Journal of Statistics: Advances in Theory and Applications
Volume 19, Number 1, 2018, Pages 17-30
Available at http://scientificadvances.co.in
DOI: http://dx.doi.org/10.18642/jsata_7100121917

APPLICATION OF MICROSOFT EXCEL IN UNDERSTANDING THE DEGREES OF FREEDOM OF LINEAR REGRESSION

JIAN-HUA DING[1], LING-YING SHUAI[1], JIE-BO XUE[2] and JUN LI[1]

[1]School of Life Sciences
Huaibei Normal University
Huaibei 235000
P. R. China
e-mail: healthlicn@chnu.edu.cn

[2]School of Accountancy
Fujian Jiangxia University
Fuzhou 350108
P. R. China

Abstract

In this article, the degrees of freedom of the residual variance of simple linear regression are simulated by using the VBA (Visual Basic for Application) of Microsoft Excel 2010. The simulation file dynamically demonstrates why the residual variance should be calculated by dividing the sum of squared errors by $n-2$ rather than n, which can be displayed in class and is helpful for students to grasp the meaning of degrees of freedom.

2010 Mathematics Subject Classification: 62J05.
Keywords and phrases: degrees of freedom, linear regression, excel simulation.
Received January 7, 2018; Revised February 2, 2018

1. Introduction

Although most of the statistical tests encountered during a course on inferential statistics depend on degrees of freedom, many introductory textbooks present the concept in a strictly formulaic manner, often without further explanation of why the formula is given [1]. Furthermore, some of the definitions offered in the literature are inconsistent with one another. For example, Clapham [2] states that the number of degrees of freedom is a positive integer, in contrast, Kotz and Johnson [3] point out that the number of degrees of freedom may be fractional in some approximations. After summarizing many views, Eisenhauer [1] think of degrees of freedom as the number of pieces of information that can be freely varied without violating any given restrictions.

Take simple linear regression as an example. A fit linear regression equation is obtained according to n pairs of measurements (x_i, Y_i) by applying the least square method:

$$Y_{\text{fit}}(x) = a + bx. \tag{1}$$

In order to test if the regression has a significance, we should analyze the components of total sum of squares and total degrees of freedom of response variable Y, then we have SST = SSR + SSE, as in formula

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_{\text{fit}}(x_i) - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - Y_{\text{fit}}(x_i))^2, \tag{2}$$

and have the corresponding degrees of freedom $df_T = df_R + df_E$, i.e., $n - 1 = 1 + (n - 2)$. Why does the SSE have $n - 2$ degrees of freedom? The answer is that a restriction $Y_{\text{fit}}(x_i)$ is used to calculate the sum of squared errors here. According to Equation (1), two response variables cannot be freely assigned due to the fact that a straight line is specified by two points.

In mathematics, it can be justified easily that residual variance of a sample obtained by dividing the residual sum of square by $n-2$ rather than n is just the unbiased estimator of the residual variance of the population. However, this is difficult to be understood by most students [1]. With the popularization of multimedia teaching technology, more and more people use computer simulation methods to help students understand some difficult or abstract concepts in statistics, such as the Sampling Distributions [4], Central Limit Theorem [5], Law of Large Numbers [6], etc. And many simulation processes were compiled into applets on the Internet by using the Java language for students' exercises, such as Rossman/Chance Applet Collection (http://rossmanchance.com/applets/index.html), Rice Virtual Lab in Statistics (http://onlinestatbook.com/stat_sim/index.html), and so on. But, so far, we have not found an applet for simulating degrees of freedom. Sullivan [7] once tried to understand the $n-2$ degrees of freedom by a short computer program written in BASIC. But his program is not suitable to demonstrate in class. Microsoft Excel, as a commonly used software, has a wide range of applications in multimedia teaching [8-11]. In this article, we simulated the $n-2$ degrees of freedom of simple linear regression by using Excel 2010, which can be easily demonstrated in class and thus overcome the shortcoming of Sullivan's program.

2. The Principle of Simulation

It is postulated that in a population of double variables, variable Y is related to variable x by the mathematical model

$$Y_{\text{true}}(x) = \alpha + \beta x. \tag{3}$$

However, this model does not specify the numerical values of α and β. Measurements of (x, Y) pairs are made to estimate the unknown parameters α and β by applying the least square method. It is assumed that x is the predictor variable with negligible error, while Y is the response variable with random error, so we have

$$Y_{\text{meas}}(x) = \alpha + \beta x + \varepsilon, \tag{4}$$

where ε is the residual of population from the Normal distribution $N(0, \sigma^2)$. If n pairs of measurements are randomly extracted from the population as a sample at a time, we can obtain the fit regression equation of the sample, i.e., Equation (1). Then we have

$$Y_{\text{meas}}(x) = a + bx + e, \tag{5}$$

where e is the residual of sample and the estimator of ε. The $Y_{\text{true}}(x)$, $Y_{\text{meas}}(x)$, $Y_{\text{fit}}(x)$, ε and e are shown in Figure 1.

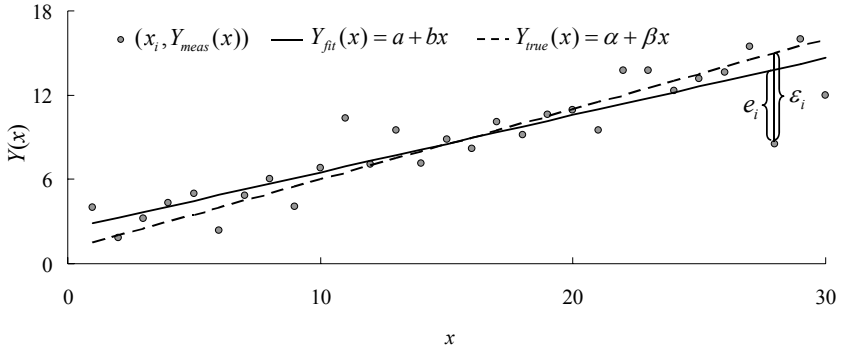


Figure 1. The relationships between the $Y_{\text{true}}(x)$, $Y_{\text{meas}}(x)$, $Y_{\text{fit}}(x)$, ε and e.

It is assumed that α equals 1, β equals 0.5, the sample size equals n, and the independent variable x equals 1, 2, 3, …, n, in turn, then we can obtain n values of $Y_{\text{true}}(x)$ according to Equation (3). The function formula NormInv(Rnd( ), 0, σ) in Excel can generate a set of random numbers from the Normal distribution $N(0, \sigma^2)$, which can be used to generate n random numbers to simulate the residual ε. Then n values of $Y_{\text{meas}}(x)$ are produced according to Equation (4). Those n pairs of measurements $(x, Y_{\text{meas}}(x))$ are fitted to generate Equation (1). Then we have a sum of squared deviations between measured and true values

$$\text{SMT} = \sum_{i=1}^{n} [Y_{\text{meas}}(x) - Y_{\text{true}}(x)]^2. \tag{6}$$

According to Equation (3) and (4), we have

$$\mathrm{SMT} = \sum_{i=1}^{n} \varepsilon_i^2. \tag{7}$$

Similarly, we have a sum of squared deviations between measured and fitted values

$$\mathrm{SMF} = \sum_{i=1}^{n} [Y_{\mathrm{meas}}(x) - Y_{\mathrm{fit}}(x)]^2. \tag{8}$$

According to Equation (1) and (5), we have

$$\mathrm{SMF} = \sum_{i=1}^{n} e_i^2. \tag{9}$$

Repeating the extracting process T times will generate T samples, as well as T values of SMT and T values of SMF. As a result, the average of T values of SMT is

$$\mathrm{SMT}_{\mathrm{mean}} = (\sum_{j=1}^{T} \sum_{i=1}^{n} \varepsilon_i^2)/T, \tag{10}$$

and the average of T values of SMF is

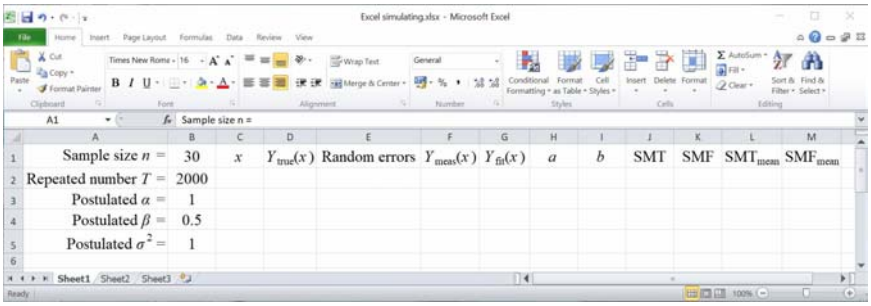
$$\mathrm{SMF}_{\mathrm{mean}} = (\sum_{j=1}^{T} \sum_{i=1}^{n} e_i^2) / T. \tag{11}$$

As T becomes large enough, both $\mathrm{SMT}_{\mathrm{mean}}$ and $\mathrm{SMF}_{\mathrm{mean}}$ will approach their mathematical expectations, i.e., $n\sigma^2$ and $(n-2)\sigma^2$ in theory, respectively.

3. The Process of Simulation

The first step: Open Excel 2010 to create a new file. In cells from A1 to A5 of Sheet1, input "Sample size n =", "Repeated number T =", "Postulated α =", "Postulated β =", "Postulated σ^2 =" in turn. Then

input the corresponding values of the five above-mentioned items in cells from B1 to B5 (in this article, take 30, 2000, 1, 0.5, 1 as samples in turn). At last, input "x", "$Y_{\text{true}}(x)$", "Random errors", "$Y_{\text{meas}}(x)$", "$Y_{\text{fit}}(x)$", "a", "b", "SMT", "SMF", "SMT_{mean}", and "SMF_{mean}" in cells from C1 to M1 (Figure 2).



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Sample size n =	30	x	$Y_{\text{true}}(x)$	Random errors	$Y_{\text{meas}}(x)$	$Y_{\text{fit}}(x)$	a	b	SMT	SMF	SMT_{mean}	SMF_{mean}
2	Repeated number T =	2000											
3	Postulated α =	1											
4	Postulated β =	0.5											
5	Postulated σ^2 =	1											
6													

Figure 2. The first step of the simulation process.

The second step: Press Alt+F11 to enter the Visual Basic interface, and double-click the Sheet1 icon on the left to enter the writing code interface (or enter this interface by clicking the View of menu bar → Macros → View Macros → input "memeda" in the dialog box as the macro name (example in this article) → Create), then copy the VBA code in the Appendix 1 of this article into the writing code interface (Figure 3).

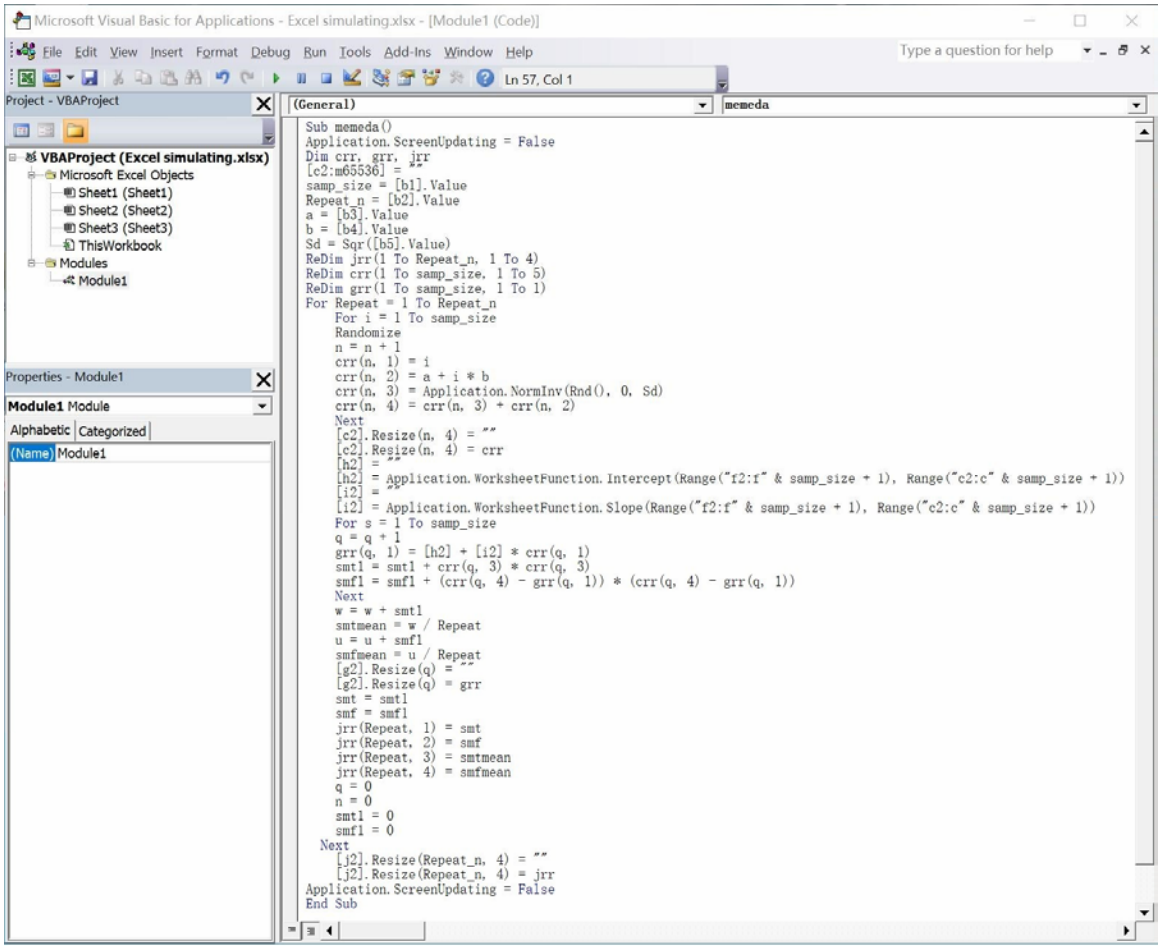


Figure 3. Inputting VBA code in excel.

The last step: Define Ctrl+q as the shortcut key to execute the macro by clicking the View of menu bar → Macros → View Macros → Options → input "q" in the shortcut key box → Ok, then close the Visual Basic window, and save the file as an Excel macro-enabled workbook, i.e., a file with ".xlsm" suffix (Appendix 2). The video of above operating process is attached as Appendix 3. The VBA code is compatible for use in the other versions such as Excel 2007, Excel 2013 and so on.

Supplementary Materials

Appendices associated with this article can be found, in online version, analchemres.org.

4. The Outcome of Simulation

To press Ctrl+q will trigger the simulation program to run automatically. According to the example of this article (Appendix 2), the first step is to generate 30 independent variables in cells from C2 to C31. According to the Equation (3) and the postulated values in cells B3 and B4, 30 values of $Y_{\text{true}}(x)$ are generated in cells from D2 to D31. Then the function formula $\text{NormInv}(\text{Rnd}(), 0, \sigma)$ generates 30 random numbers from the Normal distribution $N(0, \sigma^2)$ in cells from E2 to E31 to simulate ε (the value of σ^2 is defined by the number in cell B5). According to the Equation (4), we have 30 values of $Y_{\text{meas}}(x)$ in cells from F2 to F31. Based on the data in column C and F, we can obtain a fit regression equation, and its intercept a and coefficient b are shown in cells H2 and I2, respectively. According to the data in column C and the Equation (1), then 30 values of $Y_{\text{fit}}(x)$ are presented in cells from G2 to G31. By the Equations (6)-(11), the values of SMT, SMF, SMT_{mean}, and SMF_{mean} are calculated, as shown in cells from J2 to M2. Repeating the above process 2000 times (as defined by the number in cell B2) would consequently generate 2000 sets of those values, as presented in cells from J2 to M2001.

Scatter diagrams are then drawn according to the data of column J and K (Figure 4) and the data of column L and M (Figure 5). Because each repetition represents a random sample, the values of SMT and SMF fluctuate randomly (Figure 4). However, with the increase of repetitions, the averages, i.e., SMT_{mean} and SMF_{mean}, will approach 30 and 28, respectively (Figure 5). If the number in cell B1 is changed to 40 and the number in cell B5 still equals 1, the averages approach 40 and 38; change cell B1 to 20 and B5 to 3, the averages approach 60 and 54. In summary, the difference between the two averages always approaches $2\sigma^2$, since two variables in the sample cannot be assigned freely.

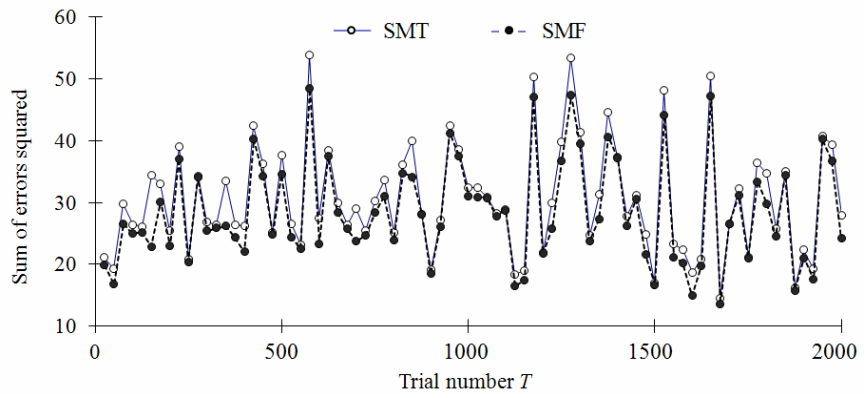


Figure 4. Individual trial values of SMT and SMF (Note: for convenience of demonstration, values are only plotted every other 24 trials, as well in Figure 5).

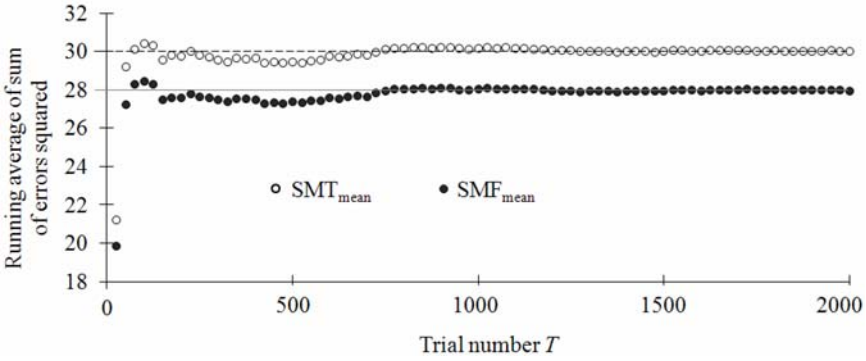


Figure 5. Running values of SMT_{mean} and SMF_{mean}.

5. Discussion

In statistics, some difficult or abstract concepts can be simulated by using computer so as to save time for students to focus on understanding the concepts themselves instead of expending much time in carrying out many calculations [12]. Consequently, many researchers have recommended using computer simulation methods to help teach statistics concepts, particularly for those difficult or abstract concepts [13], which is believed to be able to improve student learning of statistics obviously [2].

The degrees of freedom of SSE of simple linear regression are simulated by using MS Excel 2010 in this article. Once the simulation file is saved, it can be displayed at any time in class. Figure 5 will be changed accordingly after modifying the numbers in cells B1 and B5, which brings students a vividly dynamic graphical display and an interactive experience. Compared with those Web applets, the Excel simulation has many advantages: do not need to keep network connection; can be copied easily, while most applets cannot be downloaded and the Web addresses are not always viable even if the network is connected; every step of the Excel simulation can be explored by students, while the Web applets just offer a kind of black box operation of data in/out without the intermediate step.

It is worth noting that the execution of the simulation file involves a large number of operations and therefore may be time-consuming. The larger the numbers in cells B1 and B2, the better the effect of simulating, and yet the longer time it will take. In order not to affect the normal teaching rhythm, the sample size and the trial number are suggested to be below 30 and 3000, respectively, so that the simulation process will be completed in 20 seconds. Besides, the simulation can also be terminated by pressing the Esc key at any time.

Acknowledgement

This work was supported by the Teaching Quality Project of Huaibei Normal University (jy2016147, 2017kfkc152).

References

[1] J. G. Eisenhauer, Degrees of freedom, Teaching Statistics 30(3) (2008), 75-78.

DOI: https://doi.org/10.1111/j.1467-9639.2008.00324.x

[2] C. Clapham (ed.), The Concise Oxford Dictionary of Mathematics, 2nd Edition, Oxford, UK: Oxford University Press, 1996, pp 65-66.

[3] S. Kotz and N. L. Johnson (ed.), Encyclopedia of Statistical Science (Vol. 2), New York: John Wiley and Sons, 1982, pp. 293-294.

[4] M. G. Marasinghe, W. Q. Meeker, D. Cook and T. Shin, Using graphics and simulation to teach statistical concepts, The American Statistician 50(4) (1996), 342-351.

[5] R. W. West and R. T. Ogden, Interactive demonstrations for statistics education on the world wide web, Journal of Statistics Education 6(3) (1998), 1-8.

https://ww2.amstat.org/publications/jse/v6n3/west.html

[6] V. M. Ng and K. Y. Wong, Using simulation on the internet to teach statistics, The Mathematics Teacher 92(8) (1999), 729-733.

[7] J. Sullivan, Understanding the degrees of freedom concept by computer experiments, The American Statistician 50(3) (1996), 234-237.

[8] P. C. Bell, Teaching business statistics with Microsoft excel, INFORMS Transactions on Education 1(1) (2000), 18-26.

DOI: https://doi.org/10.1287/ited.1.1.18

[9] C. B. Warner and A. M. Meehan, Microsoft Excel™ as a tool for teaching basic statistics, Teaching of Psychology 28(4) (2001), 295-298.

DOI: https://doi.org/10.1207/S15328023TOP2804_11

[10] J. C. Nash, Teaching statistics with Excel 2007 and other spreadsheets, Computational Statistics and Data Analysis 52(10) (2008), 4602-4606.

DOI: https://doi.org/10.1016/j.csda.2008.03.008

[11] A. Naseri, H. Khalilzadeh and S. Sheykhizadeh, Tutorial review: Simulation of oscillating chemical reactions using Microsoft excel macros, Anal. Bioanal. Chem. Res. 3(2) (2016), 169-185.

DOI: https://doi.org/10.22036/abcr.2016.15812

[12] B. Chance, D. Ben-Zvi, J. Garfield and E. Medina, The role of technology in improving student learning of statistics, Technology Innovations in Statistics Education 1(1) (2007).

http://escholarship.org/uc/item/8sd2t4rr

[13] J. D. Mills, Using computer simulation methods to teach statistics: A review of the literature, Journal of Statistics Education 10(1) (2002).

https://ww2.amstat.org/publications/jse/v10n1/mills.html

■

Appendix 1

```
Sub memeda( )
Application.ScreenUpdating = False
Dim crr, grr, jrr
[c2:m65536] = ""
samp_size = [b1]. Value
Repeat_n = [b2]. Value
a = [b3]. Value
b = [b4]. Value
Sd = Sqr ([b5]. Value)
ReDim jrr(1 To Repeat_n, 1 To 4)
ReDim crr(1 To samp_size, 1 To 5)
ReDim grr(1 To samp_size, 1 To 1)
For Repeat = 1 To Repeat_n
    For i = 1 To samp_size
    Randomize
    n = n + 1
    crr(n, 1) = i
    crr(n, 2) = a + i * b
    crr(n, 3) = Application. NormInv (Rnd( ), 0, Sd)
    crr(n, 4) = crr(n, 3) + crr(n, 2)
    Next
    [c2] .Resize(n, 4) = ""
    [c2] .Resize(n, 4) = crr
    [h2] = ""
```

```
[h2] = Application.WorksheetFunction.Intercept(Range("f2:f" & samp_size + 1),
Range("c2:c" & samp_size + 1))

[i2] = ""

[i2] = Application.WorksheetFunction.Slope(Range("f2:f" & samp_size + 1),
Range("c2:c" & samp_size + 1))

For s = 1 To samp_size

q = q + 1

grr(q, 1) = [h2] + [i2] * crr(q, 1)

smt1 = smt1 + crr(q, 3) * crr(q, 3)

smf1 = smf1 + (crr(q, 4) - grr(q, 1)) * (crr(q, 4) - grr(q, 1))

Next

w = w + smt1

smtmean = w / Repeat

u = u + smf1

smfmean = u / Repeat

[g2].Resize (q) = ""

[g2].Resize (q) = grr

smt = smt1

smf = smf1

jrr(Repeat, 1) = smt

jrr(Repeat, 2) = smf

jrr(Repeat, 3) = smtmean

jrr(Repeat, 4) = smfmean

q = 0

n = 0

smt 1 = 0

smf 1 = 0
```

```
    Next
      [j2] .Resize (Repeat_n, 4) = ""
      [j2] .Resize (Repeat_n, 4) = jrr
    Application.ScreenUpdating = False
    End Sub
```