TEKA Kom. Mot. Energ. Roln. – OL PAN, 2008, 8, 277–281

REMARKS ON THE FAULTY REGRESSION IN EXCEL

Joanna Tarasińska, Zofia Hanusz

Department of Applied Mathematics and Informatics
University of Life Sciences in Lublin

Summary. In the paper we focus on the regression problem solving by Excel. Some faults concerning R-square coefficient are presented. The difference between R-squares in two models with and without intercept is considered.

Key words: regression model, statistics in Excel, R-square.

INTRODUCTION

Microsoft Excel is very widespread as it is integrated within the Microsoft Office that increases its availability. It has got many build-in functions. Thus it is tempting to make statistical analysis with it. However, users should realize that this is NOT the statistical package. There are many publications pointing numerous faults and errors in statistical analysis in Excel (Cook et al. 1999,Goldwater 1999, Simonoff 2000, Cryer 2001, Heiser 2006, Knűsell 1998, Knűsell 2004). An especially extensive elaboration focused on this issue is given in Heiser (2006). What is worse, many of these well-known faults are fixed in subsequent upgrades of Excel. Almost in all papers concerning statistical procedures in Excel there is advice that the program should not be used for serious statistical analysis. However, many practitioners , even statisticians, use Excel for quick and simple calculations. The aim of this paper is to warn that results of such calculations should be considered with great caution, especially if they could have serious impact on human life or health.

In the literature many papers are focused on numerical inaccuracy of Excel intrinsic function, especially the one connected with probability distributions (Cox et al. 1999, Pottel 2001, Knűsell 2005). But even for such a simple and widely used function as Standard Deviation the calculations can be misleading in the case of untypical data with deviations of small relative to the absolute value of data. Heiser (2006), Cryer (2001), Simon (2000) point out such absurd errors in statistical analysis in Excel 2000 as negative sum of squares, negative R-square, wrong degrees of freedom etc. Some of them are fixed in Excel 2003 and 2007. Some errors are the most visible in the case of untypical "non-easy" data for example when standard deviations are small relative to the absolute level of data.

In this paper we are going to focus on a very simple model of linear regression. This model is very often used by practitioners who maybe are not warned that you can not rely on Excel results.

STATISTICAL BACKGROUND

Let us remind some very well known statistical formulas used in linear regression. Let us consider the following model:

$$y_i = \alpha + \beta x_i + e_i \quad (i = 1,\ldots,n), \tag{1}$$

where y_i is an observation of dependent variable Y, x_i is a value of independent variable X, e_i is random error. We assume that errors are mutually independent and normally distributed with null mean and unknown variance σ^2, α and β are constants to be estimated. The least squares estimators of them, for which $\min\limits_{a,b}\sum_{i=1}^{n}(y_i - a - bx_i)^2$ is achieved are equal to: $\hat{\beta} = b = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ and $\hat{\alpha} = a = \bar{y} - b\bar{x}$, where $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$. The predicted value for dependent variable Y is $\hat{y} = a + bx$. Statistical test for the regression can be made by means of analysis of variance in which total sum of squares $SSY = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is decomposed into sum of squares for the regression $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ and sum of squares for error . The test is based on the test statistics $F = \dfrac{MSR}{MSE}$ where $MSR = SSR$ and $MSE = \dfrac{SSE}{n-2}$. As a measure of fit we take:

$$R^2 = 1 - \frac{SSE}{SSY} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

The coefficient R^2 gives the information on how much of total variability of Y is explained by the regression. However, this coefficient has a disadvantage because it does not take into account the number of observations and in the extreme, trivial case of only two observations it takes the value one. Thus, the so called adjusted R-square is also used in statistics where instead of sum of squares, the mean squares are taken: $adj\,R^2 = 1 - \dfrac{SSE/\nu_e}{SSY/\nu_{total}}$ with $\nu_e = n - 2$ and $\nu_{total} = n - 1$.

The other linear model which can be sometimes considered is the one with intercept α being zero i.e. the regression line through origin:

$$y_i = \beta x_i + e_i \quad (i = 1,\ldots,n). \tag{2}$$

In the model (2) the least square estimate for β, which minimizes $\sum_{i=1}^{n}(y_i - bx_i)^2$ is equal to $\hat{\beta} = b = \dfrac{\sum_{i}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ and the predictor $\hat{y} = bx$. In analysis of variance the total sum $SSY = \sum_{i=1}^{n} y_i^{\,2}$ is decomposed into a sum of squares for errors $SSE = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ and a sum of squares for the regression $SSR = SSY - SSE = \sum_{i=1}^{n} y_i \hat{y}_i$. In this case, as a measure of fit the following coefficient is taken:

$$R^2 = 1 - \frac{SSE}{SSY} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} y_i^2} = \frac{\sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n} y_i^2}.$$

It should be noted that in the model (2) the coefficient R^2 does not have this 'friendly' interpretation as in the model (1) and R^2's should not be compared in both models. Now the variability about the origin is taken into account. The adjusted R-square is: $adj\,R^2 = 1 - \frac{SSE/\nu_e}{SSY/\nu_{total}}$, with $\nu_e = n-1$ and $\nu_{total} = n$.

FAULTY RESULTS OF REGRESSION THROUGH ORIGIN IN EXCEL 2003

The simplest and the quickest way to find linear regression equation and R^2 in the model (2) is to make X-Y scatter graph and add a trend line by right-clicking the data points in the graph with options 'Display R^2 and equation on the chart' and 'null intercept'.

The other way is to use the tool 'Regression' in Analysis ToolPak (Tools-Data Analysis-Regression) with option 'Constant is zero'. In this case one get not only the equation and R^2 but also adjusted R^2, ANOVA table for the regression and the confidence interval for the coefficient β. Unfortunately, R^2 is badly computed in the trend line of the chart. In Analysis ToolPak the adjusted R^2 and *p*-value for test *F* are badly computed.

As an example let us take 4 pairs of observations given in Table 1. The X-Y graph together with trend line and computed R^2 is given in Fig. 1.

Table 1. Data set for an analysis

x_i	1	2	3	4
y_i	1.7	3.1	3.3	4.4

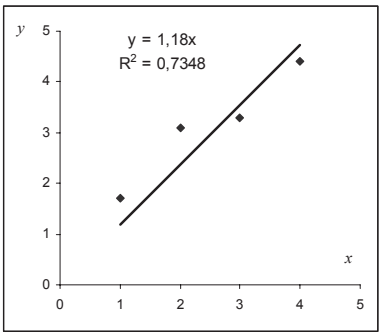

Fig. 1. X-Y graph with the trend line and R^2

The Analysis ToolPak calculations are given in Tables 2, 3 and 4, respectively.

Table 2. Regression statistics

Regression Statistics	
multiple R	0.988495
R square	0.977123
adjusted R square	0.643789
standard error	0.570964
number of observations	4

Table 3. ANOVA table

	df	*SS*	*MS*	*F*	*p-value*
Regression	1	41.772	41.772	128.135	0.007714
Error	3	0.978	0.326		
Total	4	42.75			

Table 4. Detailed results for regression

	coefficients	*standard error*	*t Stat*	*p-value*	*lower 95%*	*upper 95%*
intercept	0	#N/D!	#N/D!	#N/D!	#N/D!	#N/D!
slope	1.18	0.104243	11.31967	0.001479	0.848251	1.511749

It is easy to check that R^2 given in the chart in Fig. 1 is badly calculated because it should be equal to 0.977123. This value is well calculated in Table 2. In trend line Excel 'mixes' models (1) and (2) taking $R^2 = 1 - \frac{SSE}{SSY}$ where *SSE* is taken from the model (2) whereas *SSY* is taken from the model (1), i.e. $SSY = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 3.6875$.

It should be noted that wrong formula for R^2 in the trend line can even lead to an absurd negative value R^2 as it is presented in Fig. 2 in which the point (4,4.4) from the previous example is replaced by the point (4,3.4).

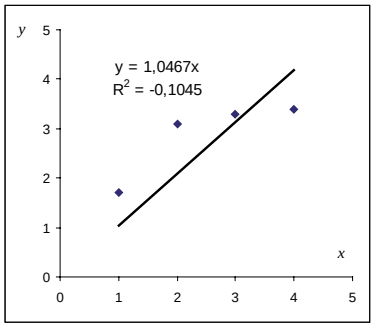

Fig. 2. The example of an absurd negative R^2

In Table 2 the adjusted R^2 is badly calculated. It should be $adj\,R^2 = 1 - \frac{0.978/3}{42.75/4} \cong 0.9695$, the value got for example in STATISTICA. Let us also notice that *p*-values in Table 3 and 4 should be the same. The proper one is enclosed in Table 4. The value 0.007714 in Table 3 results again from faulty mixing models (1) and (2). Namely, it is $Pr(F > 128.135)$ where *F* is the random value distributed as *F* with (1,2) degrees of freedom as it is in the model (1), instead of (1,3) degrees of freedom according to the model (2).

Similar problems with faulty calculations as described above can occur of course in multiple or multinomial regression models. Fig. 3 shows both regression lines with and without intercept fitted to real experimental data concerning dependence of cohesion against hardness of bisquits (Grzegorczyk 2008). In analysis of variance table for the quadratic regression without an intercept R^2 is approximately equal to 0,83 and is bigger then R-square for the curve with the intercept al-

though the first one doesn't fit the points. Thus, as it was mentioned in Introduction the coefficients R^2 in both models can not be compared.

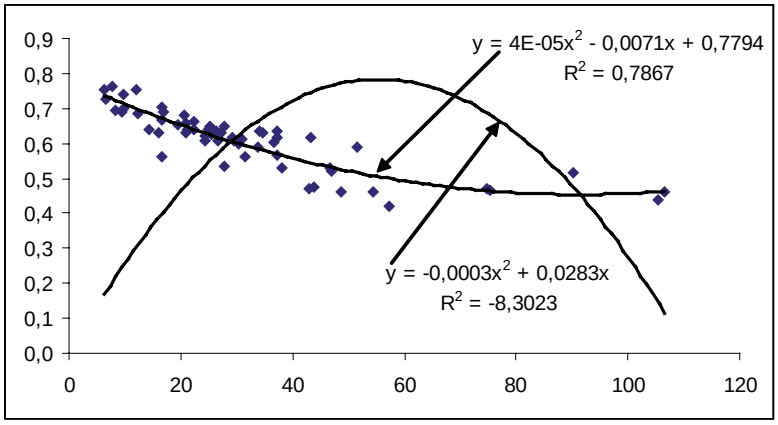

Fig. 3. Parabolas fitting experimental points with and without an intercept

Additionally in the case of multiple regression many authors report bad results of calculations in earlier Excel versions in presence of near-singularity (collinearity) of the design matrix. However McCullogh and Wilson (2005) point that this problem is corrected in Excel 2003.

CONCLUSIONS

The paper has pointed out that Microsoft Excel is not good statistical program to make statistical calculation concerning regression analysis. Especially in the case where users apply drawing experimental points and put a regression line without an intercept.

Moreover we wanted to underline that R-squares in models with and without intercepts cannot be compared because they have different interpretation. Greater R-square for the model without an intercept does not mean that this model is better. In most situations this model does not fit the experimental sets of data.

REFERENCES

Cook H.R, Cox M.G., Dainton M.P.,Harris P.M., 1999, Testing the intrinsic functions of Excel, Report to the National Measurement System Policy Unit, Department of Trade & Industry, http://publications.npl.co.uk/npl_web/pdf/cise27.pdf

Cox M.G, Dainton M.P., Harris P.M., 1999, Testing functions for linear regression, Report to the National Measurement System Policy Unit, Department of Trade & Industry, http://publications.npl.co.uk/npl_web/pdf/cmsc08.pdf

Cryer J.D. 2001,Problems with using Microsoft Excel for statistics, Joint Statistical Meetings, August 2001, Atlanta Ga, http://www.stat.uiowa.edu/~jcryer/JSMTalk2001.pdf

Goldwater E., 1999,Using Excel for statistical data analysis,
www-unix.oit.umass.edu/~evagold/excel.html