} & F_y^T(y) = \mathbb{P}(Y \leq y) \\ & \Leftrightarrow F_y^T(Y) = \int_{-\infty}^{y} f(y)\mathrm{d}x \begin{cases} lim_{y \to -\infty}\ F_Y(y) = 0 \\ lim_{y \to +\infty}\ F_Y(y) = 1 \end{cases} \end{aligned}$$

The empirical distribution function is given by :

$$F_y^E(y) = \frac{1}{T} \sum_{t=1}^{T} I_{]-\infty,y]} Y_t \begin{cases} I_{]-\infty,y]} Y_t = 1 \\ if\ Y_t \leq y \\ else\ 0 \end{cases}$$

[7]Lecoutre, Jean-Pierre. Statistique et probabilités. Dunod, 2002.

Then, with the derivative, we get its statistics :

$$D = \max_{-\infty<y<+\infty} \left|F_y^T(y) - F_y^E(y)\right|$$

Thus, the test verifies if the data follows a known distribution law, checking if D is greater than the critical value (with the Kolmogorov-smirnov table).

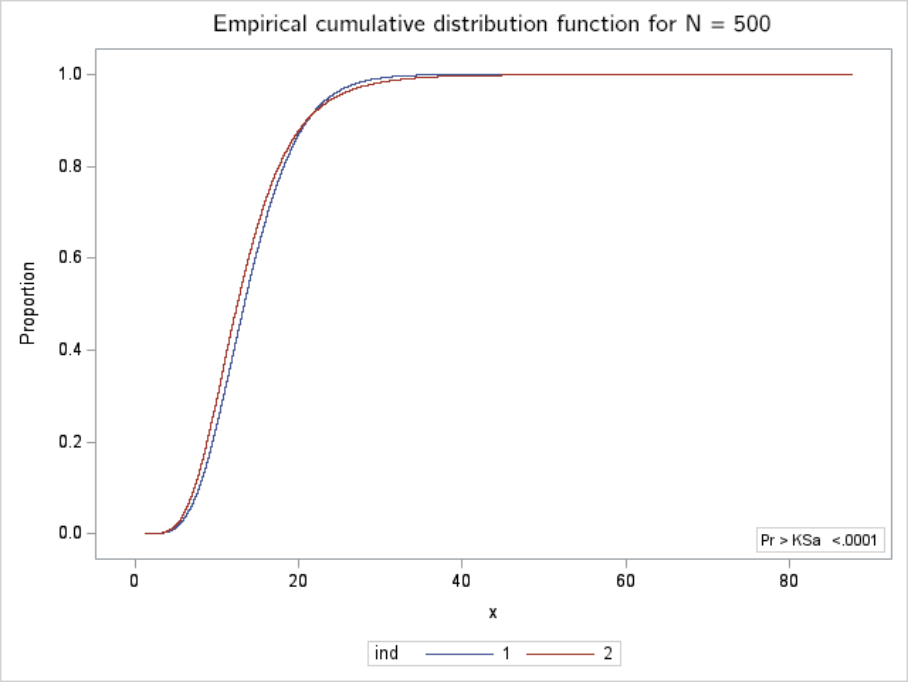

Figure 2: ECDF for $N = 500$ compared to the Chi-squared distribution law

The distributions are indeed very close, however it is not-inferentially conclusive and the test rejects the null hypothesis : they are fundamentally different.

Kolomogorov-Smirnov Test			
ind	N	EDF at maximum	Deviation from mean at maximum
1	150000	0.394413	-6.733827
2	50000	0.463960	11.663331
Total	200000	0.411800	
Maximum deviation occured at observation 173198			
Value of x at maximum = 12.009659			

Table 1: Kolmogorov-Smirnov : Inferential gap between empirical and theorical distributions

Kolmogorov-Smirnov two-sample test (Asymptotic)			
KS	0.030115	D	0.069547
KSa	14.467654	Pr >KSa	<.0001

Table 2: Statistical Kolmogorov-Smirnov test results

It seems that Kolmogorov-Smirnov rejects the null hypothesis. We tried another test, wich is the Kuiper one. This test is also used to verify a distribution by comparing it to a given known law. It was invented by the mathematician Nicolaas Kuiper.

$$Kuiper = D_+ + D_- = \max_{-\infty<y<+\infty} \left|F_y^T(y) - F_y^E(y)\right| + \max_{-\infty<y<+\infty} \left|F_y^E(y) - F_y^T(y)\right|$$

Kuiper two-sample test (Asymptotic)					
K	0.080707	Ka	15.628779	Pr >Ka	<.0001

Table 3: Statistical Kuiper test results

Kuiper indeed confirms Kolmogorov-Smirnov's results. As a matter of fact, it is not surprising knowing that to obtain real statistical significance, it would probably have been necessary to generate more data ($N = 500$ here).

Therefore, we ran the algorithm with a much larger sample ($N = 10,000$) to see if the the score test ECDF would converge into the Chi-squared distribution :

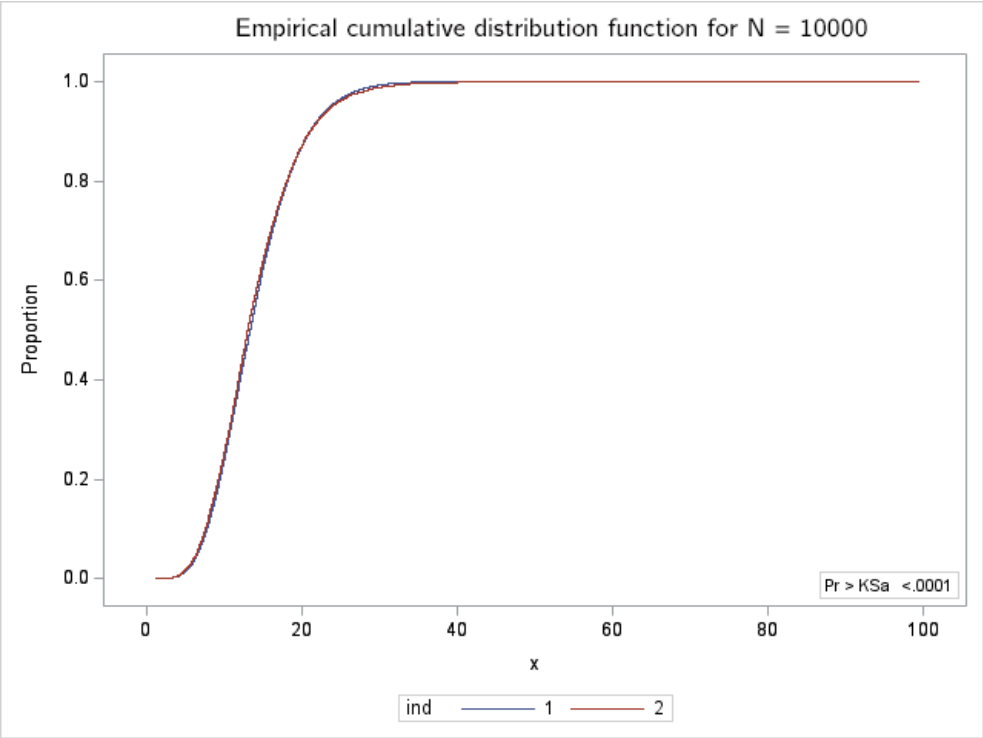

Figure 3: ECDF for $N = 10000$, compared with the Chi-squared distribution

Kolmogorov-Smirnov two-sample test (Asymptotic)			
KS	0.006340	D	0.026193
KSa	2.536159	Pr >KSa	<.0001

Table 4: Kolmogorov-Smirnov with $N = 10000$

Kuiper two-sample test (Asymptotic)					
K	0.032287	Ka	3.126143	Pr >Ka	<.0001

Table 5: Kuiper with $N = 10000$

As expected, the data is gradually converging to a Chi-squared distribution but we still are rejecting the null-hypothesis. It would certainly take an extremely large number of observations to obtain the desired distribution, but still, the results are satistying as we could really expect it to be almost the same asymptotically.

3.2 Correlation when $\rho = 0$

3.2.1 The reproduction of uncorrelated data with Iman Conover

Running Iman Conover's algorithm, it seems to work well. The correlations are indeed close to 0 according to the HeatMap and the scatter plots :

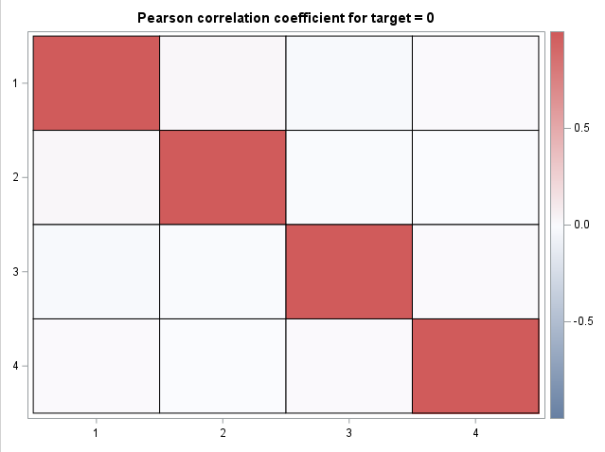

Figure 4: Heatmap with $\rho = 0$

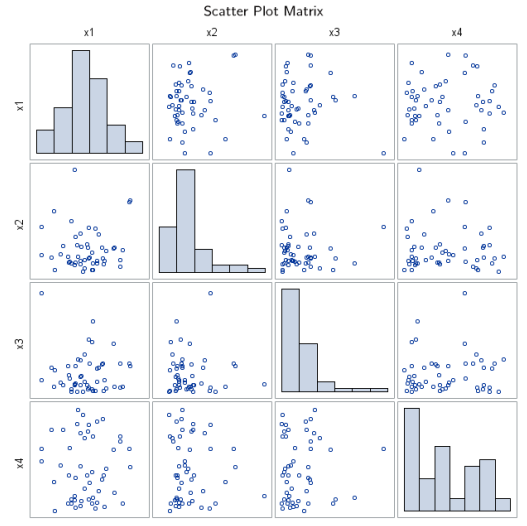

Figure 5: Scatter plot matrix with $\rho = 0$

The correlation matrix initially filled with zeros :

$$C = \mathbb{I}_4 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{15}$$

Our variables seem to have been reproduced successfully with the correlation desired. A target matrix with no correlation between the variables return us a very weakly correlated data.

Similarly, the histograms appear to have the distributions specified previously (Normal, LogNormal, Exponential and Uniform) (4).

3.2.2 White test Behavior when the correlation is close to zero

In this section, White is expected to be efficient because explanatory variables are uncorrelated.

Correlation = 0							
				Type I Error			
Observations	Mean NR^2	Median NR^2	Mean R^2	p = 0.85	p = 0.90	p = 0.95	p = 0.99
N = 50	13.34988	12.56435	0.26699	0.12354	0.0803	0.04266	0.00772
N = 100	13.27685	12.28085	0.13276	0.1292	0.0906	0.05238	0.01574
N = 150	13.31620	12.24420	0.08877	0.13152	0.09332	0.05538	0.01828
N = 250	13.39975	12.33852	0.05359	0.1337	0.09288	0.05622	0.02108
N = 500	13.51012	12.47353	0.02702	0.13658	0.09214	0.05542	0.01858

Table 6: Results with an homoscedastic model, when $\rho = 0$

As expected, White rejects the alternative hypothesis in 95% of cases for $\alpha = 0.05$, thus confirming the presence of homoscedasticity, regardless of the number of variables generated. The test's statistics are generally stable, whether it is their means or their median. However, it is noted that they increase very slowly as the number of data generated increases. This progression is not proportional, in fact the more the number of variables generated increases, the more the R^2 decreases. The more precise p is, the less frequent the Type I errors will occur.

3.3 Correlation when $\rho = 0.5$

3.3.1 Graphic representation

Previous results were expected because of the absence of multicollinearity. In this section, we introduce a correlation of 0.5 in the Iman Conover Algorithm.

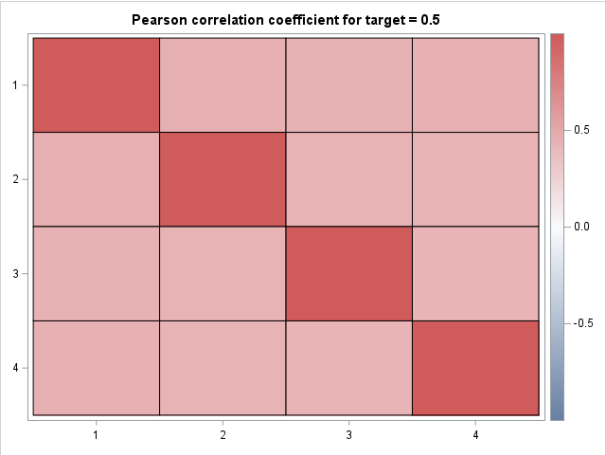

Figure 6: Heatmap when $\rho = 0$

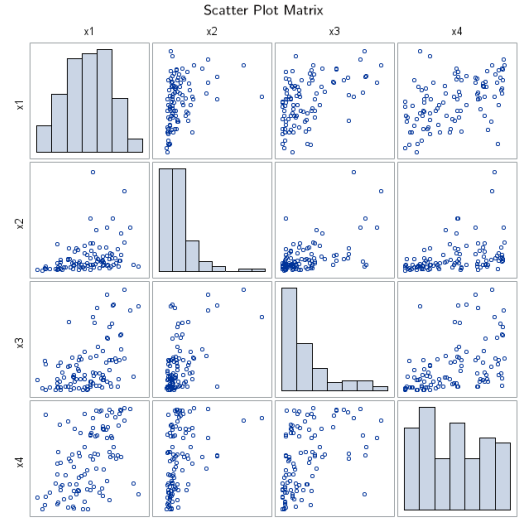

Figure 7: Scatter plot matrix when $\rho = 0.5$

The heatmap shows a correlation already more pronounced than before although far from being very high. The scatterplots shows the beginning of linear relationships between our variables.

3.3.2 White behavior when the correlation is medium

Correlation = 0.5							
				Type I Error			
Observations	Mean NR^2	Median NR^2	Mean R^2	p = 0.85	p = 0.90	p = 0.95	p = 0.99
N = 50	13.29915	12.48093	0.26598	0.1234	0.0838	0.04352	0.0082
N = 100	13.21140	12.16863	0.13211	0.1268	0.091	0.05178	0.0172
N = 150	13.19761	12.13487	0.08798	0.1322	0.0874	0.0531	0.019
N = 250	13.29609	12.23860	0.05318	0.1334	0.0988	0.05462	0.0196
N = 500	13.42614	12.40593	0.02481	0.1312	0.0934	0.5496	0.0224

Table 7: Results with an homoscedastic model, when $\rho = 0.5$

It appears that our conclusions do not differ than the uncorrelated case. Indeed, White does not seem affected by a moderate correlation, and the test statistics adopts the same behavior as before: the R^2 also decreases as the number of observations increases.

3.4 Correlation when $\rho = 0.75$

3.4.1 Graphic representation

As the previous test did not bear fruit, we start to import a severe correlation into our model in order to study its behavior in the presence of strong collinearity, or even multicollinearity.

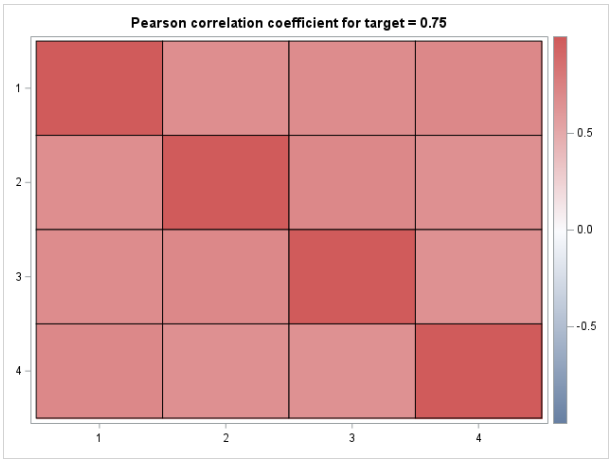

Figure 8: Heatmap when $\rho = 0.75$

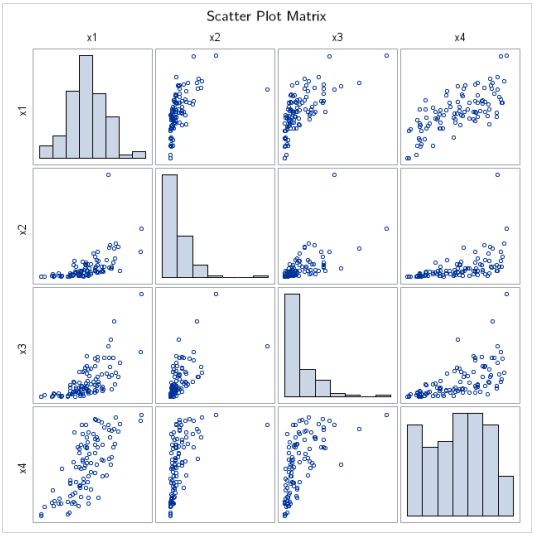

Figure 9: Scatter plot matrix when $\rho = 0.75$

Both graphs clearly show us that the correlation of our variables is very high, so we expect a strong reaction on White's test.

3.4.2 White behavior when the correlation is high

Correlation = 0.75							
				Type I Error			
Observations	Mean NR^2	Median NR^2	Mean R^2	p = 0.85	p = 0.90	p = 0.95	p = 0.99
N = 50	13.28719	12.46136	0.26574	0.1174	0.0838	0.04078	0.0082
N = 100	13.15843	12.14810	0.13158	0.1154	0.0932	0.0496	0.0138
N = 150	13.14204	12.06829	0.08761	0.1238	0.092	0.05194	0.0192
N = 250	13.27551	12.24257	0.05310	0.1376	0.0984	0.05356	0.0186
N = 500	13.39315	12.36259	0.02678	0.1328	0.0926	0.05408	0.018

Table 8: Results with an homoscedastic model, when $\rho = 0.75$

Against all expectations, White's test seems to withstand this collinearity test perfectly. Indeed, the results of rejecting the alternative hypothesis (and therefore confirming homoscedasticity) are almost the same as when the variables were not at all correlated.

3.5 Correlation when $\rho = 0.99$

The results obtained previously seem very surprising. Given those previous results, White is expected to resist this level of correlation (as long as the multicollinearity is not perfect, and the matrix $(X'X)$ remains invertible). A final test under the null hypothesis is carried out, and this time, with an excessively large spearman correlation, to make sure that White will resist anyway.

3.5.1 Graphic representation

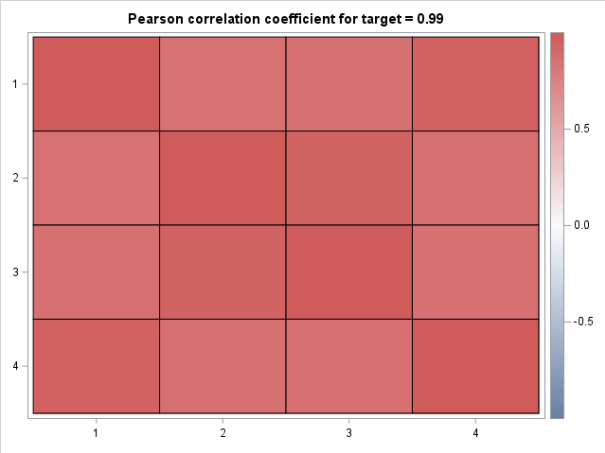

Figure 10: Heatmap when $\rho = 0.99$

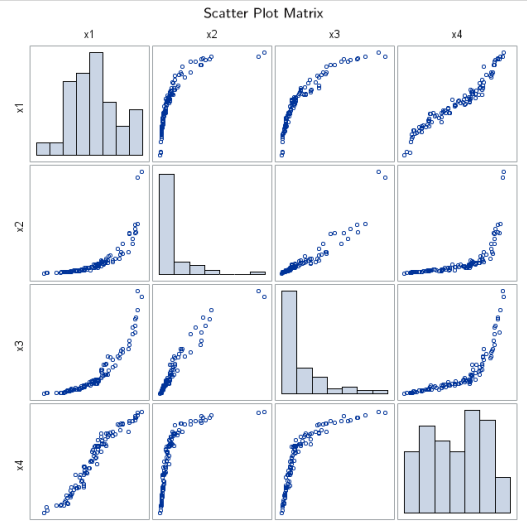

Figure 11: Scatter plot matrix when $\rho = 0.99$

This degree of correlation is almost absurd, the spearman correlation reaches almost 1, the scatterplots show us almost perfect monotonic relationship between each variable and a very high (HeatMap) Pearson's coefficient: it is an almost perfect multicollinearity.

3.5.2 White behavior when the correlation is very high

Correlation = 0.99							
				Type I Error			
Observations	Mean NR^2	Median NR^2	Mean R^2	p = 0.85	p = 0.90	p = 0.95	p = 0.99
N = 50	12.99093	12.16975	0.25981	0.1078	0.063	0.03912	0.0094
N = 100	12.78860	11.79985	0.12788	0.115	0.0796	0.04438	0.0152
N = 150	12.82243	11.78980	0.08548	0.1154	0.0812	0.04586	0.0154
N = 250	12.86746	11.79892	0.05146	0.1096	0.0776	0.04704	0.0148
N = 500	13.39315	12.36259	0.02678	0.1274	0.0886	0.05408	0.0186

Table 9: Results with an homoscedastic model, when $\rho = 0.99$

White's test resists multicollinearity remarkably well. The Type I error does not change, regardless of the level of correlation between the explanatory variables, as well as the number of data generated. White's test is very robust to collinearity under the null hypothesis.

We check in the next section whether this ultra-resistant behavior to multicollinearity will continue to be observed when the generated data are no longer homoscedastic, but heteroscedastic.

4 The robustness of White's test to multicollinearity when the generated data is heteroscedastic

We performed our estimations under the null hypothesis of homoscedasticity, we will now test the efficiency of White's test by adding heteroscedasticity. Basically, we multiply an explanatory variable of our model by ε such as $\mathbb{E}(\varepsilon_t) \neq \sigma_t^2$. In our case, more precisely $x_3 \sim exp(\lambda)$. Given a positive distribution, this will facilitate our analysis, as well as our conclusions.

We note $y = X\beta + \sqrt{x_{3_t}}\varepsilon$. We logically expect the test statistic to reject H_0.

In this section, the scatter plots and heatmaps are not re-added because Iman Conover generates the correlation before the model specification: thus, the correlations will be identical to the previous section, regardless of whether the model is homoscedastic or heteroscedastic.

4.1 Correlation when $\rho = 0$

Correlation = 0							
				Type II Error			
Observations	Mean NR^2	Median NR^2	Mean R^2	p = 0.85	p = 0.90	p = 0.95	p = 0.99
N = 50	30.12362	29.94349	0.60247	0.111	0.1652	0.2602	0.4652
N = 100	51.10119	49.31100	0.51101	0.0032	0.0048	0.0122	0.01792
N = 150	70.72777	67.55017	0.47151	0	0	0.00104	0.003
N = 250	107.5938	102.2428	0.43037	0	0	0	0
N = 500	194.3541	185.1034	0.38871	0	0	0	0

Table 10: Results with an heteroscedastic model, when $\rho = 0$

There is a large Type II error difference between samples $N = 50$ and $N = 100$. Indeed, once the threshold of $N = 100$ is passed, the Type II error (wrongly rejecting H_1 when it is true) converges to 0. The precision of White's test then seems to sharpen between these two critical values. By observing the median of the test statistics we can indeed observe a gradual increase, so by considering a sample twice as large we would go from a slight rejection for a model that is clearly heteroscedastic to an absolute rejection at any treshold. In this case, the less precise p is, the less White makes Type II error.

4.2 Correlation when $\rho = 0.5$

Correlation = 0.5							
				Type II Error			
Observations	Mean NR^2	Median NR^2	Mean R^2	p = 0.85	p = 0.90	p = 0.95	p = 0.99
N = 50	31.07439	31.05279	0.62114	0.0808	0.1266	0.20896	0.4142
N = 100	52.71148	51.20415	0.52041	0.0028	0.0052	0.0133	0.0472
N = 150	72.75075	69.73705	0.485	0	0.0004	0.00072	0.003
N = 250	111.0383	105.8471	0.44415	0	0	0	0
N = 500	199.3685	189.7861	0.39873	0	0	0	0

Table 11: Results with an heteroscedastic model, when $\rho = 0.5$

Moving to a median correlation of 0.5 we can observe an almost similar trend. Everything seems to play out between the 50th and the 100th observation with a decisional convergence taking shape, still rejecting the hypothesis of homoscedasticity with good reason. In the same way if we observe the median of R^2 the differences between the current and previous correlation threshold do not seem significant. The margin could be perfectly explained by the random nature of our data sets, with the possibility of considering a perfect equivalence in asymptotics.

4.3 Correlation when $\rho = 0.75$

With a very high correlation, the model behaves as follows:

Correlation = 0.75							
				Type II Error			
Observations	Mean NR^2	Median NR^2	Mean R^2	p = 0.85	p = 0.90	p = 0.95	p = 0.99
N = 50	31.62532	31.71541	0.6235	0.0838	0.1178	0.19152	0.3906
N = 100	53.52523	52.03640	0.53525	0.0018	0.0044	0.01072	0.0412
N = 150	74.31087	71.50849	0.4954	0	0.003	0.00054	0.0048
N = 250	113.2530	108.0269	0.4530	0	0	0	0
N = 500	203.9823	194.3682	0.4079	0	0	0	0

Table 12: Results with an heteroscedastic model, when $\rho = 0.75$

The trends observed previously are confirmed for a heteroscedastic model, even with a "realisticly" strong correlation. With a large enough number of observation, White never makes Type II error, regardless of p accuracy.

4.4 Correlation when $\rho = 0.99$

Correlation = 0.99							
				Type II Error			
Observations	Mean NR^2	Median NR^2	Mean R^2	p = 0.85	p = 0.90	p = 0.95	p = 0.99
N = 50	32.41087	32.58279	0.6482	0.0722	0.1108	0.17744	0.3768
N = 100	55.32393	53.91722	0.5532	0.0018	0.0042	0.00764	0.0356
N = 150	77.18828	74.42352	0.5145	0	0.0001	0.0003	0.0016
N = 250	118.5656	113.2397	0.47426	0		0	0
N = 500	215.4817	204.5994	0.43096	0		0	0

Table 13: Results with an heteroscedastic model, when $\rho = 0.99$

Correlation does not appear to affect Type II errors since the results shown a similar Type II error for each correlation threshold: the only determinant in this type of configuration will be having a sample N large enough to converge to 0.

5 Interpretation of the results

Obvisouly, the results previously obtained were not expected. Before proceeding to the estimations, everything suggested that White's LM would have been greatly affected by multicollinearity phenomenon.

We will try to provide explanatory elements in the following section in order to better understand these results.

5.1 White's test behavior when confronted with multi collinearity on homoscedastic and heteroscedastic models

5.1.1 Similar resultats between the two models, but..

It seems that the only notable difference between the generation of homoscedastic and heteroscedastic models is the sensitivity to N, the number of observations.

The previous results seem to indicate the following conclusions:

- **Homoscedastic case** : When an homoscedastic model is generated, the test's statistic NR^2 stabilizes itself: when N increases, White's R^2 decreases, thus making the average and the median stable (even if it's actually increasing very weakly) as N increases. Regardless of the correlation threshold chosen, the more precise p is, the less Type I error occurs. However, with a maximum number of observations of 500, the Type I errors will never decrease to zero. White's test is very precise (95%) in his rejection of the alternative hypothesis (heteroscedasticity). He is not affected by collinearity at all.
- **Heteroscedastic case** : When the generated model is heteroscedastic, the NR^2 doesn't stabilize itself : indeed when N increases, R^2 decreases, but not sufficiently. Moreover, unlike the homoscedastic case, White's test never wrongly rejects the alternative hypothesis (heteroscedasticity) as soon as N is large enough. Something fascinating seems to happen: regardless of the collinearity threshold, the less precise p is, the less Type II errors occurs, this is the exact opposite of the homoscedastic case. Another surprising conclusion is that collinearity makes White's test even more efficient: the more variables are correlated and collinear, the less Type II errors occurs.

5.1.2 The White test model: an homeostatic model

This phenomenon of "compensation" (or stabilization) of the White's test statistic NR^2 can be explained by the "negative feedback" effect of the model, mentioned in the study on spurious regressions by Chatelain and Ralf, 2012[8] They called it an homeostatic model. Indeed, the test's statistic will resist an aberrant number of observations by lowering the value of the coefficient of determination R^2 of the model : this is why there is no "wrong" decisions when it comes to the decision rule : Type I and 2 errors are very rare.

[8]Jean-Bernard Chatelain Kirsten Ralf, 2012. "Spurious Regressions and Near-Multicollinearity, with an Application to Aid, Policies and Growth," Université Paris1 Panthéon Sorbonne (Post-Print and Working Papers) halshs-00802579, HAL.

Halbert White explained in his original paper[9], that the R^2 coefficient adjusts itself with respect to the sample given to it: if the original model is homoscedastic then the coefficients in the auxiliary regression are close to 0, which explains why the R^2 value is small. Conversely, the R^2 is very high when the coefficients of the auxiliary regression are different from 0 (like in the case of heteroscadasticity), which explains why it decreases very weakly even if you increase the sample N.

The results seen previously clearly indicate an excellent robustness of White's test to collinearity. Until now, it is known that White is resisting well, but not from where this robustness comes from. We will therefore study this in more details by focusing on the VIF indicator, and an OLS regression allowing us to compare the regressed R^2, with White's one.

5.2 The Variance Inflation Factor (VIF)

It is known that, typically in the case of linear regressions, the presence of multicollinearity can cause many problems, especially in the case of the OLS where it tends to inflate the variances of the estimated parameters, thus drastically reducing their significance. However, the previous results seem to be almost insensitive to it, so that we are led to suspect that multicollinearity is only artificial, or even insignificant in our models.

Multicollinearity is not strictly speaking tested. It is a phenomenon that can be detected by studying characteristic indicators of the presence of collinearity. For example, a well-known indicator is the VIF (Variance Inflation Factor), calculated as follows:

$$VIF(\hat{\beta}_i) = \frac{1}{1 - R_i^2} \tag{16}$$

This indicator measures the extent to which the variance rate explained by the model increases when the explanatory variables are correlated. It is considered that there is no multicollinearity problem when this indicator is less than 5, that there may be a problem when it is between 5 and 10, and that there is a real multicollinearity problem when the indicator is greater than 10.

The intuition behind this indicator is very simple, the greater the coefficient of determination is, the greater the VIF will be (and will tend towards infinity).

Since no real difference was noted during the progressive analysis of the data generated at certain degrees of correlation, we limited ourselves to the case of the two extremes to confirm or not the possibility of problematic multicollinearity in our models. We therefore carried out two regressions using the case of the identity matrix and, and in an other hand, a matrix with a Spearmann correlation rank of 0.99.

The idea in this practice is to compare whether multicollinearity is indeed present in our data at the output of the algorithm, and then whether the absence of any real effect previously observed can be explained by a post-white transformation annihilation (1).

Thus we will consider two models of the form :

$$x_1 = b_0 + b_2 x_2 + b_3 x_3 + b_4 x_4 + \varepsilon \tag{17}$$

[9]WHITE, Halbert. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: journal of the Econometric Society, 1980, p. 817-838.

$$x_1 = b_0 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_1^2 + ... + b_8x_4^2 + b_9x_1x_2 + ... + b_{14}x_3x_4 + \varepsilon \tag{18}$$

That will be regressed by the ordinary least squares (OLS). We followingly perform a first test at the correlation threshold of the identity matrix, we obtain the following results.

Variable	DF	Parameter Estimate	Standard Error	t Value	$Pr > t$	Tolerance	Variance Inflation	R-squared
Intercept	1	0.07341	0.10605	0.69	0.4891	-	0	0.0044
x_2	1	-0.03591	0.02675	-1.34	0.1801	0.99995	**1.00005**	Adjusted R-squared
x_3	1	0.00916	0.04199	0.22	0.8274	0.99945	**1.00055**	-0.0016
x_4	1	0.09094	0.15592	0.58	0.5600	0.99948	**1.00052**	

Table 14: VIF associated with the x_1 OLS regression when $\rho = 0$

The above results show that with uncorrelated data generated (target matrix of iman conover filled with 0), the R^2 is close to 0. This R^2 value is rather logical : we generated an uncorrelated model, and then regressed x_1 on the other independent explanatory variables.

Variable	DF	Parameter Estimate	Standard Error	t Value	$Pr > t$	Variance Inflation	R-squared
Intercept	1	-0.03796	0.07962	-0.48	0.6338	0	0.8453
x_2	1	-0.00193	0.02491	-0.08	0.9382	10.31964	Adjusted R-squared
x_3	1	-0.03398	0.05356	-0.63	0.5261	8.55442	0.8411
x_4	1	0.14252	0.27082	0.53	0.5990	18.06502	
x_1^2	1	0.00116	0.01212	0.10	0.9241	1.09317	
x_2^2	1	-0.00193	0.00124	-1.56	0.1199	5.06528	
x_3^2	1	-0.01380	0.01128	-1.22	0.2219	6.856612	
x_4^2	1	-0.22814	0.24682	-0.92	0.3558	16.43199	
x_1x_2	1	0.05331	0.00747	7.14	$< .0001$	1.72226	
x_1x_3	1	0.17831	0.01785	9.99	$< .0001$	1.78295	
x_1x_4	1	1.12866	0.04474	25.23	$< .0001$	2.07702	
x_2x_3	1	0.01107	0.00772	1.43	0.1526	3.34346	
x_2x_4	1	0.03071	0.03193	0.96	0.3367	4.93628	
x_3x_4	1	0.11965	0.06919	1.73	0.0844	6.41027	

Table 15: Variance Inflation Factor (VIF) of the White model when $\rho = 0$

However, once White's transformation is done, the R^2 massively increases, it indeed approximates 0.84. This phenomenon is characteristic of the presence of multicollinearity, and it's confirmed by the VIF indicator : According to the VIF, some explanatory variables seem to present multicollinearity: x_2, x_4^2 and x_3x_4 indicate a relationship of strong collinearity, and we can suspect the presence of a collinearity between x_3^2 and x_3x_4.

We therefore come to the following conclusion : White is by construction a source of multicollinearity, even if the regressed variables are not multicollinear. It is therefore necessary to verify whether this behaviour is observed when the data are higly correlated.

We now perform the regression with the extreme threshold of 0.99 correlation.

Variable	DF	Parameter Estimate	Standard Error	t Value	$Pr > t$	Tolerance	Variance Inflation	R-squared
Intercept	1	-1.45651	0.02162	-67.38	< .0001	-	0	0.9534
x_2	1	0.06812	0.01256	5.42	< .0001	0.13856	**7.21720**	Adjusted R-squared
x_3	1	0.06966	0.03988	1.75	0.0813	0.06404	**15.61458**	0.9531
x_4	1	2.68199	0.07247	37.01	< .0001	0.19246	**5.19589**	

Table 16: VIF associated with the x_1 OLS regression when $\rho = 0.99$

The regression performed on the initial model reports a very high R^2. This may give an indication in the presence of multicollinerality. The VIF nevertheless partially confirms these assumptions with values that are certainly quite far from the threshold conventionally estimated at a problematic level of 10, but there is no consensus. In the literature, some people think that we can already be concerned about values close to 5, but it is interesting to show that the manipulation of the correlation matrix has played a role when compared to the previous outputs. Similarly, it is expected that the values will explode for White's post-transformation values.

Variable	DF	Parameter Estimate	Standard Error	t Value	$Pr > t$	Variance Inflation	R-squared
Intercept	1	-0.100177	0.01159	-86.41	< .0001	0	0.9994
x_2	1	0.28793	0.03802	7.57	< .0001	4023.32989	Adjusted R-squared
x_3	1	-0.53593	0.07305	-7.34	< .0001	4013.98986	0.9994
x_4	1	3.94750	0.08332	47.38	< .0001	444.18155	
x_1^2	1	-0.23682	0.00230	-102.97	< .0001	8.49420	
x_2^2	1	-0.01006	0.00099703	-10.09	< .0001	689.82354	
x_3^2	1	0.04863	0.01038	4.69	< .0001	1912.33332	
x_4^2	1	-3.80586	0.11358	-33.51	< .0001	919.69218	
x_1x_2	1	0.18666	0.01817	10.27	< .0001	6227.09022	
x_1x_3	1	-0.36469	0.04101	-8.89	< .0001	5896.91338	
x_1x_4	1	1.96575	0.03790	51.87	< .0001	390.90426	
x_2x_3	1	-0.00532	0.00880	-0.60	0.5459	4533.63986	
x_2x_4	1	-0.49054	0.05756	-8.52	< .0001	9588.26209	
x_3x_4	1	0.89769	0.11639	7.71	< .0001	10220	

Table 17: Variance Inflation Factor (VIF) of White's model when $\rho = 0.99$

Indeed, the results are all the more convincing, if the VIF remains an indicator, we observe here an explosion of the results that leave little doubt as to the presence of multicollinearity in the series, with some of them exceeding the thousand.

6 Conclusion

The reliability of White's Heteroscedasticity test has been strongly verified by the use of Iman Conover's algorithm and leads us to satisfactory conclusions. First of all, the general interpretation tends to confirms that White's test is completely valid. In the case where a model is constructed on the basis of the constancy of variance of the residuals (homoscedasticity), the precision of the test is as expected for the different tresholds, which shows a very high robustness. On the other hand, if the residuals depend on the explanatory variables of the model (heteroscedasticity), the precision of the test tends towards perfection, i.e. 100%, especially when the sample size is large. Looking at the results, it is found that regardless of the level of correlation when the number of observations is increased, the coefficient of determination R^2 of the auxiliary regression from the White test gradually decreases. Indeed, in case of homoscedasticity(H_0), we notice that the R^2 decreases enormously as the number of observations increases: here we have a compensation phenomenon: the R^2 adjusts itself with respect to the sample it is given because if the original model is homoscedastic then the coefficients in the regression are close to 0 which explains why the R^2 is small. Using a Goodness of Fit test, the graphical analysis of our data seems to indicate that the NR^2 follow a "Chi-squared" distribution. Even if not confirmed with the statistical results using the Kolmogorov-Smirnov test. Conversely, the R^2 is very high when the parameters of the regression are generally different from 0 in case of heteroscedasticity (H_1) which explains why it decreases very little even after increasing the sample size.

Also, it is an important thing to notice that when N is very large, the NR^2 score test statistic actually follows a Chi-squared distribution regardless of the level of correlation.Finally, the hypothesis of multicollinearity causing problems to the robustness of White's test may be questionable. Indeed, we observed a phenomenon of resilience to collinearity, regardless of the size or nature of the residuals of the model, the results shown through indicators such as the VIF, leads us to conclude that multicollinearity may be present, but White completely ignores it.

References

[1] Régis Bourbonnais et al. *Econométrie*. Dunod Paris, France, 2011.

[2] Jean-Bernard Chatelain and Kirsten Ralf. Spurious regressions and near-multicollinearity, with an application to aid, policies and growth. *Journal of macroeconomics*, 39:85–96, 2014.

[3] Ronald L Iman and WJ Conover. Communication in statistics-simulation and computation. *A distribution-free approach to inducing rank correlation among input variables*, 11(3):311–34, 1982.

[4] John Johnston and John DiNardo. Econometric methods. 1963.

[5] Jean-Pierre Lecoutre. *Statistique et probabilités*. Dunod, 2002.

[6] Valérie Mignon. *Econométrie : Théorie et applications*. 2008.

[7] Igor V Tetko, David J Livingstone, and Alexander I Luik. Neural network studies. 1. comparison of overfitting and overtraining. *Journal of chemical information and computer sciences*, 35(5):826–833, 1995.

[8] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, pages 817–838, 1980.

[9] Rick Wicklin. *Simulating data with SAS*. SAS Institute, 2013.