ARTICLE TEMPLATE

Variance Inflation Factor and Condition Number in multiple linear regression

ARTICLE HISTORY
Compiled April 5, 2018

ABSTRACT
The Variance Inflation Factor and the Condition Number are measures traditionally applied to detect the presence of collinearity in a multiple linear model. This paper presents the relation and the difference between both measures from theoretical and empirical perspectives by using Monte Carlo simulations and taking special interest in the computational techniques.

KEYWORDS
Variance Inflation Factor; Condition Number; multicollinearity detection; data transformation

1. Introduction

The presence of high collinearity in a multiple linear regression model implies that the conclusions of the analysis can be questioned, **for example,** because of a lack of accuracy of the estimations due to the high variances of the estimators. Thus, the detection of collinearity has to be a compulsory first step in every econometric analysis.

The measures most applied to detect collinearity are the Variance Inflator Factor (VIF) and the Condition Number (CN), although they are based on concepts not included in the most accepted definitions of collinearity given by [1]: "k variables are collinear or nearly dependent, if one of them lies almost in the space spanned by the remaining $(k-1)$ variables, that is, if the angle between one and its orthogonal projection on the others is small". What happens is that "collinearity evidently implies different things to different people. Some associate collinearity primarily with numerical problems and sensitivity, while others concentrate on variance inflation and related statistical concerns", [2].
Thus, given the following linear model with n observations and p exogenous variables,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{1}$$

where **the first column of X is composed of ones** and $\mathbf{u}$ represents the random disturbance (that is supposed to be spherical), the VIF for every exogenous variable in model (1) is obtained from the following expression:

$$VIF(i) = \frac{var\left(\widehat{\beta}_i\right)}{var\left(\widehat{\beta}_i^o\right)} = \frac{1}{1-R_i^2}, \quad i = 2, \dots, p, \tag{2}$$

being $\widehat{\boldsymbol{\beta}}$ the OLS estimator of model (1), $\widehat{\boldsymbol{\beta}}^o$ the OLS estimator of model (1) supposing that the exogenous variables are orthogonal and R_i^2 is the coefficient of determination of the following auxiliary regression:

$$\mathbf{X}_i = \mathbf{X}_{-i}\boldsymbol{\delta} + \mathbf{w},$$

where $\mathbf{X}_{-i}$ is equal to the matrix $\mathbf{X}$ after eliminating the variable $\mathbf{X}_i$ for all $i = 2, \dots, p$. Since $0 \leq R_i^2 \leq 1$, it is verified that $VIF(i) \geq 1, \forall i$.

Because the VIF is obtained as the ratio between the observed variance and the variance that will be obtained if $\mathbf{X}_i$ is uncorrelated with the rest of the exogenous variables, it shows how much the variance of the estimator is inflated as a consequence of the linear relation between the regressors. However, it should be noted that collinearity may not be related to the correlation. There can be multicollinearity between explanatory variables without there being high correlation between pairs of these variables, [3] and [4]. Belsley [1] summarized this idea with the following statement: “low VIFs do not guarantee low collinearity”.

On the other hand, given the linear model (1), the condition number (CN) is defined as:

$$K(\mathbf{X}) = \frac{\mu_{max}}{\mu_{min}}, \tag{3}$$

where μ_{max} and μ_{min} are the minimum and maximum singular values of matrix X, respectively.

From the decomposition of the singular values of matrix X given by $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^t$ where $\mathbf{U}^t\mathbf{U} = \mathbf{I}$, $\mathbf{V}^t\mathbf{V} = \mathbf{I}$ being I the identity matrix (with adequate dimensions) and $\mathbf{D} = diag(\mu_1 \dots \mu_p)$, with μ_i, $i = 1, \dots, p$, the eigenvalues of matrix X, then:

$$\mathbf{X}^t\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{U}^t\mathbf{U}\mathbf{D}\mathbf{V}^t = \mathbf{V}\mathbf{D}^2\mathbf{V}^t. \tag{4}$$

In this case, the eigenvalues of the matrix $\mathbf{X}^t\mathbf{X}$ coincide with the square of the singular values of matrix X, that is, $\xi_i = \mu_i^2$ for $i = 1, \dots, p$. Then, expression (3) is equivalent to:

$$K(\mathbf{X}) = \sqrt{\frac{\xi_{max}}{\xi_{min}}}, \tag{5}$$

where ξ_{max} and ξ_{min} are, respectively, the maximum and minimum eigenvalues of matrix $\mathbf{X}^t\mathbf{X}$, [5], [6], [1], [7], [8] and [9]. Note that data should have unit length, that is to say, the data should be divided by the square root of the sum of its squared elements. It is a measure related to the ill-conditioning of matrix $\mathbf{X}^t\mathbf{X}$ from a numerical point of view. Steward [10] noted that “one problem is that while the condition number can be very useful as a multicollinearity indicator, it may not be specific enough for statistical applications since it distils a large amount of information into a single number”.

Unfortunately, both measures are not a statistical contrast to detect collinearity. [11] questioned the Chi-square test for the existence of multi-collinearity, modified by [12]. A consequence of the skepticism developed here is a return to the position of treating multicollinearity as a numerical problem, i.e., with generally accepted thresholds. **By**

following [13] "commonly a VIF of 10 or even one as low as 4 have been used as rules of thumbs to indicate excessive or serious collinearity" [14–18]. In relation to condition number, [6] stated that values of $K(\mathbf{X})$ between "the range $0-10$ indicate weak near dependencies, $10-30$ indicate moderately strong near dependencies, $30-100$ strong near dependencies, and indices in excess of 100 are very strong". However, this should always be examined in context, [13], as there are cases where even a very high VIF (or CN) does not require corrective action, [18].

[19] distinguished between high correlation among regressors which, under certain conditions, gives rise to *systematic volatility*, and a numerical issue (the regressor data matrix $\mathbf{X}^t\mathbf{X}$ is ill-conditioned), which gives rise to *erratic volatility.* Holland [20] detailed the connection between ill-conditioning and multicollinearity: "Moreover, the terms ill-conditioning and collinearity are also sometimes used interchangeably, though ill-conditioning describes any effect in a data matrix that causes large changes in the regression estimates, due to a small change in the data, so does not involve multicollinearity alone [1]. However, multicollinearity is the primary cause of such behaviour in regression models".

From this idea, the condition number can be inserted into the problem of the stability of the $\mathbf{X}^t\mathbf{X}$ matrix, from a numerical analysis point of view, while the VIF is inspired by a statistical point of view based on the correlation between the regressors. Both parameters, although not directly related to collinearity, attempt to measure the presence of collinearity in the estimation of a linear model. This paper attempts to clarify what is measured by the VIF and the CN and to find a relation between them from a conceptual or purely arithmetical point of view. In the case that there is some kind of relationship between them, there should also be some relation between their thresholds.

This is not the first time this question has been raised. [21] stated that the VIF and the CN are related, establishing the following inequality **for standardized data:**

$$\max_{i=2,\ldots,p} VIF(i) \leq K(\mathbf{Z})^2 < (p-1) \cdot \sum_{i=2}^{p} VIF(i), \tag{6}$$

where $\mathbf{Z}^t\mathbf{Z}$ is the matrix of correlations, that is to say, the data are standardized (its mean is zero and its variance is equal to 1 divided by the number of observations). As consequence, the **square of** CN is a upper bound of the maximum VIF. Thus, the CN may include information about the grade of collinearity that is not detected by the VIF.

In the case of standardized data with $p=3$, the following is verified:

$$\mathbf{Z}^t\mathbf{Z} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \frac{1}{VIF} = 1-\rho^2 = det(\mathbf{Z}^t\mathbf{Z}) = \xi_1 \cdot \xi_2 = (1-\rho)\cdot(1+\rho),$$

where ρ is the coefficient of correlation between the exogenous variables[1], then supposing that $\rho > 0$:

$$K(\mathbf{Z}) = \frac{1}{1-\rho} \cdot \sqrt{\frac{1}{VIF}} = (1+\rho)\cdot\sqrt{VIF}. \tag{7}$$

Note that if $VIF = 10$ then $\rho^2 = 0.9$. Consequently, $VIF > 10$ is equivalent

[1] Note that the constant term disappears after the standardization of the data.

to $K(\mathbf{Z}) > 6.162$.

Both relations (6) and (7) are established when the data are standardized. However, for the calculation of the CN, the data should be expressed in unit length.

The structure of the paper is as follows: after presenting the notation in Section 2, Section 3 presents a relation between the CN and the VIF for standardized data as an alternative to the relation established by [21]. This relation's adequateness for unit length data is also analyzed by using Monte Carlo simulations. Analogously to the relation showed by [21], the **square of** CN is a upper bound of the VIF. Based on this conclusion, Section 4 analyzes if there is a relation between the regressors that is captured by the CN and not by the VIF. In section 5, we use Monte Carlo simulations to analyze whether the functional relation shown in (7) is verified for $p = 3, 4, 5$. Finally, the main contributions of the paper are summarized in section 6.

2. Notation

For model (1), X denotes the information matrix for n observations and p variables when natural units are used and a constant term is included. This is to say, $\mathbf{X} = (\mathbf{1}, \mathbf{X}_2, \dots, \mathbf{X}_p)$ where $\mathbf{1} = (1 \dots 1)^t$. Further, considering that $\mathbf{X} = (x_{ij})$ with $i = 1, \dots, n$ and $j = 1, \dots, p$:

- **Data are considered to be unit length when original uncentered data are divided by the square root of the sum of every variable squared. In this case, the information matrix is noted as $\mathbf{U} = (u_{ij})$ with $i = 1, \dots, n$ and $j = 1, \dots, p$. Then, $u_{ij} = \frac{x_{ij}}{||\mathbf{X}_j||}$, where $||\mathbf{X}_j|| = \sqrt{\sum_{i=1}^{n} x_{ij}^2}$.**
- **If data are unit length, then:**

$$\overline{\mathbf{U}}_j = \frac{\overline{\mathbf{X}}_j}{||\mathbf{X}_j||}, \quad var\left(\mathbf{U}_j\right) = \frac{var\left(\mathbf{X}_j\right)}{||\mathbf{X}_j||^2}, \quad ||\mathbf{U}_j|| = 1.$$

- **Data are considered to be typified when original centered data are divided by their standard deviation. In this case, the information matrix is noted as $\mathbf{T} = (t_{ij})$ with $i = 1, \dots, n$ and $j = 1, \dots, p$. Then, $t_{ij} = \frac{X_{ij} - \bar{X}_j}{\sqrt{var(X_j)}}$.**
- **If data are typified their mean is zero, $\overline{\mathbf{T}}_j = 0$, variance is equal to 1, $var\left(\mathbf{T}_j\right) = 1$, and the cross products matrix is equal to the correlation matrix multiplied by the number of observations, $\mathbf{T}^t\mathbf{T} = n \cdot \mathbf{R}$, where R is the correlation matrix obtained from X.**
- **Data are considered to be standardized when original centered data are divided by their standard deviation multiplied by the square root of the number of observations. In this case, the information matrix is noted as $\mathbf{Z} = (z_{ij})$ with $i = 1, \dots, n$ and $j = 1, \dots, p$. Then, $z_{ij} = \frac{X_{ij} - \bar{X}_j}{\sqrt{n \cdot var(X_j)}}$.**
- **If data are standardized their mean is zero, $\overline{\mathbf{Z}}_j = 0$, variance is equal to 1 divided by the number of observations, $var\left(\mathbf{Z}_j\right) = \frac{1}{n}$, and the cross products matrix is equal to the correlation matrix, $\mathbf{Z}^t\mathbf{Z} = \mathbf{R}$.**

Note that the VIF is invariant to origin and scale transformations, [22]. Consequently, the same value is obtained using the expression 2 for original, standardized, unit length or typified data. Contrarily, the value of the CN depends on the data transformation [22] and [23]. Although it coincides

with typified and standardized data, it differs in the rest of situations. Unless another indication is provided, in this paper the CN will be calculated from expression (5) and with unit length data.

3. Relationship between the Variance Inflation Factor and eigenvalues of $\mathbf{X}^t\mathbf{X}$

Given the linear model (1) and considering that the data are standardized, the diagonal elements of $\left(\mathbf{Z}^t\mathbf{Z}\right)^{-1}$ are the VIF. **By following [24]** and parting from (4) we obtain:

$$VIF(i) = \sum_{k=2}^{p} \frac{v_{ik}^2}{\xi_k},$$

where v_{ik} are the elements of the eigenvector corresponding to the eigenvalue ξ_k, $i, k = 2, \dots, p$. In this case,

- since $\xi_k \leq \xi_{max}$ for all k, it is verified that $VIF(i) \geq \frac{1}{\xi_{max}} \sum\limits_{k=2}^{p} v_{ik}^2 = \frac{1}{\xi_{max}}$.
- since $\xi_k \geq \xi_{min}$ for all k, it is verified that $VIF(i) \leq \frac{1}{\xi_{min}} \sum\limits_{k=2}^{p} v_{ik}^2 = \frac{1}{\xi_{min}}$.

Then, it is verified that:

$$\frac{1}{\xi_{max}} \leq VIF(i) \leq \frac{1}{\xi_{min}} \Leftrightarrow 1 \leq \xi_{max} \cdot VIF(i) \leq K(\mathbf{Z})^2. \tag{8}$$

If it is considered that the collinearity existing in model (1) is worrying when there is a variable i that leads to a $VIF(i) > 10$, it will be verified in relation (8) when:

$$\frac{\xi_{max}}{\xi_{min}} = K(\mathbf{Z})^2 \geq \xi_{max} \cdot VIF(i) > 10 \cdot \xi_{max} \Rightarrow \xi_{min} < 0.1. \tag{9}$$

Remark 1. Note that $VIF(i) > 10$ implies that $\xi_{min} < 0.1$ but $\xi_{min} < 0.1$ does not imply that $VIF(i) > 10$. However, from expression (8) is observed that $\xi_{min} > 0.1$ implies $VIF(i) < 10$.

3.1. Monte Carlo simulation

To study if Remark 1 is also verified when working with original or unit length data[2], values are simulated for:

$$\mathbf{X}_i = \sqrt{1 - \gamma^2} \cdot \mathbf{W}_i + \gamma \cdot \mathbf{W}_p,$$

where $i = 2, \dots, p$ with $p = 3, 4, 5$, $\mathbf{W}_i \sim N(10, 100)$, $\gamma \in \{0, 0.05, 0.1, 0.15, \dots, 0.95\}$ and $n \in \{15, 20, 25, 30, \dots, 200\}$. This way of generating independent variables with different grades of collinearity (γ is specified so that the correlation between any two independent variables is given by γ^2), as previously applied, for example, by [25–27].

[2] Note that, when data are standardized, the VIF and CN coincide with the result obtained from typified data.

The matrix $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \dots \ \mathbf{X}_p]$ is constructed such that $\mathbf{X}_1$ is a vector with ones (representing the constant term in model (1)). Then, the maximum VIF and the minimum eigenvalue are calculated considering that values are original ($\mathbf{X}$) or transformed to be unit length ($\mathbf{U}$).

Since this calculation is repeated 1000 times, 760000 values are obtained from the maximum VIF and minimum eigenvalue, as shown in Figure 3.1. It is observed that **Remark 1** is not verified when working with original data (first column) but it seems to be verified for unit length data (second column).

4. Differences between the Variance Inflation Factor and Condition Number

From expression (6), it can be concluded that the **square of** CN is a upper bound of the VIF. For this reason, it could be interesting to analyze if the CN is able to capture any kind of relation between the regressors that is ignored by the VIF. With **illustrative** purpose, the following examples are developed[3]:

Example 4.1. The following table shows the VIF and CN for the variables that compound the matrix given by $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \mathbf{X}_3]$, where the variables $\mathbf{X}_2$ and $\mathbf{X}_3$ are orthogonal, i.e., $\mathbf{X}_2^t\mathbf{X}_3 = 0$:

$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	VIF(2)	VIF(3)	CN	
1	1	-0.5	1	1	6.793	Original data
1	3	-0.5	1	1	5.095	Unit length data
1	2	1	1	1	1	Standardized data

First, it is observed that the VIF is invariant to origin and scale changes (property inherited from the coefficient of determination), while the CN is not invariant, leading to different results depending on the transformation of the data.

Second, since the values of VIFs are equal to 1, the orthogonality between $\mathbf{X}_2$ and $\mathbf{X}_3$ is fully detected in all cases, which occurs with the CN only when data are standardized **since in this case the constant term disappears**.

Third, in the first two cases, the CN is different from 1, showing the relation between $\mathbf{X}_2$ with $\mathbf{X}_1$ (note that $\mathbf{X}_1^t\mathbf{X}_2 = 6$ and $\mathbf{X}_1^t\mathbf{X}_3 = 0$).■

Example 4.2. From the previous example, it can be concluded that the VIF ignores the relation between the constant term and the rest of the regressors, which does not occur with the CN. In order to confirm this statement, the following table presents the values of VIF and CN for the variables that compounds the matrix given by $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \mathbf{X}_3 \ \mathbf{X}_4]$, where all the variables are orthogonal, that is to say, $\mathbf{X}^t\mathbf{X}$ is a diagonal matrix:

$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	VIF(2)	VIF(3)	VIF(4)	CN	
1	1	1	-0.833	1	1	1	1.603	Original data
1	1	0	1.166	1	1	1	1	Unit length data
1	-2	0.5	0.166	1	1	1	1	Standardized data
1	0	-1.5	-0.5					

Analogously to the previous example, the VIF captures the orthogonality between

[3] Note that these examples are not regression models since $n = p$.

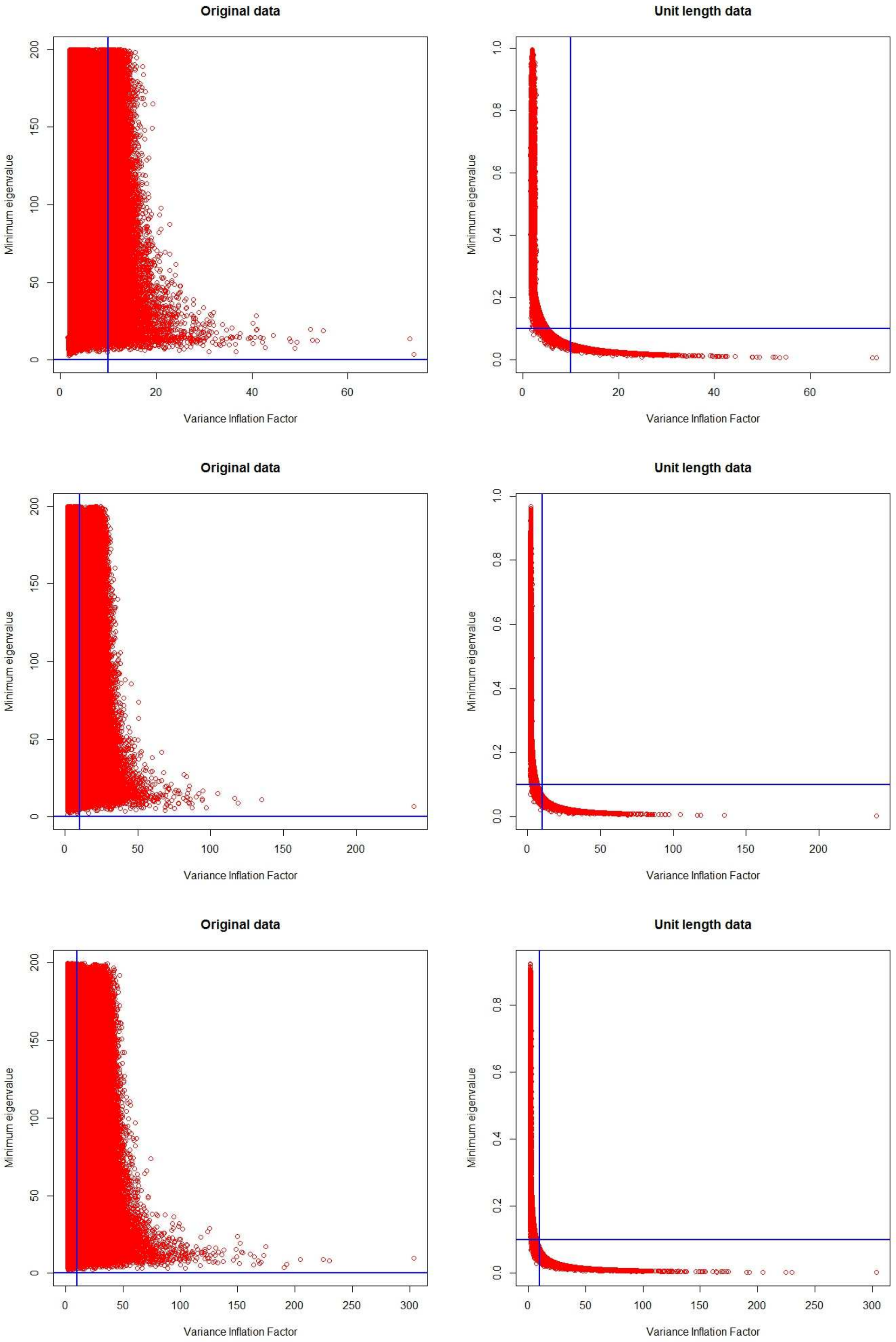

Figure 1. Dispersion diagram for the maximum VIF and the minimum eigenvalue for the 760000 simulated cases: by columns, the transformation data (original or unit length data), and by rows, the number of variables ($p = 3, 4, 5$). The vertical line shows $VIF = 10$, and the horizontal one shows $\xi_{min} = 0.1$

the regressors in all cases and the CN with data that are standardized or in unit length, but not with the original data.

Thus, it is possible to conclude that i) a data transformation is required since the CN can not be calculated from original data and ii) the standardization of the model eliminates the constant term and, for this reason, unit length data should be used.■

Example 4.3. To obtain a independent variable linearly related to the constant term, the independent variable must be almost constant, that is to say, with a very small variance. For this purpose, the following example, previously applied by [3], is developed:

$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	VIF(2)	VIF(3)	CN	
1	1	1	1	1	3160.47	Unit length data
1	1.001	1.003	1	1	1	Standardized data
1	1.002	1				

In this case, there are two variables, $\mathbf{X}_2$ and $\mathbf{X}_3$, that are almost orthogonal (its coefficient of correlation is equal to $-6.4 \cdot 10^{-14}$). Indeed, their VIFs are equal to 1. However, both variables are almost constant (their quasivariances are equal to 10^{-6} and $3 \cdot 10^{-6}$, respectively), and can be highly related to the constant term. This question is confirmed with the value obtained for the CN with unit length data.

Note that this example confirms the previous conclusion that the CN calculated with standardized data ignores the relation between the exogenous variables and the constant term since it has been eliminated.■

It has been shown that the VIF is invariant to data transformation while the CN is not (See [22] and [23]). Thus, the CN should not be calculated from the original data. A previous transformation is necessary as was recommended by [5]: ***transforms a data matrix X with mutually orthogonal columns, the standard of ideal data, into a matrix whose condition indexes could be all unity, the smallest (and therefore most ideal) condition indexes possible.*** **This goal is achieved by transforming the data to be unit length or standardized. The great difference between both transformations is that the standardization eliminates the constant term. For this reason, when working with standardized data, the relation between the exogenous variables and the constant term is not captured, in contrast to what happens with unit length data.**

4.1. Monte Carlo simulation

The previous examples have shown that the difference between the VIF and the CN is that the VIF ignores the relation of the exogenous variables with constant term, in contrast to the CN calculated with unit length data. The constant term will be related to an independent variable when this latter variable is almost constant, that is to say, with a very small variance. However, the question is, how small does the variance of the independent variable have to be to lead to a worrying relation with the constant term?

To answer this question, values have been simulated for $\mathbf{X}_2 \sim N(1, \sigma)$, where $\sigma^2 = 0.1, 0.01, 0.005, 0.001, 0.0005, 0.0001$ and $n \in \{15, 20, 25, 30, \dots, 200\}$. Then, the following matrix has been developed $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2]$, where $\mathbf{X}_1$ is a vector of ones with the appropriate dimensions. Next, the matrix is transformed to have unit length ($\mathbf{U}$),

Table 1. Minimum, mean and maximum values for the maximum VIF and the CN with unit length data

	VIF			CN		
σ^2	Minimum	Mean	Maximum	Minimum	Mean	Maximum
0.1	1	1	1	3.4306	6.596	15.971
0.01	1	1	1	12.464	20.4408	49.345
0.005	1	1	1	18.375	28.872	67.833
0.001	1	1	1	37.399	64.481	138.188
0.0005	1	1	1	58.978	91.1706	210.296
0.0001	1	1	1	107.352	203.7102	479.615

and the CN and the VIF are calculated. Since this operation is repeated 1000 times, the values of Table 1 corresponds to the 38000 values of the VIF and the CN for every case.

It is expected that the relation between $\mathbf{X}_1$ and $\mathbf{X}_2$ will become stronger as the value of σ^2 decreases. However, this expectation is not reflected in the VIF, **which is always equal to 1.**[4]. What it is verified is that the CN increases as the variance of the variable diminishes, capturing the relation between this independent variable and the constant term.

Finally, taking into account the thresholds provided by Belsley [6] (**and stated in the introduction**) for the values of the CN, it is possible to conclude that the linear regression will be worrying when $\sigma^2 < 0.005$.

5. Relationship between the Variance Inflation Factor and Condition Number: Monte Carlo simulation

In the introduction section, a functional relation is established between the VIF and the CN given by a squared root when $p = 3$ and the data are standardized; see expression (7). In this section, **we seek a relationship between the CN and maximum VIF** when $p = 3, 4, 5$ and for standardized and unit length data.

With this purpose, and similarly to section 3, values are simulated for

$$\mathbf{X}_i = \sqrt{1-\gamma^2} \cdot \mathbf{W}_i + \gamma \cdot \mathbf{W}_p,$$

where $i = 2, \dots, p$ with $p = 3, 4, 5$, $\mathbf{W}_i \sim N(10, 100)$, $\gamma \in \{0, 0.01, 0.02, 0.03, \dots, 0.99\}$ and $n \in \{15, 20, 25, 30, \dots, 200\}$. Note that the variances of the simulated variables are high, and consequently, **it is expected that there is no linear** relation between them and the constant term.

Then, the matrix $\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \dots \ \mathbf{X}_p]$ is developed where $\mathbf{X}_1$ is a vector of ones with the appropriate dimensions and the maximum VIF and CN are calculated for standardized ($\mathbf{Z}$) and unit length ($\mathbf{U}$) data. Since this operation is repeated 1000 times,

[4] **Denoting $\mathbf{X}_1 = \mathbf{1}$, the auxiliary regression to calculate the VIF is expressed as $\mathbf{X}_2 = \gamma\mathbf{1} + \mathbf{w}$, where it is verified that $\widehat{\gamma} = \overline{\mathbf{X}}_2$ and, consequently, $SSR = \sum\limits_{i=1}^{n} (X_{2i} - \overline{\mathbf{X}}_2)^2 = SST$. In this case, it is always verified that $R^2_{aux} = 1$.**
The version of the previous regression with unit length data is given by $\mathbf{X}_{2,lu} = \gamma\mathbf{1}_{lu} + \mathbf{w}$ where $\mathbf{X}_{2,lu} = \mathbf{X}/\sqrt{a}$ with $a = \sum\limits_{i=1}^{n} X^2_{2i}$ and $\mathbf{1}_{lu} = \mathbf{1}/\sqrt{n}$. In this case, $\widehat{\gamma} = \sqrt{\frac{n}{a}} \cdot \overline{\mathbf{X}}_2$ and, then, $SSR = \frac{1}{a} \sum\limits_{i=1}^{n} (X_{2i} - \sqrt{n} \cdot \overline{\mathbf{X}}_2 \cdot \frac{1}{\sqrt{n}})^2 = SST$. Thus, this situation will be similar to the initial one.

Table 2. Coefficient of determination for regressions (10) **to** (13), taking into account the number of simulated variables and the nature of the data.

		$\log(maxVIF)$		$\sqrt{maxVIF}$	
p	Data	Regression (10)	Regression (12)	Regression (11)	Regression (13)
3	Standardized	0.9358	0.959	**0.9925**	0.987
4	Standardized	0.8963	0.9478	**0.9969**	0.9937
5	Standardized	0.868	0.9308	**0.9984**	0.9953
3	Unit length	0.9404	**0.972**	0.9411	0.955
4	Unit length	0.9309	0.9627	0.9742	**0.9792**
5	Unit length	0.9099	0.9465	0.9857	**0.9871**

Figure 2 displays the 3800000 values for the maximum VIF and the CN. **From the observation of Figure 2, we consider appropriate to analyze the following relations:**

To analyze the adequateness of both relations, the following regressions are estimated:

$$
\begin{aligned}
CN &= \beta_1 + \beta_2 \cdot \log(maxVIF) + \beta_3 \cdot n + \epsilon, & (10)\\
CN &= \beta_1 + \beta_2 \cdot \sqrt{maxVIF} + \beta_3 \cdot n + \epsilon, & (11)\\
CN &= \alpha_1 \cdot \log(maxVIF) + \alpha_2 \cdot n + \epsilon, & (12)\\
CN &= \alpha_1 \cdot \sqrt{maxVIF} + \alpha_2 \cdot n + \epsilon, & (13)
\end{aligned}
$$

where the effect of the small sample on the relation between both measures has been included. **The CN is analyzed as a function of the VIF (and not vice versa), since the interpretation of the threshold of the VIF seems to be more understandable due to it is based on the linear relation of one of the independent variables as function of the rest.** Table 2 presents the coefficient of determination for the above regressions, taking into account the number of simulated variables and the nature of the data.

It is observed that the highest coefficient of determination corresponds to the relation as a function of the squared root. **For this reason, the regression (11) and (13) are estimated from the simulated values (see Table 3 and Table 4).** From **these tables**, it is possible to conclude that:

- All coefficients are individually significant with a significance level of 5%. All models are globally significant with the same level of significance.
- By considering the threshold usually established for the VIF to consider high collinearity ($VIF = 10$), it is possible to calculate a equivalent bound for the CN for a determined sample size. **Alternatively, an interested reader could easily substitute by $VIF = 4$).**
- Since in all cases, the coefficient associated with the sample size is negative, the bound equivalent between the VIF and the CN diminishes as the sample size increases.
- Since in all cases, the coefficient associated with the maximum VIF is positive, the CN increases with the VIF.
- Taking into account the bounds and the relation obtained, values of maximum VIF higher than 10 imply values of CN higher than **5** in both proposed transformations. Note that Table 4 shows that the **minimum** CN is **5.034**. Thus, by following the decision rule usually accepted for the VIF, the grade of collinearity

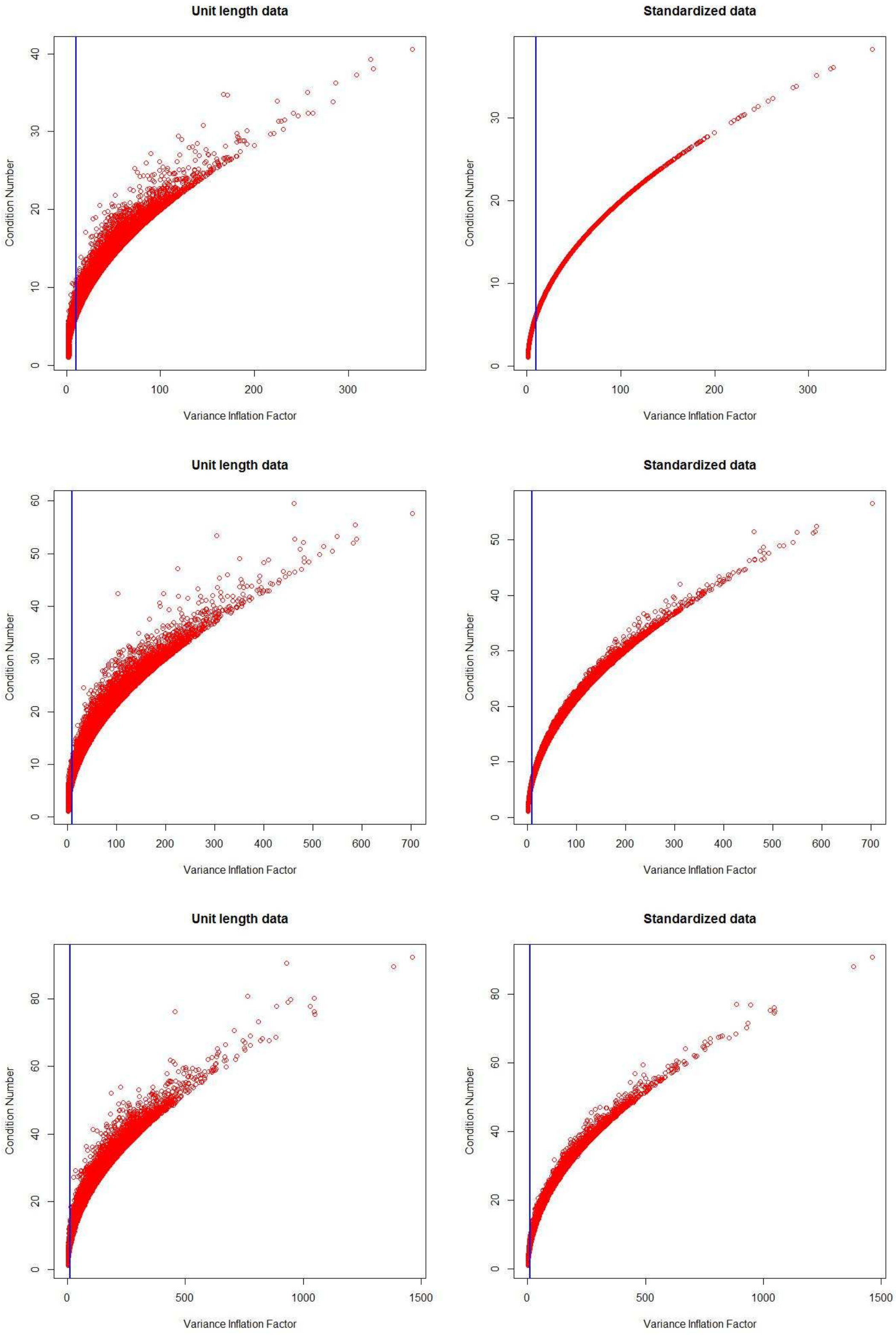

Figure 2. Dispersion diagram of the maximum VIF and the CN for the 3800000 simulated cases: data transformation in columns (unit length and standardized data) and number of variables in rows ($p = 3, 4, 5$). The vertical line $VIF = 10$ is included.

Table 3. Estimation of the regression (11) and the minimum and maximum bounds equivalents for the VIF and CN

p	Data	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$CN_{VIF=10,n=15}$	$CN_{VIF=10,n=200}$
3	Standardized	-0.843	2.186	$-1.881 \cdot 10^{-5}$	6.069457	6.065877
4	Standardized	-0.8097	2.26	$-2.719 \cdot 10^{-4}$	6.332969	6.282668
5	Standardized	-0.9326	2.435	$-5.096 \cdot 10^{-4}$	6.751902	6.665626
3	Unit length	-1.449	2.389	$-8.075 \cdot 10^{-4}$	6.093569	5.944181
4	Unit length	-1.185	2.395	$-1.346 \cdot 10^{-3}$	6.368465	6.119455
5	Unit length	-1.194	2.548	$-1.871 \cdot 10^{-3}$	6.835418	6.489283

Table 4. Estimation of the regression (13) and the minimum and maximum bound equivalents for the VIF and CN

p	Data	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$CN_{VIF=10,n=15}$	$CN_{VIF=10,n=200}$
3	Standardized	1.986	-0.00403	6.2198	5.4742
4	Standardized	2.142	-0.00467	6.7035	5.8395
5	Standardized	2.332	-0.00583	7.2869	6.2084
3	Unit length	2.003	-0.0065	6.2365	5.034
4	Unit length	2.208	-0.0073	6.8728	5.5223
5	Unit length	2.409	-0.00835	7.4926	5.9479

will be worrying when the CN is higher than **5**.

- **Note that Tables 3 and 4 lead to relations similar to those obtained from expression (7)**.

Finally, since regression (13) is most coherent with expression (7), its estimates are presented in Table 4. Tables 5-10 show equivalent values of the CN and VIF depending on the sample size.

6. Conclusions

The existence of an approximate linear relation between the regressors of a econometric model presents a widely studied problem know as collinearity. The measures commonly applied for its detection are the Variance Inflation Factor and the Condition Number. This paper analyzes these measures and their relation and obtains the following conclusions:

a) The decision rule that there exists collinearity for values of VIF higher than 10 **implies that there is** a minimum eigenvalue of $\mathbf{X}^t\mathbf{X}$ less than 0.1 when data are standardized. The Monte Carlo simulations confirm that this relation is also verified when data are expressed in unit length. This would answer the following question raised in [28]: *Kendall (1957) and Silvey (1969) have suggested using the eigenvalues of* $\mathbf{X}^t\mathbf{X}$ *as a key to the presence of collinearity: collinearity is indicated by the presence of a "small" eigenvalue. Unfortunately we are not informed what "small" is.*

b) The **square of** CN is a upper bound of the VIF. This finding is supported because the VIF does not capture the relation between the exogenous variables and the constant term, in contrast to what happens with the CN. It is possible to say that the VIF captures only the relation between the exogenous variables (from a statistical point of view), while the CN is more focused on the ill-conditioning of the matrix $\mathbf{X}$ (from a numerical point of view).

c) The fact that the VIF ignores the relation between the exogenous variables, and the constant term implies that this measure is not able to detect linear dependencies when the variance in the variables is very small. From the CN and using the bounds provided by [6], the Monte Carlo simulations indicate that there is a worrying relation between the exogenous variables and the constant term when the variance in the exogenous variable is less than 0.005.

d) The Monte Carlo simulations suggest that the VIF and the CN are related through a squared root when data are standardized or unit length (if there is no relation with the constant term). From this relation, is possible to establish equivalences between the bounds for the VIF and the CN for different samples sizes. **Thus, the values of the CN are presented from Table 5 to Table 10 for $p = 3, 4$ and 5 with standardized and unit length data considering n varying from 15 to 200 and the maximum VIF from 4 to 200.**

e) **For $p = 2$ is not possible to establish a relation between the VIF and the CN due to the first is always equal to 1. This fact supports the idea presented in conclusion b) about the VIF ignores the relation between the exogenous variables and the constant term.**

References

[1] Besley DA. Conditioning diagnostics: Collinearity and weak data in regression. New York: John Wiley; 1991.

[2] Cook R. Demeaning conditioning diagnostics through centering: Comment. The American Statistician. 1984;38(2):78–79.

[3] Chennamaneni P, Echambadi R, Hess JD, et al. How do you properly diagnose harmful collinearity in moderated regressions? Retrieved June. 2008;1:2011.

[4] Leighton TR. Introductory econometrics: theory and applications ; 1985.

[5] Belsley DA, Kuh E, Welsch RE. Regression diagnostics: Identifying influential data and sources of collinearity. Vol. 571. John Wiley & Sons; 2005.

[6] Belsley DA. Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise. Journal of Econometrics. 1982;20(2):211–253.

[7] Thisted RA. Collinearity and least squares regression: Comment. Statistical Science. 1987;2(1):91–93.

[8] Hadi AS, Wells MT. Assessing the effects of multiple rows on the condition number of a matrix. Journal of the American Statistical Association. 1990;85(411):786–792.

[9] Edelman A. On the distribution of a scaled condition number. Mathematics of computation. 1992;58(197):185–190.

[10] Stewart GW. Collinearity and least squares regression. Statistical Science. 1987; 2(1):68–84.

[11] Wichers CR. The detection of multicollinearity: A comment. The Review of Economics and Statistics. 1975;57(3):366–368.

[12] Haitovsky Y. Multicollinearity in regression analysis: Comment. The Review of economics and statistics. 1969;51(4):486–489.

[13] O'brien RM. A caution regarding rules of thumb for variance inflation factors. Quality & Quantity. 2007;41(5):673–690.

[14] Mason RL, Gunst RF, Hess JL. Statistical design and analysis of experiments: with applications to engineering and science. Vol. 474. John Wiley and Sons; 2003.

[15] Marquardt DW. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics. 1970;12(3):591–612.

[16] Snee RD. Some aspects of nonorthogonal data analysis. Journal of Quality Technology. 1973;5(2):67–79.

[17] Alin A. Multicollinearity. Wiley Interdisciplinary Reviews: Computational Statistics. 2010;2(3):370–374.

[18] Alauddin M, Nghiem HS. Do instructional attributes pose multicollinearity problems?: An empirical exploration. Economic Analysis and Policy. 2010;40(3):351.

[19] Spanos A, McGuirk A. The problem of near-multicollinearity revisited: erratic vs systematic volatility. Journal of Econometrics. 2002;108(2):365–393.

[20] Holland LM. Evaluation of estimators for ill-posed statistical problems subject to multicollinearity [dissertation]. University of Waikato; 2014.

[21] Berk KN. Tolerance and condition in regression computations. Journal of the American Statistical Association. 1977;72(360a):863–866.

[22] García J, Salmerón R, García C, et al. Standardization of variables and collinearity diagnostic in ridge regression. International Statistical Review. 2016; 84(2):245–266.

[23] Salmerón R, García J, García C, et al. Transformation of variables and the condition number in ridge estimation. Computational Statistics. 2016;:1–28.

[24] Belsley DA. A guide to using the collinearity diagnostics. Computer Science in Economics and Management. 1991;4(1):33–50.

[25] McDonald GC, Galarneau DI. A monte carlo evaluation of some ridge-type estimators. Journal of the American Statistical Association. 1975;70(350):407–416.

[26] Gibbons DG. A simulation study of some ridge estimators. Journal of the American Statistical Association. 1981;76(373):131–139.

[27] Salmerón Gómez R, García Pérez J, López Martín MDM, et al. Collinearity diagnostic applied in ridge estimation through the variance inflation factor. Journal of Applied Statistics. 2016;43(10):1831–1849.

[28] Belsley DA, Kuh E, Welsch RE. Regression diagnostics: Identifying influential data and sources of collinearity. Vol. 571. John Wiley & Sons; 2005.

Table 5. CN for standardized data with $p = 3$ and obtained from $\widehat{CN} = 1.986 \cdot \sqrt{\max VIF} - 0.00403 \cdot n$.

	Max VIF																						
n	4	5	6	7	8	9	10	15	20	25	30	40	50	60	70	80	90	100	120	140	160	180	200
15	3.91	4.38	4.8	5.19	5.56	5.9	6.22	7.63	8.82	9.87	10.8	12.5	14	15.3	16.6	17.7	18.8	19.8	21.7	23.4	25.1	26.6	28
25	3.87	4.34	4.76	5.15	5.52	5.86	6.18	7.59	8.78	9.83	10.8	12.5	13.9	15.3	16.5	17.7	18.7	19.8	21.7	23.4	25	26.5	28
35	3.83	4.3	4.72	5.11	5.48	5.82	6.14	7.55	8.74	9.79	10.7	12.4	13.9	15.2	16.5	17.6	18.7	19.7	21.6	23.4	25	26.5	27.9
45	3.79	4.26	4.68	5.07	5.44	5.78	6.1	7.51	8.7	9.75	10.7	12.4	13.9	15.2	16.4	17.6	18.7	19.7	21.6	23.3	24.9	26.5	27.9
55	3.75	4.22	4.64	5.03	5.4	5.74	6.06	7.47	8.66	9.71	10.7	12.3	13.8	15.2	16.4	17.5	18.6	19.6	21.5	23.3	24.9	26.4	27.9
65	3.71	4.18	4.6	4.99	5.36	5.7	6.02	7.43	8.62	9.67	10.6	12.3	13.8	15.1	16.4	17.5	18.6	19.6	21.5	23.2	24.9	26.4	27.8
75	3.67	4.14	4.56	4.95	5.32	5.66	5.98	7.39	8.58	9.63	10.6	12.3	13.7	15.1	16.3	17.5	18.5	19.6	21.5	23.2	24.8	26.3	27.8
85	3.63	4.1	4.52	4.91	5.27	5.62	5.94	7.35	8.54	9.59	10.5	12.2	13.7	15	16.3	17.4	18.5	19.5	21.4	23.2	24.8	26.3	27.7
95	3.59	4.06	4.48	4.87	5.23	5.58	5.9	7.31	8.5	9.55	10.5	12.2	13.7	15	16.2	17.4	18.5	19.5	21.4	23.1	24.7	26.3	27.7
105	3.55	4.02	4.44	4.83	5.19	5.53	5.86	7.27	8.46	9.51	10.5	12.1	13.6	15	16.2	17.3	18.4	19.4	21.3	23.1	24.7	26.2	27.7
115	3.51	3.98	4.4	4.79	5.15	5.49	5.82	7.23	8.42	9.47	10.4	12.1	13.6	14.9	16.2	17.3	18.4	19.4	21.3	23	24.7	26.2	27.6
125	3.47	3.94	4.36	4.75	5.11	5.45	5.78	7.19	8.38	9.43	10.4	12.1	13.5	14.9	16.1	17.3	18.3	19.4	21.3	23	24.6	26.1	27.6
135	3.43	3.9	4.32	4.71	5.07	5.41	5.74	7.15	8.34	9.39	10.3	12	13.5	14.8	16.1	17.2	18.3	19.3	21.2	23	24.6	26.1	27.5
145	3.39	3.86	4.28	4.67	5.03	5.37	5.7	7.11	8.3	9.35	10.3	12	13.5	14.8	16	17.2	18.3	19.3	21.2	22.9	24.5	26.1	27.5
155	3.35	3.82	4.24	4.63	4.99	5.33	5.66	7.07	8.26	9.31	10.3	11.9	13.4	14.8	16	17.1	18.2	19.2	21.1	22.9	24.5	26	27.5
165	3.31	3.78	4.2	4.59	4.95	5.29	5.62	7.03	8.22	9.27	10.2	11.9	13.4	14.7	16	17.1	18.2	19.2	21.1	22.8	24.5	26	27.4
175	3.27	3.74	4.16	4.55	4.91	5.25	5.58	6.99	8.18	9.22	10.2	11.9	13.3	14.7	15.9	17.1	18.1	19.2	21.1	22.8	24.4	25.9	27.4
185	3.23	3.7	4.12	4.51	4.87	5.21	5.53	6.95	8.14	9.18	10.1	11.8	13.3	14.6	15.9	17	18.1	19.1	21	22.8	24.4	25.9	27.3
195	3.19	3.65	4.08	4.47	4.83	5.17	5.49	6.91	8.1	9.14	10.1	11.8	13.3	14.6	15.8	17	18.1	19.1	21	22.7	24.3	25.9	27.3
200	3.17	3.63	4.06	4.45	4.81	5.15	5.47	6.89	8.08	9.12	10.1	11.8	13.2	14.6	15.8	17	18	19.1	20.9	22.7	24.3	25.8	27.3

Table 6. CN for standardized data with $p = 4$ and obtained from $\widehat{CN} = 2.142 \cdot \sqrt{\max VIF} - 0.00467 \cdot n$.

	Max VIF																						
n	4	5	6	7	8	9	10	15	20	25	30	40	50	60	70	80	90	100	120	140	160	180	200
15	4.21	4.72	5.18	5.6	5.99	6.36	6.7	8.23	9.51	10.6	11.7	13.5	15.1	16.5	17.9	19.1	20.3	21.3	23.4	25.3	27	28.7	30.2
25	4.17	4.67	5.13	5.55	5.94	6.31	6.66	8.18	9.46	10.6	11.6	13.4	15	16.5	17.8	19	20.2	21.3	23.3	25.2	27	28.6	30.2
35	4.12	4.63	5.08	5.5	5.9	6.26	6.61	8.13	9.42	10.5	11.6	13.4	15	16.4	17.8	19	20.2	21.3	23.3	25.2	26.9	28.6	30.1
45	4.07	4.58	5.04	5.46	5.85	6.22	6.56	8.09	9.37	10.5	11.5	13.3	14.9	16.4	17.7	18.9	20.1	21.2	23.3	25.1	26.9	28.5	30.1
55	4.03	4.53	4.99	5.41	5.8	6.17	6.52	8.04	9.32	10.5	11.5	13.3	14.9	16.3	17.7	18.9	20.1	21.2	23.2	25.1	26.8	28.5	30
65	3.98	4.49	4.94	5.36	5.75	6.12	6.47	7.99	9.28	10.4	11.4	13.2	14.8	16.3	17.6	18.9	20	21.1	23.2	25	26.8	28.4	30
75	3.93	4.44	4.9	5.32	5.71	6.08	6.42	7.95	9.23	10.4	11.4	13.2	14.8	16.2	17.6	18.8	20	21.1	23.1	25	26.7	28.4	29.9
85	3.89	4.39	4.85	5.27	5.66	6.03	6.38	7.9	9.18	10.3	11.3	13.2	14.7	16.2	17.5	18.8	19.9	21	23.1	24.9	26.7	28.3	29.9
95	3.84	4.35	4.8	5.22	5.61	5.98	6.33	7.85	9.14	10.3	11.3	13.1	14.7	16.1	17.5	18.7	19.9	21	23	24.9	26.7	28.3	29.8
105	3.79	4.3	4.76	5.18	5.57	5.94	6.28	7.81	9.09	10.2	11.2	13.1	14.7	16.1	17.4	18.7	19.8	20.9	23	24.9	26.6	28.2	29.8
115	3.75	4.25	4.71	5.13	5.52	5.89	6.24	7.76	9.04	10.2	11.2	13	14.6	16.1	17.4	18.6	19.8	20.9	22.9	24.8	26.6	28.2	29.8
125	3.7	4.21	4.66	5.08	5.47	5.84	6.19	7.71	9	10.1	11.1	13	14.6	16	17.3	18.6	19.7	20.8	22.9	24.8	26.5	28.2	29.7
135	3.65	4.16	4.62	5.04	5.43	5.8	6.14	7.67	8.95	10.1	11.1	12.9	14.5	16	17.3	18.5	19.7	20.8	22.8	24.7	26.5	28.1	29.7
145	3.61	4.11	4.57	4.99	5.38	5.75	6.1	7.62	8.9	10	11.1	12.9	14.5	15.9	17.2	18.5	19.6	20.7	22.8	24.7	26.4	28.1	29.6
155	3.56	4.07	4.52	4.94	5.33	5.7	6.05	7.57	8.86	9.99	11	12.8	14.4	15.9	17.2	18.4	19.6	20.7	22.7	24.6	26.4	28	29.6
165	3.51	4.02	4.48	4.9	5.29	5.66	6	7.53	8.81	9.94	11	12.8	14.4	15.8	17.2	18.4	19.6	20.6	22.7	24.6	26.3	28	29.5
175	3.47	3.97	4.43	4.85	5.24	5.61	5.96	7.48	8.76	9.89	10.9	12.7	14.3	15.8	17.1	18.3	19.5	20.6	22.6	24.5	26.3	27.9	29.5
185	3.42	3.93	4.38	4.8	5.19	5.56	5.91	7.43	8.72	9.85	10.9	12.7	14.3	15.7	17.1	18.3	19.5	20.6	22.6	24.5	26.2	27.9	29.4
195	3.37	3.88	4.34	4.76	5.15	5.52	5.86	7.39	8.67	9.8	10.8	12.6	14.2	15.7	17	18.2	19.4	20.5	22.6	24.4	26.2	27.8	29.4
200	3.35	3.86	4.31	4.73	5.12	5.49	5.84	7.36	8.65	9.78	10.8	12.6	14.2	15.7	17	18.2	19.4	20.5	22.5	24.4	26.2	27.8	29.4

Table 7. CN for standardized data with $p = 5$ and obtained from $\widehat{CN} = 2.332 \cdot \sqrt{\max VIF} - 0.00583 \cdot n$.

	Max VIF																						
n	4	5	6	7	8	9	10	15	20	25	30	40	50	60	70	80	90	100	120	140	160	180	200
15	4.58	5.13	5.62	6.08	6.51	6.91	7.29	8.94	10.3	11.6	12.7	14.7	16.4	18	19.4	20.8	22	23.2	25.5	27.5	29.4	31.2	32.9
25	4.52	5.07	5.57	6.02	6.45	6.85	7.23	8.89	10.3	11.5	12.6	14.6	16.3	17.9	19.4	20.7	22	23.2	25.4	27.4	29.4	31.1	32.8
35	4.46	5.01	5.51	5.97	6.39	6.79	7.17	8.83	10.2	11.5	12.6	14.5	16.3	17.9	19.3	20.7	21.9	23.1	25.3	27.4	29.3	31.1	32.8
45	4.4	4.95	5.45	5.91	6.33	6.73	7.11	8.77	10.2	11.4	12.5	14.5	16.2	17.8	19.2	20.6	21.9	23.1	25.3	27.3	29.2	31	32.7
55	4.34	4.89	5.39	5.85	6.28	6.68	7.05	8.71	10.1	11.3	12.5	14.4	16.2	17.7	19.2	20.5	21.8	23	25.2	27.3	29.2	31	32.7
65	4.29	4.84	5.33	5.79	6.22	6.62	7	8.65	10.1	11.3	12.4	14.4	16.1	17.7	19.1	20.5	21.7	22.9	25.2	27.2	29.1	30.9	32.6
75	4.23	4.78	5.27	5.73	6.16	6.56	6.94	8.59	9.99	11.2	12.3	14.3	16.1	17.6	19.1	20.4	21.7	22.9	25.1	27.2	29.1	30.8	32.5
85	4.17	4.72	5.22	5.67	6.1	6.5	6.88	8.54	9.93	11.2	12.3	14.3	16	17.6	19	20.4	21.6	22.8	25.1	27.1	29	30.8	32.5
95	4.11	4.66	5.16	5.62	6.04	6.44	6.82	8.48	9.88	11.1	12.2	14.2	15.9	17.5	19	20.3	21.6	22.8	25	27	28.9	30.7	32.4
105	4.05	4.6	5.1	5.56	5.98	6.38	6.76	8.42	9.82	11	12.2	14.1	15.9	17.5	18.9	20.2	21.5	22.7	24.9	27	28.9	30.7	32.4
115	3.99	4.54	5.04	5.5	5.93	6.33	6.7	8.36	9.76	11	12.1	14.1	15.8	17.4	18.8	20.2	21.5	22.6	24.9	26.9	28.8	30.6	32.3
125	3.94	4.49	4.98	5.44	5.87	6.27	6.65	8.3	9.7	10.9	12	14	15.8	17.3	18.8	20.1	21.4	22.6	24.8	26.9	28.8	30.6	32.3
135	3.88	4.43	4.93	5.38	5.81	6.21	6.59	8.24	9.64	10.9	12	14	15.7	17.3	18.7	20.1	21.3	22.5	24.8	26.8	28.7	30.5	32.2
145	3.82	4.37	4.87	5.32	5.75	6.15	6.53	8.19	9.58	10.8	11.9	13.9	15.6	17.2	18.7	20	21.3	22.5	24.7	26.7	28.7	30.4	32.1
155	3.76	4.31	4.81	5.27	5.69	6.09	6.47	8.13	9.53	10.8	11.9	13.8	15.6	17.2	18.6	20	21.2	22.4	24.6	26.7	28.6	30.4	32.1
165	3.7	4.25	4.75	5.21	5.63	6.03	6.41	8.07	9.47	10.7	11.8	13.8	15.5	17.1	18.5	19.9	21.2	22.4	24.6	26.6	28.5	30.3	32
175	3.64	4.19	4.69	5.15	5.58	5.98	6.35	8.01	9.41	10.6	11.8	13.7	15.5	17	18.5	19.8	21.1	22.3	24.5	26.6	28.5	30.3	32
185	3.59	4.14	4.63	5.09	5.52	5.92	6.3	7.95	9.35	10.6	11.7	13.7	15.4	17	18.4	19.8	21	22.2	24.5	26.5	28.4	30.2	31.9
195	3.53	4.08	4.58	5.03	5.46	5.86	6.24	7.89	9.29	10.5	11.6	13.6	15.4	16.9	18.4	19.7	21	22.2	24.4	26.5	28.4	30.2	31.8
200	3.5	4.05	4.55	5	5.43	5.83	6.21	7.87	9.26	10.5	11.6	13.6	15.3	16.9	18.3	19.7	21	22.2	24.4	26.4	28.3	30.1	31.8

Table 8. CN for unit length data with $p = 3$ and obtained from $\widehat{CN} = 2.003 \cdot \sqrt{\max VIF} - 0.0065 \cdot n$.

	Max VIF																					
n	4	5	6	7	8	9	10	15	20	25	30	40	50	60	70	80	90	100	120	140	160	180	200
15	3.91	4.38	4.81	5.2	5.57	5.91	6.24	7.66	8.86	9.92	10.9	12.6	14.1	15.4	16.7	17.8	18.9	19.9	21.8	23.6	25.2	26.8	28.2
25	3.84	4.32	4.74	5.14	5.5	5.85	6.17	7.6	8.8	9.85	10.8	12.5	14	15.4	16.6	17.8	18.8	19.9	21.8	23.5	25.2	26.7	28.2
35	3.78	4.25	4.68	5.07	5.44	5.78	6.11	7.53	8.73	9.79	10.7	12.4	13.9	15.3	16.5	17.7	18.8	19.8	21.7	23.5	25.1	26.6	28.1
45	3.71	4.19	4.61	5.01	5.37	5.72	6.04	7.47	8.67	9.72	10.7	12.4	13.9	15.2	16.5	17.6	18.7	19.7	21.6	23.4	25	26.6	28
55	3.65	4.12	4.55	4.94	5.31	5.65	5.98	7.4	8.6	9.66	10.6	12.3	13.8	15.2	16.4	17.6	18.6	19.7	21.6	23.3	25	26.5	28
65	3.58	4.06	4.48	4.88	5.24	5.59	5.91	7.34	8.54	9.59	10.5	12.2	13.7	15.1	16.3	17.5	18.6	19.6	21.5	23.3	24.9	26.5	27.9
75	3.52	3.99	4.42	4.81	5.18	5.52	5.85	7.27	8.47	9.53	10.5	12.2	13.7	15	16.3	17.4	18.5	19.5	21.5	23.2	24.8	26.4	27.8
85	3.45	3.93	4.35	4.75	5.11	5.46	5.78	7.21	8.41	9.46	10.4	12.1	13.6	15	16.2	17.4	18.4	19.5	21.4	23.1	24.8	26.3	27.8
95	3.39	3.86	4.29	4.68	5.05	5.39	5.72	7.14	8.34	9.4	10.4	12.1	13.5	14.9	16.1	17.3	18.4	19.4	21.3	23.1	24.7	26.3	27.7
105	3.32	3.8	4.22	4.62	4.98	5.33	5.65	7.08	8.28	9.33	10.3	12	13.5	14.8	16.1	17.2	18.3	19.3	21.3	23	24.7	26.2	27.6
115	3.26	3.73	4.16	4.55	4.92	5.26	5.59	7.01	8.21	9.27	10.2	11.9	13.4	14.8	16	17.2	18.3	19.3	21.2	23	24.6	26.1	27.6
125	3.19	3.67	4.09	4.49	4.85	5.2	5.52	6.95	8.15	9.2	10.2	11.9	13.4	14.7	15.9	17.1	18.2	19.2	21.1	22.9	24.5	26.1	27.5
135	3.13	3.6	4.03	4.42	4.79	5.13	5.46	6.88	8.08	9.14	10.1	11.8	13.3	14.6	15.9	17	18.1	19.2	21.1	22.8	24.5	26	27.4
145	3.06	3.54	3.96	4.36	4.72	5.07	5.39	6.82	8.02	9.07	10	11.7	13.2	14.6	15.8	17	18.1	19.1	21	22.8	24.4	25.9	27.4
155	3	3.47	3.9	4.29	4.66	5	5.33	6.75	7.95	9.01	9.96	11.7	13.2	14.5	15.8	16.9	18	19	20.9	22.7	24.3	25.9	27.3
165	2.93	3.41	3.83	4.23	4.59	4.94	5.26	6.69	7.89	8.94	9.9	11.6	13.1	14.4	15.7	16.8	17.9	19	20.9	22.6	24.3	25.8	27.3
175	2.87	3.34	3.77	4.16	4.53	4.87	5.2	6.62	7.82	8.88	9.83	11.5	13	14.4	15.6	16.8	17.9	18.9	20.8	22.6	24.2	25.7	27.2
185	2.8	3.28	3.7	4.1	4.46	4.81	5.13	6.56	7.76	8.81	9.77	11.5	13	14.3	15.6	16.7	17.8	18.8	20.7	22.5	24.1	25.7	27.1
195	2.74	3.21	3.64	4.03	4.4	4.74	5.07	6.49	7.69	8.75	9.7	11.4	12.9	14.2	15.5	16.6	17.7	18.8	20.7	22.4	24.1	25.6	27.1
200	2.71	3.18	3.61	4	4.37	4.71	5.03	6.46	7.66	8.72	9.67	11.4	12.9	14.2	15.5	16.6	17.7	18.7	20.6	22.4	24	25.6	27

Table 9. CN for Unit length data with $p = 4$ and obtained from $\widehat{CN} = 2.208 \cdot \sqrt{\max VIF} - 0.0073 \cdot n$.

	Max VIF																						
n	4	5	6	7	8	9	10	15	20	25	30	40	50	60	70	80	90	100	120	140	160	180	200
15	4.31	4.83	5.3	5.73	6.14	6.51	6.87	8.44	9.76	10.9	12	13.9	15.5	17	18.4	19.6	20.8	22	24.1	26	27.8	29.5	31.1
25	4.23	4.75	5.23	5.66	6.06	6.44	6.8	8.37	9.69	10.9	11.9	13.8	15.4	16.9	18.3	19.6	20.8	21.9	24	25.9	27.7	29.4	31
35	4.16	4.68	5.15	5.59	5.99	6.37	6.73	8.3	9.62	10.8	11.8	13.7	15.4	16.8	18.2	19.5	20.7	21.8	23.9	25.9	27.7	29.4	31
45	4.09	4.61	5.08	5.51	5.92	6.3	6.65	8.22	9.55	10.7	11.8	13.6	15.3	16.8	18.1	19.4	20.6	21.8	23.9	25.8	27.6	29.3	30.9
55	4.01	4.54	5.01	5.44	5.84	6.22	6.58	8.15	9.47	10.6	11.7	13.6	15.2	16.7	18.1	19.3	20.5	21.7	23.8	25.7	27.5	29.2	30.8
65	3.94	4.46	4.93	5.37	5.77	6.15	6.51	8.08	9.4	10.6	11.6	13.5	15.1	16.6	18	19.3	20.5	21.6	23.7	25.7	27.5	29.1	30.8
75	3.87	4.39	4.86	5.29	5.7	6.08	6.43	8	9.33	10.5	11.5	13.4	15.1	16.6	17.9	19.2	20.4	21.5	23.6	25.6	27.4	29.1	30.7
85	3.8	4.32	4.79	5.22	5.62	6	6.36	7.93	9.25	10.4	11.5	13.3	15	16.5	17.9	19.1	20.3	21.5	23.6	25.5	27.3	29	30.6
95	3.72	4.24	4.71	5.15	5.55	5.93	6.29	7.86	9.18	10.3	11.4	13.3	14.9	16.4	17.8	19.1	20.3	21.4	23.5	25.4	27.2	28.9	30.5
105	3.65	4.17	4.64	5.08	5.48	5.86	6.22	7.79	9.11	10.3	11.3	13.2	14.8	16.3	17.7	19	20.2	21.3	23.4	25.4	27.2	28.9	30.5
115	3.58	4.1	4.57	5	5.41	5.78	6.14	7.71	9.03	10.2	11.3	13.1	14.8	16.3	17.6	18.9	20.1	21.2	23.3	25.3	27.1	28.8	30.4
125	3.5	4.02	4.5	4.93	5.33	5.71	6.07	7.64	8.96	10.1	11.2	13.1	14.7	16.2	17.6	18.8	20	21.2	23.3	25.2	27	28.7	30.3
135	3.43	3.95	4.42	4.86	5.26	5.64	6	7.57	8.89	10.1	11.1	13	14.6	16.1	17.5	18.8	20	21.1	23.2	25.1	26.9	28.6	30.2
145	3.36	3.88	4.35	4.78	5.19	5.57	5.92	7.49	8.82	9.98	11	12.9	14.6	16	17.4	18.7	19.9	21	23.1	25.1	26.9	28.6	30.2
155	3.28	3.81	4.28	4.71	5.11	5.49	5.85	7.42	8.74	9.91	11	12.8	14.5	16	17.3	18.6	19.8	20.9	23.1	25	26.8	28.5	30.1
165	3.21	3.73	4.2	4.64	5.04	5.42	5.78	7.35	8.67	9.84	10.9	12.8	14.4	15.9	17.3	18.5	19.7	20.9	23	24.9	26.7	28.4	30
175	3.14	3.66	4.13	4.56	4.97	5.35	5.7	7.27	8.6	9.76	10.8	12.7	14.3	15.8	17.2	18.5	19.7	20.8	22.9	24.8	26.7	28.3	29.9
185	3.07	3.59	4.06	4.49	4.89	5.27	5.63	7.2	8.52	9.69	10.7	12.6	14.3	15.8	17.1	18.4	19.6	20.7	22.8	24.8	26.6	28.3	29.9
195	2.99	3.51	3.98	4.42	4.82	5.2	5.56	7.13	8.45	9.62	10.7	12.5	14.2	15.7	17	18.3	19.5	20.7	22.8	24.7	26.5	28.2	29.8
200	2.96	3.48	3.95	4.38	4.79	5.16	5.52	7.09	8.41	9.58	10.6	12.5	14.2	15.6	17	18.3	19.5	20.6	22.7	24.7	26.5	28.2	29.8

Table 10. CN for unit length data with $p = 5$ and obtained from $\widehat{CN} = 2.409 \cdot \sqrt{\max VIF} - 0.00835 \cdot n$.

	Max VIF																						
n	4	5	6	7	8	9	10	15	20	25	30	40	50	60	70	80	90	100	120	140	160	180	200
15	4.69	5.26	5.78	6.25	6.69	7.1	7.49	9.2	10.6	11.9	13.1	15.1	16.9	18.5	20	21.4	22.7	24	26.3	28.4	30.3	32.2	33.9
25	4.61	5.18	5.69	6.16	6.6	7.02	7.41	9.12	10.6	11.8	13	15	16.8	18.5	19.9	21.3	22.6	23.9	26.2	28.3	30.3	32.1	33.9
35	4.53	5.09	5.61	6.08	6.52	6.93	7.33	9.04	10.5	11.8	12.9	14.9	16.7	18.4	19.9	21.3	22.6	23.8	26.1	28.2	30.2	32	33.8
45	4.44	5.01	5.53	6	6.44	6.85	7.24	8.95	10.4	11.7	12.8	14.9	16.7	18.3	19.8	21.2	22.5	23.7	26	28.1	30.1	31.9	33.7
55	4.36	4.93	5.44	5.91	6.35	6.77	7.16	8.87	10.3	11.6	12.7	14.8	16.6	18.2	19.7	21.1	22.4	23.6	25.9	28	30	31.9	33.6
65	4.28	4.84	5.36	5.83	6.27	6.68	7.08	8.79	10.2	11.5	12.7	14.7	16.5	18.1	19.6	21	22.3	23.5	25.8	28	29.9	31.8	33.5
75	4.19	4.76	5.27	5.75	6.19	6.6	6.99	8.7	10.1	11.4	12.6	14.6	16.4	18	19.5	20.9	22.2	23.5	25.8	27.9	29.8	31.7	33.4
85	4.11	4.68	5.19	5.66	6.1	6.52	6.91	8.62	10.1	11.3	12.5	14.5	16.3	18	19.4	20.8	22.1	23.4	25.7	27.8	29.8	31.6	33.4
95	4.02	4.59	5.11	5.58	6.02	6.43	6.82	8.54	9.98	11.3	12.4	14.4	16.2	17.9	19.4	20.8	22.1	23.3	25.6	27.7	29.7	31.5	33.3
105	3.94	4.51	5.02	5.5	5.94	6.35	6.74	8.45	9.9	11.2	12.3	14.4	16.2	17.8	19.3	20.7	22	23.2	25.5	27.6	29.6	31.4	33.2
115	3.86	4.43	4.94	5.41	5.85	6.27	6.66	8.37	9.81	11.1	12.2	14.3	16.1	17.7	19.2	20.6	21.9	23.1	25.4	27.5	29.5	31.4	33.1
125	3.77	4.34	4.86	5.33	5.77	6.18	6.57	8.29	9.73	11	12.2	14.2	16	17.6	19.1	20.5	21.8	23	25.3	27.5	29.4	31.3	33
135	3.69	4.26	4.77	5.25	5.69	6.1	6.49	8.2	9.65	10.9	12.1	14.1	15.9	17.5	19	20.4	21.7	23	25.3	27.4	29.3	31.2	32.9
145	3.61	4.18	4.69	5.16	5.6	6.02	6.41	8.12	9.56	10.8	12	14	15.8	17.4	18.9	20.3	21.6	22.9	25.2	27.3	29.3	31.1	32.9
155	3.52	4.09	4.61	5.08	5.52	5.93	6.32	8.04	9.48	10.8	11.9	13.9	15.7	17.4	18.9	20.3	21.6	22.8	25.1	27.2	29.2	31	32.8
165	3.44	4.01	4.52	5	5.44	5.85	6.24	7.95	9.4	10.7	11.8	13.9	15.7	17.3	18.8	20.2	21.5	22.7	25	27.1	29.1	30.9	32.7
175	3.36	3.93	4.44	4.91	5.35	5.77	6.16	7.87	9.31	10.6	11.7	13.8	15.6	17.2	18.7	20.1	21.4	22.6	24.9	27	29	30.9	32.6
185	3.27	3.84	4.36	4.83	5.27	5.68	6.07	7.79	9.23	10.5	11.6	13.7	15.5	17.1	18.6	20	21.3	22.5	24.8	27	28.9	30.8	32.5
195	3.19	3.76	4.27	4.75	5.19	5.6	5.99	7.7	9.15	10.4	11.6	13.6	15.4	17	18.5	19.9	21.2	22.5	24.8	26.9	28.8	30.7	32.4
200	3.15	3.72	4.23	4.7	5.14	5.56	5.95	7.66	9.1	10.4	11.5	13.6	15.4	17	18.5	19.9	21.2	22.4	24.7	26.8	28.8	30.7	32.4